Psychological Methods
1997, Vol. 2, No. 2, 161-172

1082-989X/97/$3.00

On the Logic and Purpose of Significance Testing

Jose M. Cortina
George Mason University

William P. Dunlap
Tulane University

There has been much recent attention given to the problems involved with the traditional approach to null hypothesis significance testing (NHST). Many have suggested that, perhaps, NHST should be abandoned altogether in favor of other bases for conclusions such as confidence intervals and effect size estimates (e.g., Schmidt, 1996). The purposes of this article are to (a) review the function that data analysis is supposed to serve in the social sciences, (b) examine the ways in which these functions are performed by NHST, (c) examine the case against NHST, and (d) evaluate interval-based estimation as an alternative to NHST.

The topic of this article is null hypothesis significance testing (NHST; Cohen, 1994). By this we mean the process, common to the behavioral sciences, of rejecting or suspending judgment on a given null hypothesis based on a priori theoretical considerations and p values in an attempt to draw conclusions with respect to an alternative hypothesis. We should begin by saying that we agree with J. Cohen, G. Gigerenzer, D. Bakan, W. Rozeboom, and so on with respect to the notion that the logic of NHST is widely misunderstood and that the conclusions drawn from such tests are often unfounded or at least exaggerated (Bakan, 1966; Cohen, 1990, 1994; Gigerenzer, 1993; Rozeboom, 1960). Nevertheless, we think it important that the extent and likely direction of such problems be carefully examined, because NHST, when used and interpreted properly, is useful for certain purposes and is only partially problematic for others.

This article is divided into three parts. The first part is devoted to an examination of the role played by data analysis in the social sciences and the extent to which NHST supports this role. The second part presents and critiques the case against NHST as an analytic tool. The third part investigates the viability of confidence intervals as an alternative to NHST.

Jose M. Cortina, Department of Psychology, George Mason University; William P. Dunlap, Department of Psychology, Tulane University.

We are most grateful for the comments of John Hollenbeck, Richard P. DeShon, Mickey Quinones, Sherry Bennett, Frank Schmidt, Jack Hunter, and the members of the Behavioral Science Research Group, without whom this article would have suffered.

Correspondence regarding this article should be addressed to Jose M. Cortina, Department of Psychology, MSN 3F5, 4400 University Drive, Fairfax, Virginia 22030. Electronic mail may be sent via Internet to jcortina @gmu.edu.

Purpose of Data Analysis

Perhaps the best way to begin is to consider the rationale of the typical research project in the social sciences and the role played by data analysis in the structure of the project. Every research project begins with a research question. Let us assume that the question has to do with a bivariate relationship that has never been examined empirically before. The first step in the project typically involves the generation of a theoretically based answer to the question, that is, a hypothesis. This hypothesis is often based on a combination of reason and previous empirical work in related areas and gives us evidence of a certain kind for a particular answer to the research question. This form of evidence is invaluable, but scientific tradition holds that corroboration, whether in the form of falsification or justification, is desirable (Lakatos, 1978; Popper, 1959; Serlin & Lapsley, 1985). If that corroboration is independent of the theoretical basis for the initial answer to the research question, then the corroboration is all the more impressive. This independent corroboration often takes the form of empirical data and allows us to attack the research question from directions that are largely orthogonal to one another.

The stronger the theoretical basis for the initial answer, the less reliance one need put on the data. For example, it is said that Albert Einstein had no interest in empirical tests of his theory of relativity. The

theory was so strong that any competently collected empirical evidence would support the theory. Thus, the empirical evidence was largely superfluous.

On the other hand, we in the social sciences are rarely if ever justified in placing so much confidence in theoretical answers. Nevertheless, we can certainly think of such hypotheses as lying on a continuum of supportability that ranges from weak to strong. If multiple, well-respected theoretical perspectives suggest the same hypothesis, then the theoretical support for that position is strong, and one is less reliant on the data. If, on the other hand, the theoretical support for a given position is weak, then the burden of proof shifts to the data.

Let us assume for the moment that the theoretical basis for a given hypothesis is of average strength. Let us assume further that the experiment designed to provide corroboration for this hypothesis involves sound measures, reasonable procedures, and so on, and that data are collected within this design. The Neyman-Pearson framework on which much of modern significance testing is based suggests that the desired outcome of data analysis is the adjustment of our confidence in our hypotheses so that we can behave as if a given hypothesis is true or false until further evidence is amassed (Neyman & Pearson, 1928, 1933). But the question remains, how does one decide whether and the extent to which the data allow us to increase our confidence in the theoretically based answer to the research question?

