Chapter 4

There Is a Time and a Place for Significance Testing

Stanley A. Mulaik
Georgia Institute of Technology

Nambury S. Raju
Illinois Institute of Technology

Richard A. Harshman
University of Western Ontario

We expose fallacies in the arguments of critics of null hypothesis significance testing who go too far in arguing that we should abandon significance tests altogether. Beginning with statistics containing sampling or measurement error, significance tests provide prima facie evidence for the validity of statistical hypotheses, which may be overturned by further evidence in practical forms of reasoning involving defeasible or dialogical logics. For example, low power may defeat acceptance of the null hypothesis. On the other hand, we support recommendations to report point estimates and confidence intervals of parameters, and believe that the null hypothesis to be tested should be the value of the parameter given by a theory or prior knowledge. We also use a Wittgensteinian argument to question the coherence of concepts of subjective degree of belief underlying subjective Bayesian alternatives to significance testing.

INTRODUCTION

An accumulating literature (Bakan, 1966; Carver, 1978; Cohen, 1994; Gigerenzer, 1993; Guttman, 1977, 1985; Meehl, 1967, 1978; Oaks, 1986; Pollard, 1993; Rozeboom, 1960; Serlin and Lapsley, 1993; Schmidt, 1992, 1996) has called for a critical reexamination of the common use of "null hypothesis significance testing" (NHST) in psychological and social science research. Most of these articles expose misconceptions about significance testing common among researchers and writers of psychological textbooks on statistics and measurement. But

the criticisms do not stop with misconceptions about significance testing. Others like Meehl (1967) expose the limitations of a statistical practice that focuses only on testing for zero differences between means and zero correlations instead of testing predictions about specific nonzero values for parameters derived from theory or prior experience, as is done in the physical sciences. Still others emphasize that significance tests do not alone convey the information needed to properly evaluate research findings and perform accumulative research. For example, reporting that results are significant at some pre-specified significance level (as in the early Fisher, 1935 or Neyman-Pearson, 1933, significance testing paradigms) or the p level of significance (late Fisher, 1955, 1959) do not indicate the effect size (Glass, 1976; Hays 1963; Hedges, 1981), nor the power of the test (Cohen, 1969, 1977, 1988), nor the crucial parameter estimates that other researchers may use in meta-analytic studies (Rosenthal, 1993; Schmidt, 1996). A common recommendation in these critiques is to report confidence interval estimates of the parameters and effect sizes. This provides data usable in meta-analyses. The confidence interval also provides a rough and easily computed index of power, with narrow intervals indicative of high power and wide intervals of low power (Cohen, 1994). A confidence interval corresponding to a commonly accepted level of significance (e.g. .05) would also provide the information needed to perform a significance test of pre-specified parameter values.

Other than emphasizing a need to properly understand the interpretation of confidence intervals, we have no disagreements with these criticisms and proposals.

But a few of the critics go even further. In this chapter we will look at arguments made by Carver (1978), Cohen (1994), Rozeboom (1960), Schmidt (1992, 1996), and Schmidt and Hunter (chapter 3 of this volume), in favor of not merely recommending the reporting of point estimates of effect sizes and confidence intervals based on them, but of abandoning altogether the use of significance tests in research. Our focus will be principally on Schmidt's (1992, 1996) papers, because they incorporate arguments from earlier papers, especially Carver's (1978), and also carry the argument to its most extreme conclusions. Where appropriate, we will also comment on Schmidt and Hunter's (chapter 3 of this volume) rebuttal of arguments against their position.

Our position with respect to Schmidt (1992, 1996), Schmidt and Hunter (chapter 3 of this volume), and Carver (1978) is that their opposition to significance testing arises out of confusion regarding a number of things: (a) that significance testing is the same as misconceptions held by many researchers about significance testing, (b) that a null hypothesis is necessarily a statistical hypothesis of zero difference, zero effect, or zero correlation, (c) that significance testing is principally concerned with testing a null hypothesis of zero difference, zero effect, or zero correlation, (d) that proponents of significance testing believe sig-

nificance tests are supposed to yield absolute and final determinations of the truth or falsity of a statistical hypothesis, (d) that meta-analysis not only replaces significance testing, but has no need ever of significance tests, (f) that because in small samples significance tests have very little power to detect small effects, they are useless, (g) that because in large samples significance tests have very large power to detect small effects, they will always do so, and thus are useless, (h) that significance tests per se are, or should be, concerned with the accumulation of knowledge, (i) that knowing the power of a test implies that one knows the probability that the hypothesis will be rejected, (j) that confidence intervals around point estimates of parameters should be reported and are not or should not be used for significance testing, (k) that the physical sciences do not use significance tests but instead compute confidence intervals and perform meta-analyses. We answer these criticisms. We place significance testing in the context of seeking to make objective judgments about the world. We also defend significance testing against criticisms raised by others based on the idea that while significance tests concern making a dichotomous decision, that is, the statistical hypothesis is either true or false, we should instead focus on determining how our degrees of belief in the hypothesis are affected by the evidence. We follow-up this essay with a brief appendix on some of the historical controversies in the area of significance testing, for this history has bearing on the current controversy involving significance testing.

PRELIMINARY ARGUMENTS AGAINST SIGNIFICANCE TESTING

Schmidt (1992, 1996) draws heavily on Carver (1978) in focusing his attack on significance testing from the perspective that it "has systematically retarded the growth of cumulative knowledge in psychology" (Schmidt, 1996, p. 115). Schmidt (1996) believed that authors like Carver have "carefully considered all ... arguments [for retaining significance testing] and shown them to be logically flawed and hence false" (p. 116). Our own reading of Carver suggests that "significance testing" for him refers primarily to testing a statistical "null hypothesis" of zero differences between means, zero effect sizes, or zero correlations. He did not consider point hypotheses involving possibly nonzero values for parameters, which is the more general case considered by mathematical statisticians for "significance testing." In fact, most statisticians who do significance testing, regard the "null hypothesis" as simply the hypothesis to be tested or "nullified" (Gigerenzer, 1993). Much of Carver's argument also involved exposing what are actually misconceptions about significance testing, for example, interpreting the p value of the significance level as an unconditioned probability that you would be wrong in accepting the null hypothesis. Criticisms of these

misconceptions are not actually arguments against significance testing properly conceived and are somewhat tangential to the issue of significance testing. (We discuss some of these arguments further on).

Corrupt Scientific Method

Carver (1978) also described how significance testing of the null hypothesis involves a "corrupt scientific method" (p. 387). According to Carver, researchers begin with a research hypothesis about the efficacy of some experimental treatment. Proponents of the "corrupt scientific method" recommend that researchers perform experiments in which differences between experimental and control treatments are compared to differences one would expect under a hypothesis of random measurement and sampling error. A statistical "null hypothesis" is then proposed of no difference between experimental treatments. The null hypothesis is to be rejected and results regarded as "significant" only if a difference as large or larger than some specified amount occurs that would occur only rarely under a hypothesis of chance. "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" (Carver, 1978, p. 387). Carver did not oppose conducting experiments, but giving emphasis to the null hypothesis as opposed to one's research hypothesis. He was troubled by the fact that in small samples one might fail to detect a large, real difference, and yet in very large samples one would almost always reject the null hypothesis, but the effects detected as merely significant might be small or theoretically negligible, so the outcome depends on the sample size. Furthermore, one is almost always guaranteed, Carver felt, of rejecting the null hypothesis with very large samples and he cited Bakan (1966) as indicating that it is unlikely that two groups represent *exactly* the same population with respect to the variable being measured. But Bakan's example seems to concern natural groups being compared rather than groups of experimental units assigned by randomization to experimental treatments, where there are reasonable grounds to expect no differences unless there are effects, either experimental effects or systematic error. Nevertheless, Carver did put his finger on problems in the logic of "null hypothesis significance testing". But his analysis of these problems leaves many issues confounded, especially the issue of significance testing per se versus the issue of the kinds of hypotheses to test and the inferences to be drawn from tests of them.

Criticism of the "Nil" Hypothesis.

One of the problems of "null hypothesis significance testing" is with the null hypothesis of zero difference between means or zero correlation (known as the "nil

hypothesis" (Cohen, 1994)). Meehl (1967) incisively showed how routine and exclusive use of this hypothesis in research prevents progress by inhibiting researchers from formulating and testing hypotheses about specific nonzero values for parameters based on theory, prior knowledge and/or estimates of parameters based on accumulated data, in situations where they have such knowledge and theory to go on. And the logic of "nil hypothesis testing" seems askew, because if a researcher has a theory that a certain treatment has an effect, his theory is supported by rejecting another hypothesis (that there is no effect) rather than by making a successful specific prediction that is within the bounds of measurement error of the observed value. It seems unreasonable to regard as support for a theory that some other hypothesis is rejected in favor of an alternative hypothesis that is so vague in its content ("there is a difference") that it would be compatible with almost any substantive hypothesis predicting almost any size of difference. At best a test of the hypothesis of no difference can provide evidence against the null hypothesis of no difference, no effect, or no correlations. But it provides little evidence for any particular alternative hypothesis.

Meehl (1967) contrasted the "nil hypothesis" approach in the behavioral sciences to hypothesis testing in physics where proponents of theories are required to make specific predictions about a parameter based on theories, and the theories are provisionally accepted only if the outcomes are within measurement error of the predicted value, and no other theories make predictions that also fall within the range of measurement error around the estimate of the parameter. Furthermore, in physics as more and more data accumulate and standard errors of parameter estimates get smaller and smaller, tests of a theory become more and more stringent, because to retain support, predicted values must stay within an increasingly narrower range of measurement error around the estimated parameter as gauged by the standard error. But in "nil hypothesis significance testing" almost any theory that predicts an unspecified nonzero effect will have greater possibilities of being "supported" as measurement precision and sample sizes increase, because the range of measurement and/or sampling error around the zero value of the null hypothesis will get narrower and narrower and the power to detect any small effect increases.

So, one issue concerns the hypotheses to test statistically and whether there are ways to formulate a statistical hypothesis so that it takes into account what is currently known or theorized. There may be times when the no-difference and no-relation hypothesis is appropriate to test, and others when it is not. But, as we shortly argue, the issue of what hypothesis to test is distinct from the issue of significance testing itself, and criticisms of the testing of improper hypotheses should not be taken as criticisms of the concept of significance testing.

Building on Previous Studies in Hypothesis Testing.