Let us consider for a moment what it is that we have to work with in the social sciences, particularly in comparison with the physical sciences. Largely because of the complexity of the variables studied in the social sciences, our theories are not powerful enough to generate point hypotheses. Physical scientists often scoff at our attempts to make general statements of the form "A increases B" without specifying the precise degree to which A increases B. The example of 32 ft./s/s has been used in arguing that we need to focus on parameter estimation (e.g., regression weights) instead of saying things such as "Gravity makes things fall." Although we agree with this sentiment in principle, it strikes us as being a bit naive when applied to the study of human behavior. The constraints placed on population values by theory are much weaker in the social sciences than they are in the physical sciences (Serlin & Lapsley, 1985). The example of 32 ft./s/s is a more or less immutable law of nature (on Earth of course). It is always true, so it makes sense to apply a specific number to it. What happens if we change the question a bit? Suppose we wish to know the rate at which stock prices drop after interest rates are raised. In some cases, a given increase in interest rates produces a precipitous drop; in other cases the same increase produces little or no drop. We can compute an average and use this to draw some general conclusions, but it would be difficult to predict the drop rate in any given instance with a great deal of accuracy. The point is that the vagaries of the stock market and other such phenomena are created by human decisions and behavior. There is often too much complexity in such situations and perhaps too much variability in people for a theoretically based point parameter estimate to make much sense. We suggest that there are many situations in which we are better off sticking to more conservative, general predictions (and conclusions) such as A increases B.

Regardless of the form of the hypothesis, it must be exclusionary in order to make a contribution. This usually, though by no means necessarily, takes the form of a dichotomous prediction of some kind. For example, the hypotheses "A affects B" and "The 95% confidence interval will not contain zero" are dichotomous and exclusionary in the sense that A either affects B or does not, or the interval contains zero or not, and our hypotheses exclude the alternative possibilities. Empirical evidence is then gathered and evaluated in terms of the extent to which we can adjust our confidence in the relevant hypotheses. In other words, we use empirical corroboration (or lack thereof) to adjust a degree of rational belief (Keynes, 1921). Where it is necessary to make a decision based on these levels of confidence, we then behave as if the null were true or we behave as if the alternative were true. This behavior can take the form of policy change, as is often the case in the applied social sciences, or it can represent one step in a series of steps associated with the test of a theory (e.g., path analysis or measurement development). As always, it behooves us to keep in mind that the underlying basis for the decision is continuous.

What is required of empirical corroboration? There are at least three interrelated requirements for empirical corroboration: (a) objectivity, (b) exclusion of alternative hypotheses (e.g., a hypothesis of a relationship with a different sign or an alternative structural model), and (c) exclusion of alternative explanations (e.g., confounds or sampling error). It can be shown that proper experimental design followed by significance testing allows one to address all three of these requirements. No analysis procedure, including

significance testing, can meet these requirements perfectly, but significance testing, for a variety of reasons, addresses these critical issues as well as or better than do the alternatives for many types of research questions. Before discussing these three requirements, however, let us take a moment to consider the reason that corroboration itself is critical.

In the perfect world, perhaps every piece of research would be published somewhere. In a slightly less perfect world, any study with compelling theory and an adequate method would be published somewhere. Given limited journal space, however, we are forced to choose from among submissions those articles that advance knowledge with the greatest efficiency. The article that provides a certain theoretical answer to a question, but also presents data that suggest a different answer, does advance knowledge. However, all else being equal, it does not advance knowledge as far as does the article with theoretical and empirical answers that are in agreement. The first type informs us that either the theory is wrong or the method is flawed. Thus, we are unable to adjust our confidence in any particular answer to the relevant research question. The second type, by contrast, allows us to increase our confidence in a given answer by virtue of the independent sources suggesting that answer. Given that corroboration is important, significance testing is useful because it gives us a mechanism for adjusting our confidence in certain answers. This can be shown in the context of the three requirements for corroboration mentioned above.

One of the most critical requirements of corroboration is objectivity. The desire for objective verification can be traced at least as far back as Kant's (1781) *Critique of Pure Reason* and his references to intersubjectivity, but the modern notions of objectivity and value freedom in the social sciences appear to stem from M. Weber (Miller, 1979). Empirical corroboration should be objective in the sense that it should be as independent of the theoretically based answer to the research question as is possible (Dellow, 1970; Miller, 1987). In this way, it can complement entirely the theoretical evidence. While perfect objectivity may not be possible (Kuhn, 1962; Lakatos, 1978), it is important that we strive for separation of theoretical and empirical evidence.

It is largely because of this need to separate theoretical from empirical evidence that Bayesian statistics, with its reliance on arbitrary prior probabilities, has been used seldom to test scientific hypotheses (Putnam, 1981). Instead, we ask that theoretical answers, which are typically developed by the researcher that asked the research question and are therefore susceptible to intrasubjectivity, be separated as much as possible from those empirical answers. In proper significance testing with adequate levels of power, we compare observed results with a priori cutoffs to decide whether or not we will adjust our confidence in the hypothesis at hand. These cutoffs are somewhat arbitrary, but they are also intersubjective in the sense that they are conventional, and they are chosen before the data are examined. Thus, while it can be argued that the choice of cutoff is somewhat whimsical, it is largely the whim of others (as opposed to the experimenter at hand) that determines the cutoff, and whatever impact the whim of the experimenter has on the choice of cutoff, this impact takes place prior to examination of data. This allows one to meet the Popperian requirement that the conditions necessary for the drawing of certain conclusions are determined beforehand, ceteris paribus clauses notwithstanding (Serlin & Lapsley, 1985). Of course, observed probability values should be reported so that the reader can draw his or her own conclusions, but the conclusions of the experimenter are to be based on predetermined criteria.