It seems that the most appropriate time to test the null hypothesis of no-effect or no-correlation is when one has no prior knowledge or theory of what value to expect and subjects have been assigned at random to experimental treatment conditions so that the expectation would be, failing an experimental effect or systematic error, of no effect. However, once previous studies have been conducted, and the no-effect/no-correlation hypothesis rejected, then there will be estimates of the parameter based on the prior data and these can be used as the hypothesized value for the parameter in a significance test with a new set of data. (Or a hypothesis formulated on the basis of examining a number of prior estimates may be used.) The hypothesis to test concerns the value of the parameter—not that previous samples and a new sample come from populations having equal (but unspecified) population values for the parameter, which is a less informative result, when confirmed.

For example, suppose the estimate of a population mean based on prior data is 50.1. We now construct a confidence interval estimate of the mean in a new sample. Let $\overline{X}$ designate a sample mean based on the new data. Let $\sigma_{\overline{X}}$ designate the standard error of the sample mean. Then assuming further that $\overline{X}$ is normally distributed, a random interval

$$\left[\overline{X} - 1.96\sigma_{\overline{X}}, \overline{X} + 1.96\sigma_{\overline{X}}\right]$$

may be constructed round the sample mean for which

$$P\left(\overline{X} - 1.96\sigma_{\overline{X}} \leq \mu \leq \overline{X} + 1.96\sigma_{\overline{X}}\right) = .95.$$

In other words, the lower and upper limits of this interval are *random variables*, based on the random sample mean plus or minus 1.96 times the standard error (which we presume is known, to simplify our illustration). Such *random intervals* based on the sample means will contain the true population mean in 95% of all samples. Now, if $\overline{X}$ computed from the new data is 49 and $\sigma_{\overline{X}} = 2$, then a 95% level confidence interval is given by

$$\left[49 - 1.96(2), 49 + 1.96(2)\right],$$

which in this case is [45.08, 52.92]. Because 50.1 is contained within this random interval, we provisionally accept the hypothesis that the mean equals 50.1 in the population from which we have drawn the new data. If 50.1 fell outside the sample-based interval, we would provisionally reject the hypothesis that the

population mean is 50.1. This illustrates a use of a confidence interval to perform a significance test of a hypothesis based on previously collected data.

We need to add that instead of using for one's hypothesis previous estimates for the parameter, one may also develop a theory of what the "true" value is and test this theoretical value in new studies. Sometimes there may be several theories and more than one theoretical value to test (as we illustrate later with a case in physics that illustrates significance testing with confidence intervals in choosing between theories.) In some cases, when one theoretical value falls within the confidence interval and the other outside it, one can readily take support for the theory whose value falls within the confidence interval and reject the theory whose value falls without. But sometimes both theories may hypothesize values that fall within the confidence interval. In that case, the proper course may be to suspend judgment, collect more data with tighter standard errors so as to be able to exclude one or the other or both hypothesized values.

Proper Interpretation of a Confidence Interval

It is important to note that it is improper to interpret a specific confidence interval constructed around a sample estimate of a parameter as itself containing the population parameter with the specified probability (Kendall & Stuart, 1979; Neyman, 1941). The specific interval either contains the population parameter or it does not, so it contains the population parameter with a probability of unity or of zero. The probability associated with a confidence interval concerns the class of random intervals so constructed around the sample estimates of the parameter and not any specific interval. In contrast, assuming normality and a known standard error, an interval, constructed to have endpoints at 1.96 standard errors above and below a hypothesized value for the population parameter can be said to have a probability of approximately .95 of containing the sample estimate of the parameter, *if* the hypothesized value is correct. (Kendall & Stuart, 1979; Neyman, 1941).

If theory dictates a specific value for a parameter, then all available data that are independent of the formulation of the theoretical value can be used to estimate the parameter with a confidence interval estimate. If the theoretical value is contained within the confidence interval, that is provisional support for the theory. If not, that is provisional evidence against the theory.

The Purpose of a Significance Test

It is important to realize that in contrast to the issue of what hypothesis to test, significance testing arises because of the presumption that statistical estimates of parameters contain random errors of measurement and/or sampling error. Error in our parameter estimates introduces an element of uncertainty in inferences

about the parameter values from those estimates. A significance test is a way of applying a rational criterion of what values of a test statistic are to be regarded provisionally and defeasibly as inconsistent with and unsupportive of a hypothesized value (or range of values) because they would be too extremely different from the hypothesized value and too improbable under a hypothesis of random error combined with the hypothesized parameter's value. Thus a frequent use for significance testing is distinguishing (provisionally) whether a difference between observed and hypothesized values results from effects of random errors of measurement and/or sampling error. This is what significance testing is principally about. The "statistical significance" of a deviation is a problem that has to be resolved in some way and at some point in a research program, especially if one seeks to evaluate theoretical predictions against data in an efficient and economical way.

The Argument that Meta-Analysis Should Replace Significance Testing

Nevertheless, Schmidt (1992, 1996) built on Carver's (1978) arguments. Being an advocate of the use of meta-analytic procedures, he and his collaborator, John Hunter, ". . . have used meta-analysis methods to show that these traditional data analysis methods [significance testing] militate against the discovery of the underlying regularities and relationships that are a foundation for scientific progress (Hunter & Schmidt, 1990)" (Schmidt, 1996, pp. 115–116). Schmidt (1996) argued that "we must abandon the statistical significance test. In our graduate programs we must teach that for analysis of data from individual studies, the appropriate statistics are point estimates of effect sizes and confidence intervals around these point estimates. And we must teach that for analysis of data from multiple studies, the appropriate method is meta-analysis" (p. 116). Schmidt believed that the development and widespread use of meta-analysis methods "reveals more clearly than ever before the extent to which reliance on significance testing has retarded the growth of cumulative knowledge in psychology" (p. 116).

Schmidt claims that even ". . . these few defenders of significance testing (e.g., Winch and Campbell, 1969) agree that the dominant usages of such tests in data analysis in psychology are misuses and they hold that the role of significance tests in data analysis should be greatly reduced" (Schmidt, 1996, p. 116). (Misuses, however dominant they may be in practice, are nevertheless not evidence against the proper use of significance tests. Whether significance tests will be deemed to have a more limited application than now believed by rank-and-file researchers is certainly a legitimate question we would be willing to entertain. At the same time we think there is a legitimate role to be played by signifi-

cance testing). Schmidt issued the following challenge to statisticians who still believe in significance tests: "Can you articulate even one legitimate contribution that significance testing has made (or makes) to the research enterprise (i.e., any way in which it contributes to the development of cumulative scientific knowledge)? I believe you will not be able to do so" (p. 116).

We feel this challenge stacks the deck against significance testing, because it asks one to cite a contribution for significance testing that significance testing per se is not designed to make and for which it is not directly relevant. Significance testing is not directly concerned with accumulating scientific knowledge. We feel that function is served by the formulation of hypotheses to be tested, which, to lead to accumulating knowledge, should incorporate prior knowledge and theory into the formulation of the hypothesis. In contrast, significance testing per se concerns drawing defeasible inferences from data at hand as to the validity of a statistical hypothesis. A defeasible inference is an inference that ". . . is subject to defeat (nullification, termination, or substantial revision) by further considerations (e.g., later facts or evidence)" (Finnis 1995, p. 181). Significance testing concerns a framework for deciding (provisionally or defeasibly) whether observed results (under presumptions of randomness, sampling, and error of measurement) that differ from hypothesized values are to be treated as consistent with chance error outcomes under the assumption that the hypothesis is true, or to be regarded as so divergent and different as well as improbable under the assumption of the truth of the hypothesis as to provide little or no support for the hypothesis (Fisher, 1959). That's all a significance test provides, no more, no less. It doesn't accumulate anything. That is not its function. There are no accumulative operations intrinsic to a significance test.

On the other hand, significance testing contributes to the cumulative research enterprise in allowing one to assess whether differences from predicted values under an integrative hypothesis are more reasonably regarded as due to random measurement errors and sampling errors or not. For example, suppose you are seeking to accumulate knowledge by synthesizing findings from various studies. At some point the data "at hand" shifts from the data found in individual studies to the data across numerous studies, as in a meta-analysis, where one presumes that outcomes of each study are random and independent of one another along with whatever other assumptions needed to make the meta-analytic inferences. Significance testing comes back into play in deciding whether the data at hand across the studies are consistent with and hypothetically probable under a prespecified statistical hypothesis of interest to the researcher, or so different and so hypothetically improbable under that hypothesis as to cast doubt on the hypothesis by being possibly not a chance result at all. In other words, in statistical studies with probabilistic outcomes there will always be criteria for deciding (defeasibly) whether differences from expectations are to be treated as real dif-

ferences or as due to chance error, and those criteria will represent significance tests. But we do not claim that significance tests encompass all that is important in evaluating statistical hypotheses. We certainly support journals requiring the reporting of confidence interval estimates of parameters and effect sizes because these convey more of the crucial information about the results that may be joined with other findings in developing hypotheses and conducting meta-analyses.

Misconceptions About the Misconceptions About Significance Testing

Our concerns with these essays critical of hypothesis testing is that a major portion of their arguments to do away with significance testing are based on criticisms of abusive misconceptions of the use and interpretation of significance testing by researchers. They can hardly be regarded as criticisms of significance testing properly understood and applied. It is important to note that these critical essays rarely quote critically and accurately Fisher or Neyman and Pearson, the founding fathers of the major schools of significance testing, or mathematical statisticians well trained in their methods. But if one reads these eminent statisticians' writings, he or she often will come across passages in which they are critical of the very same abuses and misconceptions of significance testing that the current crop of critics of significance testing cite as evidence against significance testing. So, if we are to clear the air and get to the heart of their criticisms of significance testing, we need to stipulate what these misconceptions about significance testing are and show how irrelevant they are to the issue of whether to abandon or retain significance tests in our research.

Misconceptions About Significance Testing

Carver (1978), Cohen (1994) and Schmidt (1996) all cite critically variants of the following fallacious misinterpretations of significance testing. We add one or two of our own:

1. *The p value of a significant test statistic is the probability that the research results are due to chance* (Carver 1978). It is hard to imagine how someone would arrive at this conclusion. Perhaps, because the statistic is significant, one wonders how this might occur by chance and then thinks of the statistical hypothesis tested, H_0, as the hypothesis of chance. One reasons that if H_0 is true, a result D as extreme or more extreme than the critical value could occur by chance only with a conditional probability $P(D \mid H_0)$ equal to the significance level. That is, the significance level is always the *conditional* probability of getting a result D as extreme or more

extreme than some critical value *given* the hypothesis H_0 is true. Of course, we never know for certain that it is true, but only reason to this probability hypothetically. But the fallacy is that the statement asserts a different kind of conditional probability. What is the probability that the null hypothesis generated this data, given that we have observed a significant result, that is what is $P(H_0 \mid D)$? But without knowledge of the prior probability for H_0 and each of the various alternative hypotheses, we cannot work this out (using Bayes' theorem). No mathematical statistician would be so naive as to confuse these kinds of probabilities.