A second demand of empirical corroboration is that it allows us to rule out alternative hypotheses. If our prediction is that "A has a positive effect on B," then it is desirable that our data allow us to assess the plausibility of alternative hypotheses such as "A has a negative effect on B" or "A has a negligible effect on B." In this way, we can show that our theory explains phenomena to a greater extent than do alternatives (Lakatos, 1978). Significance testing allows the comparison of theoretical hypotheses to a null hypothesis. The term *null* has typically been used to represent the hypothesis that is to be nullified (Cohen, 1994). This null hypothesis can be that there is no effect or no relationship, but this need not be the case. Regardless of the form of the null, significance testing involves a comparison of hypothetical distributions; specifically, it involves the comparison of the sampling distribution associated with the observed result to the distribution associated with the null. If it is highly unlikely that the null distribution would have produced the observed result, and if this discrepancy is in the anticipated direction, then we adjust our confidences such that we tentatively rule out the null hypothesis (as well as distributions that are even less probable than the null) in favor of the theoretically based hypothesis.

A third demand of empirical corroboration is that it be subject to as few alternative explanations as possible. Thus, we hope to be able to point to our data and claim not only that it is as one would expect given the theory, but also that its alignment with the theory is not due to extraneous factors. If one accepts the approach of John Stuart Mill (1872), then there are three criteria for the inference of cause: temporal precedence, covariation, and the elimination of alternative explanations for the covariation. The first two present relatively few problems. It is the problem of alternative explanations that gives us the most trouble. For any given instance of covariation, there are likely to be a multitude of possible alternative explanations such as the impact of unmeasured variables, the choice of sample, and so on. One potential alternative source of covariation that has received a great deal of attention in recent years is sampling error.[1] The p value in NHST gives the probability that the data would have occurred given the truth of the null hypothesis (Carver, 1978). In other words, p is the probability that the departure of the test result from null would have resulted from sampling error alone.[2] Although our cutoffs (.05 and .01 typically) are arbitrary, they allow us to evaluate the extent to which sampling error is a viable alternative explanation. If our p value is less than the preset cutoff, then we can be reasonably confident (assuming all else is in order) that sampling error was not the reason that our test statistic differed from our null value. Thus, we have moved toward satisfying the third of Mill's criteria for inferring causality. We have effectively ruled out sampling error as an alternative explanation for the departure of our test result from null.

In summary, the purpose of data analysis is to provide corroboration (or fail to provide corroboration) of theoretical answers to research questions. This corroboration is most convincing when it is intersubjective and disconfirming of alternative hypotheses and explanations. We clearly show that NHST does in fact provide a form of corroboration that contains each of these elements.

Case Against NHST

The case against NHST has taken various forms. The most common attack of recent years has involved pointing out the interpretational problems associated with NHST when conducted in the presence of methodological flaws (e.g., small samples; Schmidt, 1996) and experimenter ignorance (e.g., interpretation of conditional probabilities of empirical results as conditional probabilities of hypotheses; Cohen, 1994). There can be no doubt that the presence of methodological flaws in a study limits the conclusions that can be drawn from NHST. Of course, such flaws limit the conclusions that can be drawn from any procedure, so we see no point in holding NHST or any other procedure accountable for the shortcomings of the data to which they are applied.

On the other hand, we would like to address what we consider to be the most glaring problems associated with the criticisms involving experimenter ignorance. The criticisms are that (a) many experimenters commit the error of interpreting the conditional probability of the empirical result, $P(D|H)$, also known as the p value, as the conditional probability of the hypothesis, $P(H|D)$; (b) the probabilistic nature of NHST creates logical problems; and (c) NHST is misleading in that it focuses on control of Type I errors when the probability of such errors is zero.

Interpretation of p

The p value gives the probability that the observed empirical result would have occurred given a certain hypothetical distribution. We agree completely with Cohen (1994), Gigerenzer (1993), Rozeboom (1960), and others in their observation that many researchers tend to misinterpret the p value as the probability of a hypothesis given the data. However, this is the fault of those who are doing the interpreting, not the tools that they choose. Also, it is important to recognize that there are many situations in which interpretation of a small value of $P(D|H_0)$ as indicating a small value of $P(H_0|D)$ can make sound, practical sense. To show this, let us first examine Cohen's (1994) excellent example from clinical/abnormal psychology. It goes like this.

The base rate for schizophrenia in adults is roughly 2%. Suppose a given test will identify 95% of people

[1] The term *sampling error* is used here to represent any difference between statistics based on samples drawn from the same population (cf. Hunter & Schmidt, 1990).

[2] While no one argues with the previous sentence, the present sentence may rankle for some. It seems to us, however, that a discussion of the probability of departure from null implies that the null is taken to be true, and such a discussion with such an implication is equivalent to a discussion of the probability of a result given the truth of the null.

with schizophrenia as being schizophrenic and will identify 97% of "normal" individuals (in the clinical sense) as normals. The data or empirical outcomes, in this case, are the results of the test for schizophrenia. The hypotheses correspond to the true nature of the testee. So, P(normal test result|normal testee) = .97, while P (schizophrenic test result|schizophrenic testee) = .95. Stated formally, $P(D_0|H_0) = .97$ and $P(D_1|H_1) = .95$, while $P(D_1|H_0) = .03$ and $P(D_0|H_1) = .05$. We also have the prior probability or base rate of occurrence for the null hypothesis, $P(H_0) = .98$. What we want to know, of course, is the probability that a given person truly is schizophrenic in light of the test result. Formally, we want $P(H_0|D_1)$ or $P(H_1|D_0)$. The Bayesian equation for $P(H_0|D_1)$ is[3]

$$P(H_0|D_1) = \frac{P(H_0) * P(D_1|H_0)}{P(H_0) * P(D_1|H_0) + P(H_1) * P(D_0|H_1)}.$$

Using this equation, Cohen (1994) showed that $P(H_0|D_1) = .607$. Substantively, if the test says that a person is schizophrenic, then the person will actually turn out to be schizophrenic only 39% of the time. Of the positive test results, 61% are wrong!