2. *The probability of rejecting H_0 is α.* Again the fallacy is to confuse an unconditioned statement of probability with a conditioned statement of probability. α is the conditional probability of rejecting the hypothesis H_0 given H_0 is true—regarded hypothetically. In contrast, without further prior knowledge, we have no idea what the actual probability is of rejecting the null hypothesis at a given significance level. It all depends upon what is the case in the world, and we would not perform a significance test if we knew what the true effect was. Schmidt (1992, 1996) makes a similar error when he imagined scenarios in which the true effect size is .50 and then declared that the probability of making a Type I error is not .05 but zero because, he says, in this scenario the null hypothesis is always false and so there is no error to be made in rejecting the null hypothesis. The only error to be made, he said, is a Type II error, accepting the null hypothesis when it is false. In this scenario Schmidt further computed the probability of rejecting the null hypothesis of no effect to be .37 and not .05. Furthermore, he said the true error rate in making decisions from tests of a hypothesis of no effect in studies against an effect size of .50 is $1-.37=.63$, not .05, as he claimed many researchers believe. Within Schmidt's scenario the actual error rates he considered are correct, but Schmidt failed to see that Type I and Type II error rates are never actual probabilities of making these errors. Having no knowledge about the true effects when setting out to perform a significance test, we have no way of knowing what the true error rates will be. So these error rates are hypothetical probabilities considered conditionally under the case where the null hypothesis is true and under a case where the null hypothesis is false, respectively. These hypothetical error rates are used, for example, in reasoning hypothetically to set a critical value of a significance test and to evaluate the power of the significance test against hypothetical alternatives in establishing what will represent prima facie evidence for or against a null hypothesis. Type I and Type II error rates should never be thought of as unconditional probabilities.

3. *Replicability fallacy: A hypothesis accepted as significant at the α level of significance has the probability of $1 - \alpha$ of being found significant in future replications of the experiment.* Carver (1978) cited Nunnally (1975, p. 195) as asserting this fallacy. The fallacy is to presume that because one has accepted the hypothesis H_0, it is therefore true, and therefore according to the conditional probability distribution of the statistic when H_0 is true, the probability of observing a value of the test statistic again within the region of acceptance will be $1 - \alpha$. But, as Fisher (1935) was careful to point out, accepting the null hypothesis (H_0 for Neyman and Pearson) does not determine that the hypothesis is true. It still might be false. Consequently the rest of the inference fails.
4. *Validity fallacy: A hypothesis accepted as significant at the α level of significance has a probability of $1 - \alpha$ of being true.* This is a gross misinterpretation. As Cohen (1994) points out, a statement about the conditional probability $P(D \mid H_0)$ that a result will fall in the region of rejection (D) given one assumes the hypothesis H_0 is true is not the same as a statement about the probability $P(H_0 \mid D)$ of the hypothesis being true given the result has fallen in the region of rejection D, nor even unconditional probabilities about the truth of the hypothesis $P(H_0)$. As Fisher (1959) stated, the significance level tells nothing about probabilities in the real world. All the significance level tells us is a hypothetical probability that D will occur given the hypothesis H_0 is true, and that is not sufficient to allow us to infer the actual probability of the truth of the hypothesis in a real-world setting.
5. *The size p of the significance level of a result is an index of the importance or size of a difference or relation.* Schmidt (1996) cited this fallacy. An example would be to regard a finding significant at the .05 level to be not as important as a finding significant at the .001 level. The fallacy is to confuse size or magnitude of an effect with the improbability of an effect under the null hypothesis. The p value does not tell you the size or magnitude of an effect. In large samples a p value of .001 may represent a small magnitude effect, which practically speaking may be of negligible importance. On the other hand, it is true that a result significant at the .05 level is not as deviant from the hypothesized value as a result significant at the .001 level, although the p values alone tell you nothing about the difference in magnitude between them. There is also a danger of interpreting the p value as a measure of the improbability of the truth of the null hypothesis and then inferring that results with smaller p values indicate that the null hypothesis is even more improbable. Remember that the p value is the conditional probability of observing a result as deviant or more deviant from

the hypothesized value given that the hypothesized value is the true value. It is not a measure of the probability of the hypothesized value. It is, however, a measure of the plausibility of the hypothesis, because a very small value for p indicates an observed value for the statistic that would be very improbable were the hypothesized value the true value of the parameter. Some of the confusion resides in confusing the logical "improbability" of a hypothesis when evidence quite inconsistent with or improbable according to it is observed—which may have no clear quantitative value—with the "probability" of probability theory that the p value under the null hypothesis represents. In this regard this is the same as the fallacy given in case 1 above.

6. *A statistically significant result is a scientifically significant result.* This fallacy plays on the ambiguity of the word "significant." Knowledgeable statisticians recognize that regarding a result as statistically significant does not imply its size or importance scientifically. It is well known that the standard error of a test statistic varies inversely as the square root of the size of the sample so that in larger and larger samples the power to detect smaller and smaller differences from the hypothesized parameter as "significant" increases. Thus in very large samples a difference significant at the .001 level may still be very small in both absolute terms and in relative terms with respect to the initial variance of the variable. Thus no inference may be drawn as to the size or importance of a result from a knowledge that it is a significant result.
7. *If a result of a test of a hypothesis about a parameter is not significant, then the parameter equals the hypothesized value.* This fallacy is a variant of the fallacy of presuming that if a result is not significant then this means the null hypothesis is true. Schmidt (1996) believed this assumption is the most devastating to the research enterprise. He claimed it prevents researchers who get nonsignificant results with small samples from reporting and pooling their data with data from other studies in meta-analytic studies, which may be able to detect small effects with greater power that were overlooked in the individual studies because of lack of power. We agree that it is reasonable to suggest people should suspend judgment from small-sample studies because they lack power to detect meaningful differences. We agree that it is reasonable to criticize them for presuming without warrant they have made an indefeasible and final judgment in order to get them to seek additional evidence. However, again, none of the original proponents of significance tests, neither Fisher nor Neyman and Pearson would interpret significance tests as determining absolutely the validity of the hypothesized parameter, Fisher least of all. So, again, it is a misinter-

pretation of the results of a significance test. As Fisher (1935) put it, not rejecting the null hypothesis does not mean one has proven the null hypothesis to be true.

8. *The fallacy that a statistical hypothesis is the same as one's theoretical hypothesis.* Frequently, a theoretical prediction will state that a parameter has a certain value. This value is then made the value of a statistical hypothesis to be tested with a significance test. If we observe results that would be too different from the hypothesized value and too improbable according to sampling and/or measurement error under the tested hypothesis, we may be led to reject that hypothesis. But this need not imply that the theoretical hypothesis on which the statistical hypothesis is based is necessarily to be rejected. There may be in the experimental setting experimental artifacts that produce effects different from those anticipated by the theory. So, rejecting the statistical hypothesis may lead to a search for experimental artifacts instead of rejection of the theory.
9. *The argument that nothing is concluded from a significance test.* Schmidt (1996) stated "If the null hypothesis is not rejected, Fisher's position was that nothing could be concluded. But researchers find it hard to go to all the trouble of conducting a study only to conclude that nothing can be concluded" (p. 126). We think this is an unfair reading of Fisher. The issue is what inference is reasonable, although defeasible, to draw from the empirical findings of a study.

A major difference between Fisher and Neyman and Pearson was over the idea that significance testing involves an automated decision-making procedure forcing a researcher to choose one of several predetermined alternative choices. Fisher did not want to lose the freedom to exercise his own judgment as a scientist in whether or not to accept (provisionally) a tested hypothesis on the basis of a significance test. "A test of significance contains no criterion for 'accepting' a hypothesis. According to circumstances it may or may not influence its acceptability" (Fisher, 1959, p. 42). Fisher's attitude seems to reflect, furthermore, many physicists' suspicions of the Neyman-Pearson approach to significance testing, that a decision to accept or reject a theory or hypothesis can be completely encapsulated in the automated significance test. For example, a significant result that runs counter to well-established theory, may not be regarded as evidence against the theory but possibly evidence for an experimental artifact, which the researcher must then isolate. An option for Fisher was to suspend judgment.

A bitter debate between Fisher and Neyman and Pearson followed their (1933) alluding to the relevance of their paradigm to sampling inspection prob-

lems in mass-production industry. In discussing the difference between acceptance decisions in manufacturing and opinions and judgments based on significance tests formed in scientific settings, Fisher (1959) held, "An important difference is that [Acceptance] Decisions are final, while the state of opinion derived from a test of significance is provisional, and capable, not only of confirmation, but of revision" (p. 100). Fisher's belief that Acceptance Decisions might be final seems reasonable, because a decision to return goods to the manufacturer or to stop the assembly line, once implemented, is final. But the point is that from Fisher's point of view a provisional scientific opinion can be formed about the null hypothesis from the results of a significance test, and one may even seek to confirm it with additional evidence, or revise one's opinion on the basis of that additional evidence. What Fisher seemed to sense is that such concepts as "truth" and "falsity" and "logical inference," that work very well in geometry and other areas of mathematics where one presumes one has in the axioms all one needs to arrive at a final decision regarding the truth or falsity of some proposition, do not work very well in science when one is making generalizations or inferences from experience. Our information is incomplete, so our opinions formed from experience will be provisional and defeasible by additional experience. This is different from saying that one forms no conclusions, no opinions at all if a test is not significant. The conclusions, opinions, "decisions" are not final, only provisional. (See Mulaik & James, 1995; Pollock, 1986).

THE NULL HYPOTHESIS IS ALWAYS FALSE?