The point that we wish to make, however, is that this result is problematic from only one perspective, namely, that of the person who wants to find schizophrenics. Consider another perspective: that of a person in charge of hiring police officers. Such a person would, in most cases, try to avoid hiring people whose schizophrenia (or any other attribute) would be debilitating with respect to job performance. In other words, the employer's purpose is to make sure that a given applicant is, in fact, normal in the clinical sense of the word. The test from Cohen's (1994) example identifies 95 of every 100 people as normal, of whom 94.9 really are normal. Therefore, the conditional accuracy of a normal result from the test (i.e., $P(H_0|D_0)$) is 94.9/95, or .999! Only 1 out of every 1,000 people with a normal result would actually suffer from schizophrenia.

So, is this test useful to an employer? Without the test, the employer would hire 20 people with schizophrenia in every 1,000 hiring decisions. With the test, the employer would get 1 person with schizophrenia in every 1,000 hiring decisions. Thus, the odds of hiring a person with schizophrenia without the test are greater by a factor of 20! The point is that, not surprisingly, the context largely determines the problems caused by interpreting the results of NHST in a certain way. Of course, we are only echoing statements made almost 40 years ago. Rozeboom (1960) pointed out that the probabilities associated with our hypotheses are not the only considerations when deciding whether or not to accept or reject hypotheses. We must also consider the "utilities of the various decision outcomes" (Rozeboom, 1960, p. 423). For certain types of decisions, a procedure that is prone to mistakes of one kind can be devastating because of the utilities associated with those mistakes, whereas a different procedure that is prone to different kinds of mistakes can be quite useful because the mistakes that it makes result in "missing on the safe side."

Syllogistic Reasoning and Probabilistic Statements

It has been suggested that the logic of NHST is, if you will, illogical. Consider the issue as presented by Cohen (1994), who set up various syllogisms representing different ways of viewing the logic of hypothesis testing. Cohen (1994) pointed out that while the rule of Modus Tollens can be universally applied to premises of the form "If A then B, not B," resulting in the conclusion "Not A," it cannot be universally applied to the premises, "If A then probably B, not B" to conclude "Probably not A." This sequence of statements is meant as an analog for the statements that are implicit in NHST. If the null were true, then a sample taken from the population associated with the null would probably produce a statistic within a certain range (i.e., If A, then probably B). The statistic from our sample is not within that range (i.e., not B). Ergo, a population associated with the null value probably did not produce our sample (i.e., probably not A). Cohen then gave an intriguing example that highlights one of the problems that can arise as a result of applying the Modus Tollens to probabilistic statements. The example is as follows:

> If a person is an American, then that person is probably not a member of Congress.
> This person is a member of Congress, therefore,
> This person is probably not an American.

In this case, the two premises are perfectly true, and yet Modus Tollens fails to lead us to a reasonable conclusion.

As we stated above, this example is intriguing, but

[3] Conditional probabilities of other hypotheses can be computed using similar equations.

it is limited in its generalizability for two reasons. It is important that these reasons be understood so that the situations in which the logic of NHST is and is not questionable can be identified. First, the consequent of the first premise, "That person is probably not a member of Congress," is true in and of itself. Any given person is probably not a member of Congress. As a result, one could use almost anything as the antecedent of this premise without damaging the accuracy of the premise. "If one out of every three cows is blue, then this person is probably not a member of Congress" is just as valid as the first premise from Cohen's (1994) example. We explain the importance of this feature of the example momentarily.

The second limiting aspect of the first premise is that while the first premise is true as it stands, it is also the case that being an American is a necessary condition of being a member of Congress. In other words, while it is true that "If a person is an American, then that person is probably not a member of Congress," it is also true if a person is a member of Congress, then that person has to be an American. It is because of these two aspects of the particular example chosen that the Modus Tollens breaks down. Consider a different example, one that is more representative of psychology:

> If Sample A were from some specified population of "normals," then Sample A probably would not be 50% schizophrenic.
> Sample A comprises 50% schizophrenic individuals; therefore,
> Sample A is probably not from the "normal" population

In this example, the consequent of the first premise still stands by itself, that is, a given sample of people probably will not comprise 50% people with schizophrenia. However, this statement is particularly true if the antecedent holds; whereas, in the Cohen example, the consequent is particularly true if the antecedent does not hold. To clarify this, consider a third example:

> If the planets revolve around the sun, then Sample A probably would not be 50% schizophrenic.
> Sample A comprises 50% schizophrenic individuals; therefore
> The planets probably do not revolve around the sun.