Cohen (1994), influenced by Meehl (1978), argued that "the nil hypothesis is always false" (p. 1000). Get a large enough sample and you will always reject the null hypothesis. He cites a number of eminent statisticians in support of this view. He quotes Tukey (1991, p. 100) to the effect that there are always differences between experimental treatments—for some decimal places. Cohen cites an unpublished study by Meehl and Lykken in which cross tabulations for 15 Minnesota Multiphasic Personality Inventory (MMPI) items for a sample of 57,000 subjects yielded 105 chi-square tests of association and every one of them was significant, and 96% of them were significant at $p < .000001$ (Cohen, 1994, p. 1000). Cohen cites Meehl (1990) as suggesting that this reflects a "crud factor" in nature. "Everything is related to everything else" to some degree. So, the question is, why do a significance test if you know it will always be significant if the sample is large enough? But if this is an empirical hypothesis, is it not one that is established using significance testing?

But the example may not be an apt demonstration of the principle Cohen sought to establish: It is generally expected that responses to different items responded to by the same subjects are not independently distributed across subjects, so it would not be remarkable to find significant correlations between many such items.

Much more interesting would be to demonstrate systematic and replicable significant treatment effects when subjects are assigned at random to different treatment groups but the *same* treatments are administered to each group. But in this case, small but significant effects in studies with high power that deviate from expectations of no effect when no differences in treatments are administered are routinely treated as systematic experimenter errors, and knowledge of experimental technique is improved by their detection and removal or control. Systematic error and experimental artifact must always be considered a possibility when rejecting the null hypothesis. Nevertheless, do we know a priori that a test will *always* be significant if the sample is large enough? Is the proposition "Every statistical hypothesis is false" an *axiom* that needs no testing? Actually, we believe that to regard this as an axiom would introduce an internal contradiction into statistical reasoning, comparable to arguing that all propositions and descriptions are false. You could not think and reason about the world with such an axiom. So it seems preferable to regard this as some kind of empirical generalization. But no empirical generalization is ever incorrigible and beyond testing. Nevertheless, if indeed there is a phenomenon of nature known as "the crud factor," then it is something we know to be objectively a fact only because of significance tests. Something in the background noise stands out as a signal against that noise, because we have sufficiently powerful tests using huge samples to detect it. At that point it may become a challenge to science to develop a better understanding of what produces it. However, it may turn out to reflect only experimenter artifact. But in any case the hypothesis of a crud factor is not beyond further testing.

The point is that it doesn't matter if the null hypothesis is always judged false at some sample size, as long as we regard this as an empirical phenomenon. What matters is whether *at the sample size we have* we can distinguish observed deviations from our hypothesized values to be sufficiently large and improbable under a hypothesis of chance that we can treat them reasonably but provisionally as not due to chance error. There is no a priori reason to believe that one will always reject the null hypothesis at any given sample size. On the other hand, accepting the null hypothesis does not mean the hypothesized value is true, but rather that the evidence observed is not distinguishable from what we would regard as due to chance if the null hypothesis were true and thus is not sufficient to disprove it. The remaining uncertainty regarding the truth of our null hypothesis is measured by the width of the region of acceptance or a function of the stan-

dard error. And this will be closely related to the power of the test, which also provides us with information about our uncertainty.

The fact that the width of the region of acceptance shrinks with increasing sample size, means we are able to reduce our uncertainty regarding the provisional validity of an accepted null hypothesis with larger samples. In huge samples the issue of uncertainty due to chance looms not as important as it does in small- and moderate-size samples.

THE NEED FOR SIGNIFICANCE TESTS

We cannot get rid of significance tests because they provide us with the criteria by which *provisionally* to distinguish results due to chance variation from results that represent systematic effects in data available to us. As long as we have a conception of how variation in results may be due to chance and regard it as applicable to our experience, we will have a need for significance tests in some form or another.

Schmidt and Hunter (chapter 3 of this volume) ignore the provisional way statisticians treat their decisions based on significance tests in arguing that significance tests do not reveal whether observed differences or relations in a data set are real or "just due to chance." "This objection [against doing away with significance tests] assumes," they say, "that null hypothesis significance testing can perform that feat. Unfortunately, no such method exists—or is even possible." Their argument subtly portrays significance tests as designed to determine absolutely that relations or differences are "real or due to chance." Of course, there can be no such thing. They neglect to point out that Fisher denied that the significance test yields absolute and incorrigible determinations that something is "real or due to chance." Whatever opinions one forms as to the reality or chance basis of a difference or a relation are provisional. Most statisticians, including us, interpret them in this way. So Schmidt and Hunter's argument is simply a misrepresentation of what significance tests provide. What is important to consider is that under the circumstances in which they are employed, where one has no knowledge a priori of the truth of one's statistical hypotheses, significance tests provide a reasonable way of using the data available to arrive at prima facie evidence for the truth or falsity of the statistical hypothesis. Prima facie evidence is sufficient to establish truth or falsity, but conclusions based on it may be disproved or defeated by further evidence or reasoning. Thus we may use the provisional truth or falsity of such hypotheses in forms of defeasible reasoning (Pollock, 1986, 1990), which is the way we reason from experience as opposed to the way we reason in classical logic and mathematics where truths of propositions are presumed inalterably given. In defeasible reasoning truths of

propositions may change with further reasoning or evidence. Truths are provisional.

A FALLACIOUS USE OF POWER IN CRITICISM OF SIGNIFICANCE TESTING

The Meaning of the Power of a Significance Test

Statistical power is analogous to the concept of resolving power in evaluating optical instruments. Power is the hypothetical conditional probability of rejecting the null hypothesis under some alternative hypothesis for the population parameter's value. Power is influenced by three things: (a) the size of the significance level α—increasing α increases power but also increases the probability of rejecting the null hypothesis when it is true; (b) the sample size, which with increasing sample size has the effect of reducing the size of the standard error, thereby increasing power; and (c) the difference between the value of the parameter under the null hypothesis and the value of the parameter under the alternative hypothesis—the larger the difference, the greater the power to detect it. Paradoxically one's confidence in acceptance of the null hypothesis increases with an increase in power to detect a standard difference regarded as important. On the other hand, knowing power is low to detect an effect of a size less extreme than the critical value should temper any enthusiasm for an accepted null hypothesis. We now turn to a fallacious use of the concept of power in criticizing significance testing. Schmidt and Hunter (chapter 3 of this volume) create scenarios, which they imply are realistic, that illustrate the inadequacy of significance tests. They state "The average power of null hypothesis significance tests in typical studies and research literatures is in the .40 to .60 range (Cohen, 1962; 1965, 1988, 1994; Schmidt, 1996; Schmidt, Hunter, & Ury, 1976; Sedlmeier & Gigerenzer, 1989). Suppose we take .50 as a rough average. With a power of .50, half of all tests in a research literature will be nonsignificant." They argue next that supporters of significance tests assume that if a test of a null hypothesis is not significant, it is interpreted to be a zero effect. Schmidt and Hunter then claim that this means that in half of all the studies the conclusion will be that there is no relationship. "Every one of these conclusions will be false. That is, in a research area where there really is a difference or a relation, when the significance test is used to determine whether findings are real or just chance events, the null hypothesis significance test will provide an erroneous answer about 50% of the time."

Our view is that this is fallacious reasoning. We looked up the article by Sedlmeier and Gigerenzer (1989) to see how they computed power for the studies reported in the various journal articles. They noted that they followed Cohen (1962). It is important to realize that power is one of four interrelated quantities: power, significance level, sample size *N*, and effect size. Determine any three of these quantities and you determine the fourth. So to determine the power of a study reported in the literature, you have to specify the sample size, the significance level, and an effect size to be detected. Sample size is reported in a journal article about the study. To standardize comparisons across studies, it suffices to pick an arbitrary fixed significance level .05 and an arbitrary hypothetical effect size. Cohen chose to use three arbitrary hypothetical effect sizes, a small, medium, and large effect size. He chose further to operationalize the three effect sizes as corresponding to the dimensionless Pearson correlations of .20, .40, and .60, respectively. He then computed the power of a study as the power to detect a specified correlation as significant at the .05 level given the sample size *N* of the study. It is extremely important for the reader to see that the effect size is completely hypothetical and does not represent an actual effect present to be detected by the study. No effort was made to find out what the true effect was, and even if such an effort had been made it could have only reported an estimated effect that would be subject to sampling and measurement error. Thus the only thing that varied across studies was sample size *N*, and this *N* was converted into three power figures for the study, the power to detect a small, medium and large effect, respectively, at that sample size.

Sedlmeier and Gigerenzer (1989) followed Cohen's (1962) procedure in determining power for the studies reported a decade or more later than Cohen's studies. The figures Schmidt and Hunter (chapter 3 of this volume) chose to report as typical on the basis of these studies corresponded to the powers to detect as significant a medium effect (correlation of .40) at the study's sample size. On the other hand, although citing the power values Sedlmeier and Gigerenzer (1989) reported, as illustrative of typical psychological studies, Schmidt (1996) compared those power values with the power values of an artificial scenario he constructed of normally distributed data having a true .5 standard deviation effect and a power of .37 to detect that effect against the null hypothesis of no effect at the .05 level. Because he had been discussing an artificial scenario in which the true effect was .5 standard deviations, his comparison with the Sedlmeier and Gigerenzer power figures conveyed the impression that their power figures were also powers to detect the typical true effects in the fields surveyed. By citing Schmidt (1996) and the Cohen (1962) and the Sedlmeier and Gigerenzer (1989) articles together as sources for typical effects, the same impression is created by Schmidt and Hunter (chapter 3 of this volume). And this is borne out because they then use the rough average power of .50 to conclude that in half of

the studies the researcher's conclusion will be that there is no relationship, but this conclusion, they say, will be false in every case. But in our view, to suggest that, in approximately half of the psychological studies surveyed by Cohen and Sedlmeier and Gigerenzer, the null hypothesis was accepted and that in every one of these cases the decision to do so was in error, is wrong and grossly misleading.

The power figures reported by Cohen (1962) and Sedlmeier and Gigerenzer (1989) and used by Schmidt and Hunter (chapter 3 of this volume) are not powers to detect the true effect of the respective studies. Their power figures were only the hypothetical powers to detect an arbitrary medium-sized effect if there were one, given the sample sizes of the studies. We have no idea what the true effects were in those studies. They could have all been much larger than the medium effect on which the powers had been computed, in which case the true power would have been much larger than .50 and the proportion of nonsignificant results would have been lower than .50. Or they could have all been much smaller than the medium effect and the true power would have been much less than .50 and the proportion of nonsignificant results greater than .50. So any attempt to extrapolate to what the typical error rate is in using significance tests in these fields is totally unwarranted. In effect this is the same kind of error of confusing a hypothetical conditional probability with an actual probability of an event happening in the world that critics of significance tests accuse many users of significance tests of making.