As with Cohen's example, the conclusion here is false. Modus Tollens fails to lead to a reasonable conclusion because the truth of the antecedent of the first premise is unrelated to the truth of its consequent. So, while it is the case that Modus Tollens cannot be applied to probabilistic premises when the truth of the antecedent of the first premise is unrelated or negatively related to the truth of the consequent of the premise, it is approximately correct for and can be applied to arguments, typical of psychology, in which the truth of the two components of the first premise are positively related. In other words, the typical approach to hypothesis testing does not violate the relevant rule of syllogistic reasoning to any great degree. Cohen's (1994) example was useful in that it showed why application of Modus Tollens to probabilistic statements can be problematic, but it should not be taken to mean that this rule of syllogistic reasoning is useless for psychology.

Interpretation of Error Rates

This brings us to the issue of interpretation of Type I and Type II error rates, usually represented as α and β, respectively. There seems to be some confusion as to the meaning of these values. For example, Schmidt (1996) stated repeatedly that the Type I error rate, α, is zero, as opposed to .05 or .01, or whatever the predetermined cutoff value is. Cohen (1994) made similar statements. Their reasoning is that because the hypothesis of no effect is never precisely true, it is not possible to falsely reject the null hypothesis (see Frick, 1995, for an alternative position). In other words, the null is always false, so rejecting the null cannot be an error. This may be true, but it has nothing to do with the Type I error rate.

The Type I error rate, α, is the probability that the null would be rejected if the null were true. Note that there is no suggestion here that the null is or is not true. The subjunctive *were* is used instead of *is* to denote the conditional nature of this probability. The Type I error rate is the probability that the hypothetical null distribution would produce an observed value with a certain extremeness. If this value is set at .05, then in order for the observed test result to be considered statistically significant, it would have to be a value so extreme that it (or a value more extreme) would occur 5% of the time or less if we repeatedly sampled from a null distribution. The .05 value is the Type I error rate, regardless of whether or not the null is true. Even if we know the null to be false, the Type I error rate is still .05 because it has to do with a hypothetical distribution, not the actual sampling distribution of the test statistic. Alpha is not the probability of making a Type I error. It is what the prob-

ability of making a Type I error would be if the null were true. One can, perhaps, argue that the term *Type I error rate* is misleading. A better term might be *conditional Type I error rate.* Regardless of the term used, however, the value that we choose for α is the Type I error rate regardless of the truth of the null.

This is not to say that the Type I error rate is the only error rate on which we should focus or that this error rate alone allows one to determine the importance of our empirical results. The Type II error rate, which is also a conditional error rate, is at least as important as the Type I error rate. Furthermore, as is well known, for large sample sizes, the null can be rejected regardless (almost) of the effect size and regardless of the Type I error rate that is chosen. Thus, the Type I error rate is only one of many considerations in a test of significance. Nevertheless, it is important that the meaning of these values and their conditional nature be understood so that further misinterpretation does not occur.

Another issue with respect to interpretation of error rates has to do with the "null versus nil hypothesis" distinction. As pointed out by Cohen (1994), the term *null hypothesis* receives its name by virtue of the fact that it is the hypothesis to be nullified. Thus, the value associated with this hypothesis need not be zero. It can be any value against which we wish to compare the empirical result. The *nil hypothesis,* according to Cohen (1994), would be a null hypothesis for which the value to be nullified is precisely zero. The point that Cohen (1994), Thompson (1992), and others have tried to make is that because the nil hypothesis is always false, there is no glory in rejecting it. It is a "straw man" that is set up for the purpose of being knocked down. Their point is well taken (although Frick, 1995, and others have argued that there are situations in which the nil can be precisely true), but some authors have taken it further than it can go. For example, Schmidt (1996) used the position that the nil is always false to suggest that all research that has compared a research hypothesis with the nil hypothesis is worthless (except as fodder for meta-analyses). While it is certainly true that the social sciences should expand their methodological thinking to include null hypotheses other than the nil hypothesis, it is not true that the use of the nil renders previous research worthless. For example, suppose that there are theoretical reasons for positing a positive relationship between two variables. In an attempt to investigate this relationship, data from 100 subjects are collected, and the correlation is found to be .40. This value can be converted into a t score (4.32), which is greater than any cutoff value that is likely to be relevant for our choices of test and significance level. Thus, it is highly unlikely that this result would have occurred if the nil hypothesis were true. We would, therefore, proceed as if the nil were false and the research hypothesis were true pending further information (cf. Neyman & Pearson, 1928, 1933).

However, as pointed out by others, the nil is something of a straw man. It allows one to address the question, "How likely is it that a population with a correlation of zero would produce a given sample-based correlation?" It would be more interesting to ask whether or not it is likely that a population with a trivial correlation would produce a given sample-based correlation. While it is true that one person's whopping effect is another person's trivium, let us assume for the moment that any variable that explains less than 1% of the variance in another variable explains only a trivial amount of variance. Instead of using the nil hypothesis, we might use a null value of .10 (which is the square root of .01). Thus, we would compare our observed value of .40 with the null value .10 instead of the nil value. This test requires that we convert both our observed correlation and our null value to z scores with the Fisher r to z transformation, which yields z values of 0.4236 and 0.1003. We then compute a z value representing the difference between these values. For the present example, this z value is 3.185, which is also greater than any cutoff value that is likely to be relevant for our choices of test and significance level. Thus, the outcome is the same for this test as it was for the test involving the nil hypothesis: We would proceed as if the research hypothesis were true pending further information. This is not to say that it makes no difference which value we choose as the null value. Instead, the point that we wish to make is that it is nonsensical to suggest that the use of zero as the null value has produced nothing but worthless research. The point nil can be thought of as the midpoint of some interval that (a) includes all values that might be considered trivial and (b) is small enough that calculations using the point nil give a good approximation of calculations based on other values within the interval. This interval is analogous to the "good-enough belt" described by Serlin and Lapsley (1985). Of course, the explicit use of such a belt would be preferable to simply assuming that it provides support for a given hypothesis. Nevertheless, our point is that the conclusions drawn from the vast majority of research that has focused on the nil would