On the other hand, it is legitimate to generate hypothetical scenarios in which a true effect is presumed known and then investigate the performance of a significance test as Schmidt (1996) has done. Within the framework of such hypothetical assumptions, Schmidt's conclusions are correct. But the power to detect a true effect varies with the size of the true effect, the sample size, and the significance level. For example, although a given study may have a power of only .5 to detect a medium-size effect, it may have a power greater than .8 to detect a moderately large effect. So it is misleading to generalize these scenarios with medium-size effects to all studies.

But it is important also to remember that significance testing is performed in circumstances where one does *not* have prior knowledge of the size of the true effect nor of the probability of a certain effect size's occurrence. What is important is whether a significance-testing procedure provides a reasonable way of forming a judgment about the validity of a hypothesis about a population parameter from sample data. Significance testing must be judged on those terms. Unlike in Schmidt's scenarios, a typical significance test is performed when one has no prior information about the nature of the effect. If one has such information, it must be incorporated into the hypothesis to be tested.

We urge researchers to specify the value of the parameter to be tested to a value that reflects prior knowledge about it. In experiments with randomization, one knows that there should be no differences—unless there are effects, which one does not yet know. In field studies with correlation it is somewhat problematic to say what prior knowledge dictates H_0 should be. The question then is, "Does the significance test reasonably use the information given to it to guide the researcher to a provisional judgment given one has no idea whether the hypothesis is true or not until one gets the data?"

It is not necessarily a fault of significance testing if in one of these hypothetical scenarios where the true standardized effect size is .5 one accepts the null hypothesis in 74% of the cases and rejects it in only 26%. The question is whether it was reasonable, given what one does not know other than what is given in the data, to arrive at one's decisions in this manner? After all, if the null hypothesis of a zero standardized effect were true, one would reject the null hypothesis in only 5% of the cases, which is much less than 26%. But knowing that the power to detect a small effect of .5 standardized effect units is only 26%, one might be unwilling to put too much stock in such decisions, if one is looking for effects that small. One needs to use knowledge of power to temper the confidence one has in one's decisions if one has any reason to believe the effects to be detected are at most that size.

Other indirect indices of power and corresponding uncertainty associated with an accepted null hypothesis are the standard error of the test statistic, width of the acceptance region, the standardized effect size corresponding to a critical value of the test statistic, and the confidence interval calculated around a point estimate of the effect. Prior calculations of the power to detect an expected effect size can also guide the researcher to obtain the sample sizes to reach a decision with adequate power. But in those cases where one feels one has insufficient power to resolve the issue we have no quarrel with Schmidt (1996) who argued that one can simply report point estimates of the effect size and the confidence interval estimate of the effect. One always has the option to suspend judgment while waiting to obtain sufficient evidence to reach a decision. (Editors need to evaluate articles not on the grounds of statistical significance in studies where power is low against an effect size regarded as important, but of the potential of the data's being useful in combination with data from other studies for meta-analysis). But this does not mean giving up significance testing, only postponing it.

No Need to Abandon Significance Tests

There is no need to abandon significance tests altogether as Schmidt (1996) recommended, especially in those cases where one observes significant effects that

exceed in value effects detectable with high power. For example, a true effect that is four standard errors in size has approximately a power of .975 of being detected by any significance test involving a two-tailed test with a .05 level of significance. And any true effect that is 2.84 standard errors in size may be detected by any such significance test with a power of approximately .80. Of course, in contrast, the observed effect will contain error, and one's remaining uncertainty regarding its true value will be gauged by the standard error or some function of it.

Schmidt (1996) rejected the advice that researchers should calculate sample sizes needed to achieve a specified power against a specified effect. His argument is that this requirement would make it impossible for most studies to ever be conducted. As research progresses within an area sample size requirements become increasingly larger to achieve powers commensurate to detect ever smaller effects as one moves from simply detecting the presence of an effect to determining its specific value or relative size with respect to other effects. In correlational research, he cited how sample sizes may need to be quite large—often 1,000 or more. For example, with a sample of size 1,000 one has the power to detect a correlation of .089 as significantly different from zero at the .05 level with a power of .80. He believed to make these demands on researchers would be unrealistic. But many researchers with commercial and educational tests have access to large data bases today of far more than 1,000 cases. The issue is not whether or not to do a study, for small studies, as Schmidt suggests, can be integrated with other small studies by meta-analyses, but to consider the power of detecting a certain size effect with a significance test with the sample at hand. If one is looking only for large effects, then a significance test can be taken seriously with small samples.

META-ANALYSIS AND SIGNIFICANCE TESTING

Schmidt (1996) believed the issue of power is resolved if one abandons significance testing. Power, he believed, is only relevant in the context of significance testing. But this is not so. Power concerns resolving power, and this issue will remain in any meta-analysis, especially those that investigate the presence of moderator effects and interactions or hypothesizes their nonexistence. To use an analogy, one does not discard an 8× field glass just because it cannot detect objects the size of a house on the moon. One uses it to detect objects within its resolving power. The same is true of a significance test. The point is that one must decide (provisionally) whether deviations from hypothesized values are to be regarded as chance or real deviations and with sufficient power to resolve the issue.

Schmidt and Hunter (chapter 3 of this volume) believe that ". . . no single study contains sufficient information to support a conclusion about the truth or value of a hypothesis. Only by combining findings across multiple studies using meta-analysis can dependable scientific conclusions be reached . . . ." The implication is that significance testing with a single study is thus unable to reach any conclusion about the truth or value of a statistical hypothesis.

We think this argument confuses a number of complex issues. On the one hand, one may consider pooling the sample data from several studies into one large sample to achieve adequate power to test a hypothesis in question. On the other hand, there is the issue of whether one needs to show invariant results across many laboratory settings, which is an issue entirely separate from the issue that significance testing addresses. One may, for example, regard each of the studies as representing a sample from the same population (defined by comparable experimental conditions). A meta-analysis of these studies may take the form of pooling the samples from the individual studies to obtain one large sample that one uses to compute an estimate of the effect. The question will be whether the estimated effect equals some hypothesized effect. Any deviation between the estimated effect and the hypothesized effect will raise the question of whether the deviation is so large and so improbable as to be reasonably (but provisionally) regarded as not due to chance under the hypothesis. Whatever method you use to resolve this question will correspond to a significance test. So, why are we to suppose that this one large meta-analysis allows us to resolve issues about the validity of a hypothesis that no individual study can? Is it simply that individual studies have small samples and insufficient power to resolve the issue raised by the hypothesis? But suppose the individual study has a very large sample with adequate power to detect about any size effect with adequate power. Why is it we cannot reach a (provisional) judgment about the validity and value of a statistical hypothesis from such a single study, just as we do with the single meta-analytic study that has a combined sample size equal to that of the single study? What makes a meta-analysis not itself a single study?

The assertion that one cannot establish the validity and value of a hypothesis in a single study seems to be about other issues than just the issues of sample size and the pooling of samples. One of the values replication of results across many studies conveys is the objectivity of the result. Regardless of the researcher and the researcher's biases, regardless of the laboratory in which the results are observed, regardless of the research equipment used, the same results are obtained. Objectivity just is the demonstration of invariance in what is observed that is independent of the actions and properties of the observer (Mulaik, 1995).

But meta-analysis cannot establish this invariance if it simply pools studies and gets estimates of pooled effects. The resulting estimates may mask a hodgepodge of effects in the various studies. Although careful examination of the re-

ported procedures used to perform the studies may allow one to select studies that appear effectively equivalent in their methodology and research subjects, the possibility remains that some unreported moderating variable had different effects upon the dependent variable in the studies accumulated. If plausible reasons based on prior experience for the possible existence of such a moderating variable can be given, this would undermine the assumption of invariance across the studies. The effect of moderating variables would have to be ruled out with positive evidence to allow one to proceed. Can one detect in the data themselves when a hypothesis of invariance across studies fails? The problem is analogous to an analysis of variance (ANOVA), but more complex, particularly when the parameters evaluated are not means. But even when the problem involves simply mean effects, one may not be able to presume homogeneity of variance within studies, which, with different size samples as commonly occurs in meta-analysis, will make implementation of traditional ANOVA procedures problematic. Nevertheless, an issue that will arise is the power to detect the differences between studies that would undermine the use of pooled estimates of effects across studies. The decision that there are no such differences will involve a form of significance test, if it is driven by the data at all.

Examples of Meta-Analyses

To get a better grasp of the problems of meta-analysis, let us consider the following scenario (Scenario 1) in which multiple samples are drawn from a single population. Scenario 2 will deal with samples drawn from several populations.

Scenario 1: The statistical model underlying this scenario may be written as

$$r_i = \rho + e_i \tag{4.1}$$

where r_i is a sample correlation that we may think of as the population correlation ρ to which sampling error e_i has been added. Other things being equal, if the sample size is small (e.g., $N = 68$), it is known that the statistical test associated with the null hypothesis (of $\rho = 0$) will not have adequate power for detecting nonzero, medium-size population correlations. This power can be brought to an adequate level (say, of .80) either by increasing the individual sample size or by combining data from several samples. In the example given in Schmidt (1996), the bivariate data from the 21 samples, with 68 cases per sample, can be combined in one big sample with $N = 1428$. The correlation based on the bigger sample is .22, which is also the result one would obtain using meta-analytic procedures. In this example, meta-analysis does not really provide any additional useful information. However, the same result as Schmidt's (1996) meta-analysis is obtained from this single-pooled sample using a pooled estimate

of the population correlation and determining that it is significantly different from zero with a significance test. The power of this test to detect a correlation of .20 is greater than .80.

If correlations in the 21 different samples are based on *different but comparable* measures of X and Y, it would not be possible to combine different samples into one big sample for assessing the relationship between X and Y. This could happen if X and Y represent cognitive ability and grade point average (GPA), respectively, and different but comparable measures of cognitive ability are used in the 21 samples. Because different cognitive ability measures are likely to be in different metrics, combining the *raw data* from different samples to compute a single correlation coefficient between X and Y is not advisable. One, however, may be able to average the 21 correlations (as it is done in meta-analysis) to arrive at an overall strength of the relationship between X and Y, because by definition all correlations are on a common metric. Whereas correlations, like the effect sizes, offer the common or standard metric needed in most meta-analyses, Cohen (1994) noted that they "cannot provide useful information on causal strength because they change with the degree of variability of the variables they relate" (p. 1001).

Scenario 2: In this scenario, different samples are drawn from different populations and the underlying statistical model can be expressed as

$$r_{ij} = \rho_j + e_{ij}, \tag{4.2}$$

where subscript j refers to population j and i to a sample drawn from that population. In meta-analysis, or especially in validity generalization studies, the aim is to estimate the mean and variance of the ρ_j (denoted, respectively as μ_ρ and σ_j^2). In this scenario, one can also pool the data from different samples into a single data set and compute the correlation between X and Y. Such an analysis, although adequate in Scenario 1, is inadequate in Scenario 2 because the r_js may differ from population to population. One way to estimate the needed parameters is to first note that, under the assumption that $E(e_{ij}) = 0$,

$$\mu_r = \mu_\rho \tag{4.3}$$

and

$$\sigma_r^2 = \sigma_\rho^2 + E(\sigma_{e_j}^2) \tag{4.4}$$

These two equations can be used to estimate μ_ρ and σ_ρ^2 respectively (Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Estimates of these two parameters

provide useful information in practice, with estimation of σ_ρ^2 receiving more attention when one is concerned with the generalizability of the correlation between X and Y across situations or populations. In the latter application (i.e., in validity generalization studies), Equation 4.4 plays an important role. σ_r^2 on the left-hand side of Equation 4.4 can be estimated by the variance of observed correlations, and let us denote this estimate by s_r^2; the second quantity on the right-hand side of Equation 4.4 can be estimated by the average of sampling-error variances of sample-based correlations denoted by $\overline{s}_e^2$. Then the difference, $\hat{\sigma}_\rho^2 = s_r^2 - \overline{s}_e^2$, can be used as an estimate of σ_ρ^2. Two aspects of this estimate are important in the present context.

First, even when $\sigma_\rho^2 = 0$, the estimate s_r^2 generally will not equal estimate $\overline{s}_e^2$ and, therefore, $\hat{\sigma}_\rho^2$ rarely will equal zero. In explaining the benefits of meta-analysis (and there are several), Schmidt (1996) provides an example (his Figure 3) in which $s_r^2 = \overline{s}_e^2$ to claim that "meta-analysis reaches the correct conclusion" (p. 118). Here, according to Schmidt (1996), the correct conclusion is that $\sigma_\rho^2 = 0$. This illustration is misleading because with a small finite number of samples, s_r^2 rarely equals $\overline{s}_e^2$; one typically has a nonzero residual which must be assessed for its proximity to zero. As an illustration, let us consider the example given by Schmidt in his Table 2. The variance s_r^2 of the 21 observed correlations is .0109 and the average $\overline{s}_e^2$ of the 21 sampling variances is .0133. Therefore $\hat{\sigma}_\rho^2 = s_r^2 - \overline{s}_e^2 = .0109 - .0133 = -.0024$, which is close to zero but not exactly zero. Since variance can only be nonnegative, one may treat this negative difference as zero, which is sometimes done in generalizability theory (Brennan, 1983). What if this residual σ_ρ^2 is small but positive? How does one decide when an estimate is small enough to be considered zero? Would null hypothesis testing be an appropriate tool for deciding whether a residual σ_ρ^2 or an estimate of σ_ρ^2 is significantly different from zero? We think so. Now to the second point:

Second, in validity generalization studies, one is interested in finding out if σ_ρ^2 is zero. This result has important theoretical and practical implications. Therefore, one typically estimates σ_ρ^2 and then tries to determine if that estimate is significantly different from zero. Both Hunter and Schmidt (1990) and Hedges and Olkin (1985) have proposed approximate chi-square tests for this purpose. The aim of these tests is to assess whether the observed correlations or effect sizes are significantly different from each other, a pure and simple case of null hypothesis testing. Hunter and Schmidt (1990) also proposed an ad hoc procedure, which is commonly referred to as the 75% rule. Furthermore, Hunter and Schmidt recommend the use of lower credibility values in inferring the generalizability of validity. These credibility values are derived with the help of estimates of μ_ρ and σ_ρ^2. However, the estimates of μ_ρ and σ_ρ^2 by definition contain sampling error which are reflected in the lower credibility values. Hunter and Schmidt did not take into account the associated sampling errors in estab-

lishing the lower credibility values and therefore their recommended procedure raises questions about its true practical utility. It appears that the straightforward null hypothesis testing of an estimate of σ_ρ^2 is a more defensible route to establishing the generalizability of a validity coefficient across populations.

Along the lines of null hypothesis testing, Hedges and Olkin (1985) have recommended between and within chi-square tests (which are significance tests) for those investigators looking for moderators. Despite the fact that moderators are hard to find in practice (because they are difficult to detect with adequate power when the number of studies compared is small and sample sizes are small), these chi-square tests have been useful in identifying sub-populations in which the validity is found to be generalizable.

SIGNIFICANCE TESTS IN PHYSICS

Schmidt and Hunter (chapter 3 in this volume) argue that physicists do not perform significance tests. What they do in their studies is compute an estimate of the parameter of interest and place an error band or confidence interval around the estimate. To test a theoretical hypothesis they compare the estimated value to the theoretical value. Schmidt and Hunter say specifically ". . . this comparison is not based on a significance test." Furthermore physicists, Schmidt and Hunter say, combine results from different studies in ways that are not essentially different from meta-analyses. "The tests of hypotheses and theories that are considered the most credible are those conducted based on data combined across studies. Estimates from the different studies are averaged, and the standard error of the mean is compared and used to place a confidence interval around the mean estimate. (The confidence interval is not interpreted as a significance test)." They then offer a couple of examples from physics. Of interest to us is their first example, the test of Einstein's theory of relativity which states that gravity of a large massive body like the sun will bend light by a certain amount. They say that

> ...the hypothesis was tested in a famous study that measured the amount of bending in light produced by its passing the sun by comparing the apparent position of stars at the edge of the disk of the sun during an eclipse with their apparent positions when not near the sun. Several different observatories made these measurements and the measurements were averaged. The measured amount of bending corresponded to the figure predicted by Einstein's general theory, and so the hypothesis was confirmed and hence the more general theory from which it was derived was supported. In this important study *no significance tests were used.*

Schmidt and Hunter (chapter 3 of this volume) say further "...no significance test was run to see if the amount of bending was significantly greater than zero (test of the null hypothesis) or to see if the observed amount of bending was significantly different from the amount predicted by the theory (a significance test preferred by some.)"

Although it is true that no test of the nil hypothesis (that the parameter is zero) was performed, it is not quite true to say that no test was performed to see if the observed amount of bending was significantly different from the amount predicted by the theory. As we will see, the measured amount of bending of the light did not correspond to what was predicted by either theory (Newton's or Einstein's), and so some criterion was needed to determine whether the difference between the predicted and measured values was greater than what would be typical as a result of random error of measurement.

Moyer (1979) recounts how in 1916–1917, the British astronomer Arthur S. Eddington published an article in which he showed that Newton's theory of gravity would predict that gravitation would deflect light by one-half the amount predicted by Einstein's theory of relativity. By 1918, Eddington had derived from Einstein's general theory of relativity that a ray of light from a star passing near the edge of the sun would be bent in such a way that the star's image would be shifted outward by $1''.75\ r_0/r$, where r_0 is the radius of the sun and r the closest approach of the star's light to the center of the sun when compared to the star's image without the sun. In contrast Newtonian theory predicted a shift of $0''.87\ r_0/r$. Thus a test of Einstein's and Newton's theories could be made during a total eclipse of the sun, when the disc of the moon would just cover the disc of the sun and stars next to the sun in the field of view could be observed.

In 1919 a total eclipse of the sun was predicted, and Eddington and A. C. D. Crommelin led expeditions to the island of Principe in the Gulf of Guinea, West Africa and to Sobral, northern Brazil, respectively, to observe and photograph the eclipse. In his summary of the experiment, Eddington (1920/1987) noted that Einstein's theory predicts a deflection of $1''.74$ at the edge of the sun, with the amount decreasing inversely as the distance from the sun's center. In contrast Newtonian theory predicts a deflection that is half this, $0''.87$. The final estimates of the deflection (reduced to the edge of the sun) obtained at Sobral and Principe (with their 'probable accidental errors') were: Sobral, $1''.98 \pm 0''.12$; Principe, $1''.61 \pm 0''.30$. Eddington then said, "It is usual to allow a margin of safety of about twice the probable error on either side of the mean. The evidence of the Principe plates is thus just about sufficient to rule out the possibility of the 'half-deflection,' and the Sobral plates exclude it with practical certainty" (p. 245). He then noted that because of the obscuring effects of the clouds, the value of the data obtained at Principe could not be put higher than about one-sixth of that at Sobral. Nevertheless, he felt it was difficult to criticize this confirmation

of Einstein's theory because ". . . it was obtained independently with two different instruments at different places and with different kinds of checks" (p. 245).

A probable error is .67449 of a standard error in a normal distribution (Fisher, 1925). Twice the probable error equals approximately 1.35 standard errors. This corresponds in a two-tailed test to a significance level of .177. So, if Eddington was using a margin of safety of two probable errors on either side of the mean, any hypothesized value outside of the confidence band would be rejected at the .177 level. In this case Newton's prediction of 0″.87 lies outside the two probable error confidence intervals from each site and is evidently rejected by Eddington, in favor of Einstein's prediction of 1″.74, which falls within each band. But if Eddington's probable errors are converted to standard errors by multiplying them by 1/.67449 = 1.4826, we get for Sobral 1.4826 × 0″.12 = 0″.178, and for Principe 1.4826 × 0″.30 = 0″.44. So a confidence interval of two standard errors for Sobral would be [1.624, 2.336], and for Principe [.73, 2.49]. The results obtained at Principe (which was partly covered by clouds during the eclipse) were fewer and of lower quality than those from Sobral, and Eddington gave the Principe results only 1/6 of the weight given the Sobral results. By current standards the results from Sobral clearly supported Einstein's hypothesized value and not Newton's, because Einstein's value of 1″.75 is contained in the interval, and the Newtonian value of 0″.87 is not; the Principe results were equivocal, since the predicted values of each hypothesis, 0″.87 and 1″.75, fell within the two standard-error confidence interval. We believe that Eddington used the confidence bands as a significance test, but by current standards his Type I error was greater than most statisticians would feel comfortable with today, although his power was likely fairly good for the small sample of 28 observations from Sobral because of the large size of the probability of a Type I error.