have been very much the same even if an alternative null value had been used.

Additionally, while it may be the case that the nil hypothesis is always false and that Type I errors with respect to nil hypotheses never occur, the same can be said of any hypothesis relating to a specific point in a continuum (Frick, 1995). This fact does not allow one to conclude that a point null hypothesis is a straw man. Rejection of a given null hypothesis implies the rejection not only of the particular null value in question, but also of all of the values in the end of the distribution that is opposite to the end in which the observed value resides. For example, if an observed correlation of .40 is compared with the null value of zero, and NHST leads to the rejection of the hypothesis $\rho = 0$, then it also leads to the rejection of the hypotheses ($\rho = -.01$, $\rho = -.02$, $\rho = -.10$, and so forth on to $\rho = -1$. If the null were instead $\rho = .10$, then rejection of this hypothesis would also entail rejection of hypotheses, relating to .09, .08, and so forth, on to −1. Similar reasoning can be applied to differences between means, or whatever the parameter of interest. Our point is that while the p value specifically applies to a hypothetical distribution based on the null value only, and while this distribution may never, in fact, exist, it cannot be claimed that distributions relating to all values in the opposite direction from the observed result do not exist. Since these more extreme distributions would yield even smaller p values, we are even more justified in rejecting hypotheses relating to the null values associated with these distributions than we are in rejecting hypotheses relating to the null value of direct interest.

This argument is more easily understood in the context of directional hypothesis tests.[4] If a theory suggests a negative correlation between two variables, then one might use a significance test in which only the negative end of the distribution is targeted. If the relevant statistic falls within the rejection region, then the hypothesis associated with the null value is rejected (tentatively). By implication, all of the distributions associated with positive values are also rejected. Indeed, as Meehl (1967) pointed out, there is no reason to suggest that all hypotheses associated with a given half of a distribution are always false. Thus, this significance test is not trivial. The theory places certain constraints on the parameter of interest, that is, positive or negative, and the significance test allows the ruling out of the hypotheses associated with the opposite direction.

The issue is more complicated for nondirectional tests. Because such tests contain some rejection region in both tails of the distribution, it has been argued that an empirical result in one end of the distribution does not imply rejection of all values in the opposite end of the distribution. To shed some light on this topic, let us first observe that nondirectional tests occur in two contexts. In the first, the theory is so weak that it cannot suggest a direction or pattern. Thus, there is no reason to prefer one "direction" over the other, and a nondirectional hypothesis is tested. In the second context, the theory being tested does suggest a particular pattern of relationships, mean differences, and so on, but the analysis technique does not allow for precise consideration of these patterns. It is this context to which our arguments speak. For example, the null hypothesis in a one way analysis of variance (ANOVA) has to do with equality of group means. This null can be rejected if any form of departure from equality occurs. Thus, an ANOVA represents a nondirectional test, even if the theory in question suggests a particular pattern. It is for this reason that tests subsequent to the ANOVA are performed. The NHST associated with the omnibus ANOVA addresses only the issue of equality of means. Thus, it represents a preliminary step in the process of assessing the degree to which data and theory are consistent with one another. The tests subsequent to the ANOVA, which often also include NHSTs, address the more specific questions concerning the pattern of the results. Within the context of these more specific tests, rejection of the null implies rejection of all distributions that are at least as unlikely as the null distribution.

Finally, it should be noted that even if Type I errors for point null hypotheses were technically impossible, it is entirely possible to commit errors that are similar, if not identical to, Type I errors by concluding that trivial departures from the null value justify the conclusion that the null hypothesis should be spurned and the alternative hypothesis adopted. For example, consider the relationship between extroversion and job performance for sales representatives as reported by Barrick and Mount (1991). The meta-analysis–based correlation between these two variables uncorrected

[4] Our purpose is not to endorse or recommend against directional hypothesis tests (see Harris, 1994, for a discussion of the problems associated with one-tailed tests). We mention one-tailed tests only to clarify our point with respect to rejection of sets of hypotheses.

for artifacts was .09. Given the sample size (2,316), statistical significance, namely, H_0: $\rho = 0$, was not an issue. In other words, the nil hypothesis is rejected ipso facto. However, a decision that is, for practical purposes, a Type I error is still possible. If one concludes from this empirical result that extroversion explains a meaningful amount of variance in job performance in spite of the fact that less than 1% appears to be explained, and if a correlation of .10 or less is considered trivial, then this interpretation could easily be construed as an error of some sort. In fact, the authors did conclude that extroversion was a valid predictor of job performance for sales representatives. Thus, even in those cases in which a Type I error in the strict sense is, for all intents and purposes, impossible, it is possible to evaluate conclusions with respect to Type I errors, vis à vis alternative null hypotheses.

What if Significance Tests Are Abandoned?

Various authors (e.g., Cohen, 1994; Schmidt, 1996) have argued that NHST should be outlawed and replaced by parameter estimation procedures such as confidence intervals (CIs).[5] This argument, however, becomes moot with a clear understanding of alpha. Suppose, for example, that we wish to compare a sample, with a mean of 1,180 and a standard error of the mean of 8.17, to a population with a mean of 1,110. In a test of significance, alpha is the probability that the population with the mean of the 1,110 would have produced a sample with a mean of such a size that we would reject the null hypothesis. If one discards tests of significance, alpha is effectively equal to zero because no null hypothesis will ever be rejected.

The width of a CI is determined by $1 - \alpha$ (multiplied by 100 to express it as a percentage). For the example above, this alpha value would be the probability that a population with a mean of 1,180 would produce a sample with a mean of a certain magnitude ($\pm t_\alpha * SE$). In both cases, alpha partially determines which values will fall into one class versus another. For the significance test, alpha partially determines which values are associated with a decision to conclude that the mean of the population from which the sample came is different from 1,110. For the CI, alpha partially determines which values represent means of samples that are likely, or at least not unlikely, to have been produced by a population with a mean of 1,180. In using a confidence interval, we reject, implicitly or explicitly, any hypotheses associated with values that lie outside the interval. It should be noted that this is true regardless of the way that the interval is described. Even if one's focus is on the sampling error-based band about a point parameter estimate (as opposed to focusing on the values that do not lie within the band), the boundaries of this band are defined by the values that lie outside it. The presentation of such a band necessarily implies the exclusion of certain values. Sometimes, this exclusion will be in error, and the probability of such an error is called alpha.

Consider what happens if alpha is set to zero for confidence intervals, as it would be for NHST if NHST were abolished. An alpha value of zero suggests a 100% CI, which would range from minus infinity to plus infinity (from −1 to 1 for correlations). Such CIs, although perfect in the sense of always being correct, are of no use. To make the CI useful in any sense, we must accept an imperfect CI.

Historically, with the concept of probable error, early statisticians used what were basically 50% CIs. That practice was abandoned in favor of more conservative CIs for which failure of the interval to contain the population parameter was considerably less likely than 50%. Fundamentally, there is no way to avoid making a decision regarding alpha, whether that decision involves NHST or CIs. If we set alpha at zero for NHST (refuse to do them), we should also do so for CIs, which shows the underlying fallacy of the replacement of NHST with CIs. As with significance tests, there is a probability of being wrong when forming a CI, and that probability is called alpha.

Likewise, if tests of significance are abolished, power, the probability of rejecting the null hypothesis when it is false, is zero. Power, like CIs, depends on alpha; without alpha, beta is 100% and power is $1 - \beta$. Thus, power equals zero. When these points are added to the fact that CIs and significance tests are based on precisely the same information (i.e., parameter estimates and standard error values), the only reasonable conclusion is that CIs and power estimation cannot be done instead of tests of significance but that instead they should be done in conjunction with significance tests.

[5] There are many statistics—such as goodness-of-fit indices, tests of normality, and tests of randomness—for which confidence intervals are not available (Nantrella, 1972).

Discussion

In conclusion, we would like to reiterate our agreement with many of the positions taken by the authors listed in the first paragraph of this article. Significance testing is abused. Application of Modus Tollens and other rules of syllogistic reasoning to probabilistic statements can lead to problems. Interpretation of p values as the probability of the truth of the null hypothesis given the data is inappropriate. CIs and power should be reported. The points that we wish to make here are these. First, the purpose of data analysis is to allow us to examine the extent to which the data provide corroboration for the theory-based answer to the research question. This corroboration typically comes in the form of disconfirmation of alternative hypotheses and explanations (Popper, 1959). NHST gives us an objective mechanism by which we can rule out hypotheses and explanations relating to the null. Second, the arguments against the use of NHST are built on faulty premises, misleading examples, and misunderstanding of certain critical concepts. We attempt to show that there are many cases in which drawing conclusions about hypotheses based on p values is perfectly reasonable. Indeed, a probabilistic version of the Modus Tollens rule of syllogistic reasoning can be applied to many examples typical of psychology to produce approximate probabilistic statements about hypotheses. Furthermore, a firm understanding of the nature of error rates gives insight into the fact that p values are useful regardless of the actual state of reality. Finally, the position that NHST should be replaced by CIs is nonsensical. The two are based on exactly the same information, and both involve an exclusionary decision of some kind. To criticize and revile one while advocating the other is neither consistent nor rational.

Future of Data Analysis

This is not to say that NHST is appropriate for every situation. It is not. But neither are the alternatives appropriate for every situation, and neither are they to be applied without judgment. Various suggestions have been made with respect to methods that are deemed superior to NHST such as CIs, effect size estimations, and meta-analysis. Each of these methods certainly has its advantages, but they are no less prone to abuse than any other method, and none of them is appropriate for every situation. For example, meta-analysis is not useful for many research questions that have not been addressed empirically in previous studies, and like NHST, it has assumptions that should be considered when interpreting results. Effect size estimation must also be approached with caution, as an effect size estimate is typically the amount of variance in one variable accounted for by another in the sample at hand. The problem is that effect size estimates are dependent on the variabilities of the particular measures and experimental manipulations used in the sample (Cortina and DeShon, 1997; Dooling and Danks, 1975), therefore, the use of different manipulations or measures may result in different effect size estimates. Effect size estimates are helpful but must be interpreted with caution.