We have consulted physicists on the Internet regarding their use of significance tests. Evidently they get very little formal training in statistics. So, one reason why they might not perform formal significance tests is that they have not been trained to do so. But they do use confidence bands sometimes in the way many statisticians use confidence intervals as significance tests. If a hypothesized value falls within the confidence interval, that is evidence in favor of the hypothesized value. If it falls outside of the confidence interval, that is evidence against the hypothesized value. Another reason physicists often do not do significance tests is because they are not always testing hypotheses, but rather are trying to improve their estimates of physical constants. Their journals consider reporting estimates to be beneficial to the physics community, because they may be combined with results from other studies. (This supports Schmidt and Hunter's argument in chapter 3 of this volume that one does not always have to perform a significance test to have publishable results, and we concur with this aspect of their argument.) Physicists are also very suspicious of automated deci-

sion-making procedures, and Neyman-Pearson significance testing suggests that to them. They also tend to regard results that differ significantly from predicted values to be most likely due to artifacts and systematic errors. Only after exhaustive efforts to identify the source of systematic errors has failed to turn up anything will they then take seriously such significant differences as providing lack of support for established theory and in favor of some other theory that predicts such results. Finally, as Giere (1987, p. 190) notes, physicists have long realized that the typical experiment produces results with errors of only 2%. But most theories, even good ones in physics, make predictions to only within 20% of the data. Thus if one goes by significance tests alone, one would reject almost all such theories and not pursue them further. So, measures of approximation are often much more meaningful than significance tests in physics. This does not mean there are no situations where they might be used.

Hedges (1987) (cited by Schmidt and Hunter in chapter 3 of this volume as supporting their position) indicates that physicists *do* use procedures that are comparable to significance tests. Although many of their studies are equivalent to meta-analyses in the social sciences, they use a ratio known as Birge's R to evaluate the hypothesis that the population value for a parameter is the same in all studies considered for review. When this ratio is near unity, this is evidence for the consistency of the estimates across the studies; when the ratio is much greater than unity this is evidence for a lack of consistency. Birge's ratio is given as

$$R = \frac{\sum_{i=1}^{k} \omega_i (T_i - T_\bullet)^2}{k-1}$$

where $T_1, \ldots, T_k$ are estimates of a theoretical parameter in each of k studies, $S_1, \ldots, S_k$ their respective standard errors, and $T_\bullet$, their weighted average

$$T_\bullet = \frac{\sum_{i=1}^{k} \omega_i T_i}{\sum_{i=1}^{k} \omega_i},$$

with $\omega_i = 1/S_i^2$.

Hedges (1987) notes that Birge's ratio is directly related to a chi-square statistic

$$\chi^2 = (k-1)R = \sum_{i=1}^{k} \omega_i (T_i - T_\bullet)^2$$

with $k-1$ degrees of freedom. In other words, Birge's ratio is a chi-square statistic divided by its degrees of freedom. The mean of a chi-square distribution equals the degrees of freedom of the distribution, so this ratio compares the size of the obtained chi-square to the mean of the chi-square distribution. The ratio serves to test whether differences among the estimates are greater than what would be expected on the basis of unsystematic (measurement) error.

This chi-square statistic is very similar to comparable chi-square statistics proposed for meta-analysis by Hedges (1981) and Rosenthal and Rubin (1982) in the social sciences to test whether differences in estimates of a parameter in question across studies are greater than what would be expected by sampling error. In this respect, Schmidt and Hunter (1996) are correct in saying that physicists use statistics like those used in meta-analysis. But it is clear that Hedges regards these as significance tests. Thus this is further evidence that physicists have not abandoned significance tests—nor have most meta-analysts.

Hedges (1987) also notes the availability of a comparable approximate chi-square statistic

$$\chi_k^2 = \sum_{i=1}^{k} \omega_i (T_i - \tau)^2$$

for testing whether several studies confirm a theoretically predicted value for a parameter. Instead of the estimate $T_\bullet$, the theoretical value τ of the parameter is used in the formula, and because an unknown value for the parameter is not estimated, one gains a degree of freedom so that the resulting chi-square statistic has k degrees of freedom.

These approximate chi-square statistics used in meta-analysis are appropriate only in large samples ($n_i > 30$) where estimates of the standard errors are stable and the sampling distributions of the parameter estimates are approximately normal.

THE MEANING OF OBJECTIVITY

Objectivity Derived from a Schema of Perception

We have already mentioned the objectivity of significance tests. We would like to bring out why science cannot proceed merely by estimating parameters, as

suggested by Schmidt and Hunter (1996) when they recommend reporting confidence interval estimates of parameters while denying their use in significance tests. Tests of hypotheses are essential to integrating and unifying conceptually the diversity of our observations into concepts of an objective world. The issue is not the uniformity of procedures followed, of clear-cut statements of alternatives and how to decide between them, or how to calculate the results in a uniform way. Uniformity of procedure, it is true, is relevant to establishing objective findings, but the ultimate issue concerns establishing invariant features in the observations that are independent of the actions and properties of the observer.

Mulaik (1995) regarded "objectivity" as a metaphor taken from a schema of perception, the schema involved in the perception of objects. J. J. Gibson (1966, 1979) regarded visual perception of objects to take place in the context of an organism's constantly moving within and interacting with its environment. The organism's motion through the environment produces varying information to the senses about the environment, but the transformations these motions and actions produce, in what is given to the organism perceptually, occur in certain invariant ways that are correlated with those motions and actions. The organism is thus able to factor out the effects of its own actions from the optic array of information presented to it. The term for this is *proprioception.* Objects, on the other hand, for Gibson are invariants through time in the varying optic array that are distinct from invariants of transformations in the optic array produced by the organism. The detection of these objective invariants is known as *exteroception.*

This schema of object perception serves to integrate information gathered almost continuously at adjacent instants in time and points in space into information about objects and acts of the embodied self. Perception of objects and perception of self (as body-in-action) occur simultaneously together and are two aspects of the same process.

Mulaik (1995) argued that when extended conceptually beyond what is given immediately in sensory perception to information collected at widely spaced points in time and space, the schema of objectivity serves as a metaphor to integrate this information conceptually through memory and narration into our objective knowledge about the world and our place within it. It is the driving metaphor of science and even of law. It is also the metaphor that underlies the worlds of virtual reality in computer graphics.

Relevance to Hypothesis Testing

The relevance of objectivity to hypothesis testing is that hypothesis testing is a way of integrating information conceptually into objective forms, of extending them beyond a given situation or context and independent of the observer. The

hypothesis is stated before the data are given. The hypothesis must be formulated independently of the data used to evaluate it. The reason for this is because whatever the data are to reveal must be regarded as independent of the one who formulates the hypothesis. Only then will the data be an objective basis for comparison to the hypothesis and thereby a way of conveying objectivity to the hypothesis, if the data conform to the hypothesis.

In contrast, a hypothesis, so-called, formulated by observing the data to be used in its evaluation and formulated in such a way as to conform to that data, cannot then be regarded as objective by its conforming to those same data. To begin with, more than one hypothesis can be constructed to conform perfectly to a given set of data. Thus any particular hypotheses formulated by a researcher to conform to a set of data may be regarded as relative to the researcher in reflecting the particular context, perspective, biases and even media of representation used by the researcher, imposed onto the data. On the other hand, hypotheses constructed to fit a given set of data logically might not fit an independent data set. Thus fit of a hypothesis to an independent data set can serve as a test of the independent and objective validity of the hypothesis. It is logically possible to fail such tests. But it is logically impossible for data to fail to fit a hypothesis tailored to fit them by the researcher. Thus a test of possible lack of fit cannot even be performed on a hypothesis using data that the hypothesis was constructed to fit, because there is no logical possibility of a lack of fit. On the other hand, most hypotheses are formulated by interacting with data, but not the data used to evaluate them.

We humans do not have to formulate our hypotheses out of nothing. We formulate them from past experience and test them as generalizations with new experience. Or we use one set of data for formulating a hypothesis and another set for testing it. Sometimes in models with many unspecified parameters we can use some aspects of a given data set to complete an incompletely specified hypothesis by estimating the unspecified parameters, and still test the prespecified aspects of the hypothesis with other aspects of the data not used in determining the unspecified parameter estimates (Mulaik, 1990). This conforms very well to the schema whereby we determine perceptually that an object and not an illusion stands before us, by getting new information from a different point of view and comparing it to what we expect to see given what we thought is there initially by hypothesis and according to how that should appear from the new point of view.

Significance testing is concerned with hypothesis testing. The hypothesis-testing aspect of significance testing does concern the integration and accumulation of knowledge, and it is for this reason why formulation of the proper hypothesis to test is crucial to whether or not the test will contribute to the accumulation and synthesis of knowledge. But the particular aspect of hypothesis testing that significance testing is concerned with is whether or not an observed

difference from the hypothesized value is so different and so improbable under the presumption that chance error is combined with the true value, as to cast doubt on the truth of the hypothesized value. In this regard significance testing is blind to the relevance of the hypothesis chosen to test.

We believe that a major problem with nil hypothesis significance testing that brings on the accusation that significance tests prevent the accumulation of knowledge, is that once one has gained some knowledge that contradicts one's initial null hypothesis, one does not modify one's hypotheses to reflect that new knowledge. After rejecting the null hypothesis, a replication study proceeds again to test the same null hypothesis that one now has reason to believe is false. But a hypothesis one ought to test is that the effect is equal to the value estimated in the previous study, which one judged to be significantly different from a zero effect. Or if one does not trust the results of the first study because one believes the effect is an artifact, one should eliminate suspected sources of systematic error in a new experiment and collect new data, with a sample sufficiently large to detect any remaining meaningful effects with sufficient power.

Unfortunately in the context of traditional factorial ANOVA, it is not easy in terms of methods and computer programs for researchers to specify their hypotheses in terms of earlier found effects. It is easier to test the hypothesis that there is no difference between groups and no interaction effects. But we have learned through the development of algorithms for structural equation modeling that one can fix some parameters and free others in specifying a model. Perhaps the time has come to modernize ANOVA and the general linear model, to make it easy to specify and test models with fixed nonzero parameters for certain effects.

DEGREES OF BELIEF?