Finally, many have recommended CIs as a replacement for NHST. However, confidence intervals and NHST calculations are based on precisely the same information. For example, a 95% CI about the difference between two means and a significance test at .05 both use the difference between sample means, the value from the t distribution corresponding to the degrees of freedom involved, the sample variances, and the sample sizes. The two methods simply present this information in different ways. The result is an emphasis on parameter estimation versus an emphasis on sampling error, with each emphasis having its advantages and disadvantages.

Also, there is no reason to believe that the use of CIs instead of significance tests will change anything. We are all familiar with studies in which CIs are reported, and these studies often point out whether or not the interval includes zero. That is a significance test. It has been argued that this is an inappropriate application of CIs and should be done away with (Schmidt, 1996), thus leaving only the parameter estimate and the width of the band around it. Unfortunately, no alternatives are offered for objectively determining the extent to which the data corroborate the theoretical predictions of the study. Moreover, many of those who would replace significance tests with CIs see no need for alternatives because they believe that no conclusions can be drawn from the single sample study (Schmidt, 1996). For those of us who believe that there is unique merit in the single sample study, the lack of criteria for determining the agreement between theory and data is profoundly disturbing. This lack leaves a yawning gap in the system used to evaluate the merit of most manuscripts submitted for publication.

In any case, the sorts of problems that occur with NHST, CIs, and other statistical procedures are not inherent in these procedures but instead stem from

ignorance of proper applications of these techniques. By replacing NHST with CIs, we actually do harm by giving the illusion that the problems are solved when, in fact, they have not even been addressed.

It is also worth noting that problems associated with the identification of appropriate decision criteria are likely to persist simply because of the types of questions that we ask. In the behavioral sciences, we have traditionally asked yes–no questions, Does A affect B? Does goal setting affect performance? Reason and previous work give us prior bases for expecting certain answers to these questions. If significance tests contribute support for these expectations, then we have attacked the problem, and received corroboration, from both the theoretical and empirical sides. Unfortunately, our theories are rarely precise enough to allow for predictions of parameter values. Thus, only empirical support for a given parameter value is possible. On the other hand, if we keep our questions more general, more conservative, then the possibility of having both theoretical and empirical support exists. As always, we must apply the same care and thought to the interpretation of empirical evidence that we apply to theoretical evidence.

Finally, let us not forget that judgment is required in every analysis of scientific information. The abuses of NHST have come about largely because of a lack of judgment or education with respect to those using the procedure. The cure lies in improving education and, consequently, judgment, not in abolishing the method. Mindless application of any procedure causes problems, and discarding a procedure because it has been misapplied ensures the proverbial loss of both baby and bathwater.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 1–29.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.

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

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

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

Cortina, J. M., & DeShon, R. P. (1997). *Extreme groups vs. observational designs: Issues of appropriateness for the detection of interactions.* Manuscript submitted for publication.

Dellow, E. L. (1970). *Methods of science.* New York: Universe Books.

Dooling, D. J., & Danks, J. H. (1975). Going beyond tests of significance: Is psychology ready? *Bulletin of the Psychonomic Society, 5,* 15–17.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23,* 132–138.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Harris, R. J. (1994). *ANOVA: An analysis of variance primer.* Itasca, IL: F. E. Peacock.

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

Kant, I. (1781). *Critique of pure reason.* Riga, Latvia: Friedrich Hartknoch.

Keynes, J. M. (1921). *A treatise on probability.* London: MacMillan.

Kuhn, T. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lakatos, I. (1978). Falsification and methodology of scientific research programs. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programs: Imre lakatos philosophical papers* (Vol. 1). Cambridge, England: Cambridge University Press.

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

Mill, J. S. (1872). *A system of logic: Ratiocinative and inductive* (8th ed.). London: Longmans, Green, Reader, and Dyer.

Miller, R. W. (1979). Reason and commitment in the social sciences. *Philosophy and Public Affairs, 6,* 241–266.

Miller, R. W. (1987). *Fact and method: Explanation, confirmation, and reality in the natural and the social sciences.* Princeton, NJ: Princeton University Press.

Nantrella, M. G. (1972). The relation between confidence intervals and tests of significance. In R. Kirk (Ed.), *Statistical issues: A reader for the behavioral sciences* (pp. 113–117). Monterey, CA: Brooks/Cole.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for the purposes of statistical inference. *Biometrika, 20A,* 175–263.

Neyman, J., & Pearson, E. S. (1933). On the testing of statistical hypotheses in relation to probabilities a priori.

Proceedings of the Cambridge Philosophical Society, 29, 492.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

Putnam, (1981). *Reason, truth, and history.* Cambridge, MA: Harvard University Press.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

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

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70,* 434–438.

Received July 13, 1996
Revision received November 25, 1996
Accepted December 17, 1996