Rozeboom (1960) criticized null-hypothesis significance testing because it conceptualizes the problem as one in which a decision is to be made between two alternatives. Here his critique is directed more to Neyman-Pearson (1933) rather than Fisherian (1935, 1959) conceptions of significance testing. He argued "But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested" (Rozeboom, 1960, p. 420). Rozeboom objected to the decision-theoretic concept that a motor-act is to be determined by the evidence of an experiment. Decisions are voluntary commitments to action, that is, motor sets. But ". . . acceptance or rejection of a hypothesis," he said, "is a cognitive state which may provide the basis for rational decisions, but is not itself arrived at by such a decision . . ." (p. 420). In other words:

> As scientists, it is our professional obligation to reason from available data to explanations and generalities—i.e., beliefs—which are supported by these data. But belief in (i.e., acceptance of) a proposition is not an all-or-none affair; rather it is a matter of degree, and the extent to which a person believes or accepts a proposition translates pragmatically into the extent to which he is willing to commit himself to the behavioral adjustments prescribed for him by the meaning of that proposition" (pp. 420–21).

Rozeboom seemed inclined to distinguish cognitive from behavioral states, although we think he would have been hard pressed to find criteria for attributing certain cognitive states to individuals without reference to behavioral criteria.

Nevertheless, Rozeboom's (1960) critique of the Neyman-Pearson decision-theoretic approach to significance testing (about which we also have some concerns in the context of evaluating scientific theories and hypotheses by comparing them to specific rival hypotheses) raises issues whose complexities cannot be dealt with in detail in the short space that remains in this chapter. But we can make the following comments: To some extent, when he wrote his critique, Rozeboom was entertaining Bayesian inference as a possible replacement for Neyman-Pearson significance testing. Bayesian inference, however, has long been controversial, particularly when it has been invoked in connection with the use of subjective prior probabilities. Critiques of the philosophical assumptions of Bayesian inference may be found in Giere (1987), Pollock (1986), Glymour (1981), Gillies (1973), and Hacking (1965), whereas a summary of the Bayesian argument and an attempted rebuttal of many of these criticisms from a subjective Bayesian point of view that rejects a behavioral interpretation of degrees of belief was given by Howson and Urbach (1989), which in turn has been criticized by Chihara (1994) and Maher (1996). Another source on this debate was given by the Bayesian Earman (1992) and a subsequent critique by non-Bayesian Forster (1995), who argued that the Bayesian philosophy of science cannot explain the relevance of simplicity and the unification of data via theory to confirmation, induction, and scientific inference. But in joining with the critics, we can add the following comments to the argument, which were suggested to the first author by reading commentaries on the works of the philosopher Ludgwig Wittgenstein (Budd 1989; Schulte, 1992).

The focus on belief as a cognitive state that varies in degree confuses the issue of how to justify beliefs in propositions on the basis of evidence—which is a normative issue concerned with the use of objective criteria of support for a belief—with a psychological theory concerned with describing and measuring a problematic cognitive belief state that presumably is a causal effect of evidence that varies as evidence accumulates, obeys Bayes' theorem, and which in turn is a cause of subsequent behavior. Bayesians attempt to use this implicit psycho-

logical theory, which is at best a hypothesis—and one that does not fit well with the psychological facts about probabilistic beliefs either (Kahneman & Tversky, 1972)—as a normative account of how one is to modify one's degrees of belief on the basis of evidence. But a descriptive theory or hypothesis about psychological processes involving belief states is not a normative framework for justifying beliefs. We justify beliefs in a framework of norms of what constitutes supportive or disconfirming evidence for a proposition. And norms must be understood and applicable in an objective way. In fact, objectivity is a basic norm. As we evaluate the evidence in terms of these norms, our belief states take care of themselves and are not ordinarily explicitly relevant to making the evaluation itself.

The problem with arguments designed to justify beliefs by focusing on private, introspected belief states is that these arguments are based on the same pattern of reasoning used by empiricists to justify our knowledge based on experience. Empiricists sought to ground our knowledge of the world in an incorrigible foundation of inwardly experienced sense data. But the empiricist enterprise sank on the rocks of solipsism and the realization that ultimately the sense data of logically private experience involve a logically incoherent idea that is useless in the public dialogues aimed at justifying beliefs about experience. Not only are the so-called logically private sense data of one individual logically inaccessible to others for the purposes of verification, they are intractable even to the individual who would try to bring them into a kind of private language he or she would use to reason about them. Wittgenstein's (1953) famous "private-language argument" showed the incoherence of the idea of a logically private language based on logically private experience. Because the concept of language involves an activity governed by rules, and rules demand objective criteria for their application, so that one can distinguish between thinking one is following the rules and actually following them and not making them up as one goes along, language cannot be applied to that which is logically and irretrievably private. Wittgenstein's private-language argument knocked the linchpin out of the framework of arguments designed to justify one's beliefs via an incorrigible foundation in subjective, introspected experience. What remains to justify our beliefs are reasons we locate in the world, reasons that are thus public and available to everyone, and judged by rules and norms that we share.

Wittgenstein's (1953) private-language argument rules out the possibility of using language to refer to an inner, logically private "inner state" that reflects an individual's degree of belief. Historically, the concept of "degree of belief" is a metaphor based on the idea of probability that originally developed in connection with games of chance (Hacking, 1975). The concept of probability was extended from objective settings (such as games of chance) with stable well-understood causal constraints built in to the gaming apparatus and clear, objective

criteria of how to assign the probabilities, along with predictive success in use of these probabilities reinforced by the stable physical conditions of the games, through settings in which a probability represented an individual's gut feel as to what might happen next in a poorly understood situation, to degrees of belief in any proposition. The metaphor of life as a series of games of chance in which one constantly assesses the odds extended the idea of probability beyond the settings in which assessing the odds and probability could be done in an objective, rule governed manner to produce numbers assigned to events that obey Kolmogorov's axioms of probability theory. To be sure, numbers can be produced that represent people's degrees of belief, and these may have many of the superficial appearances of probabilities, but they are not probabilities.

Whatever the numbers that Bayesian statisticians elicit from clients as their subjective prior probabilities, they are not under all the constraints needed to guarantee numbers satisfying Kolmogorov's axioms. To begin with, the quantities given must be numbers between zero and unity. Most people can be trained to provide those on request. We would also conjecture that people must have knowledge of the proper use of probabilities in situations where probabilities are objectively determined to use these cases as examples to guide them. This means they must recognize the constraints on the way quantities denoting probabilities are to be distributed over a redefined sample space of alternative possibilities. One is often at a loss to specify the sample space of alternative scientific hypotheses appropriate in a given context, which one needs to do to distribute numerical quantities properly over the possible alternatives to represent probabilities satisfying the axioms of probability theory. Furthermore there is no corresponding concept of some reason for there being a specific value for one's uncertainty about the truth of a scientific hypothesis, as there is a reason for the specific value for the probability of a specific outcome in a game of chance, which is based on the invariant physical properties of the gaming equipment and its configuration. So, even if one can prescribe to a person how in any given situation to define such a space of alternative possibilities, how then do you train them to pick the numbers to assign to the alternatives to represent the individual's subjective degrees of belief about them? How can the subjective Bayesian know that the individual is giving numbers that represent the individual's subjective degree of belief? How can the person him- or herself know he or she is giving the right numbers that represent true subjective degree of belief? What criterion would one use? It is not sufficient to say that whatever the person says is "right" is "right", for in that case we cannot speak of following a rule, which requires being able to distinguish between cases where persons say they are right and their actually being right (Wittgenstein, 1953). Unless there is some objective way to determine this, the task is an impossible one not only for us external to the individual, but for the individual as well, and the numbers given are thus

of dubious value. Unless an individual can be taught rules whose application can be objectively verified as to how to assign degrees of belief in scientific hypotheses that satisfy the axioms of probability, there is no such thing as a subjective degree of belief that an individual can learn to express that corresponds to the probability of probability theory.

But one might argue that we ordinarily do not question the subjective reports of individuals but take them at face value. That is true, but in this case we do not treat their reports as objective truths about themselves, but, as Wittgenstein (1953) suggested, verbal substitutes for observable natural reactions, which we simply accept without judgment as their reactions. There is no requirement that we regard their subjective reactions as right or wrong or descriptive of anything, for there is nothing against which we could compare their expressions to validate them. Nor can the individual privately validate his expressions, because he is in no position to distinguish between being right and thinking he is right, which we are able to do with public phenomena and public criteria. This undermines claims that the use of subjective prior probabilities in Bayesian inference leads to optimal rational inference from experience, for even if the numbers given satisfied axioms of probability in representing some subjective phenomenon, who would be able to tell? Finally the argument fails that subjective Bayesian inference can still be used as an ideal norm prescribing optimal rational inference, as if ideally we could attain subjective prior probabilities satisfying the axioms of probability (Howson & Urbach, 1989), because it is not a norm for humans if humans cannot follow it correctly and objectively.

Recognizing these difficulties for subjective Bayesian prior probabilities, some Bayesians have sought to ground "personal" prior probabilities in behavioral criteria, such as Ramsey's (1931) or de Finetti's (1937) assumption that a degree of belief p in a hypothesis h is equivalent to a disposition to bet indifferently on or against the truth of the hypothesis h at odds $p/(1-p)$, so long as the stakes are kept small (Howson & Urbach, 1989). Ramsey and de Finetti showed that if the degrees of belief did not satisfy the probability axioms and if the one taking your bet could dictate which side of the issue you were to take and the size of the stakes, then you could be made to lose no matter what. Supposedly then one should have degrees of belief that satisfy the probability axioms or one would lose bets in this situation. But the usual arguments against this view point out that there are many good reasons (or none at all) why you might be willing to bet at odds different than those dictated by one's degree of belief (Howson & Urbach, 1989). Behavior in betting situations is influenced by too many things other than degree of belief to serve as a univocal indicator of degree of belief. Besides, no practical Bayesian statistician uses this method to assess personal prior probabilities.

But the strongest argument against "degrees of belief" as propounded by both the subjective and behavioristic personal-probability Bayesians is that their concept of degree of belief confounds evidentiary reasons for belief with non-evidentiary (nonepistemological) reasons, such as the hope one has that one's way of conceiving a situation will turn out to be "right" (whatever that means), which may have no prior evidence to support it. What a theory of justified knowledge requires is evidentiary reasons for one's belief, no more and no less. Subjective and/or personal probabilities are determined by more than what the person knows to be true, and it is impossible to separate in these subjective/personal probabilities what is known from what is hoped for or purely conjectured. And different individuals will have different "subjective," non-evidentiary reasons for their belief; as a consequence, Bayesians believe Bayesian inference will yield different inferences for different individuals, although accumulating data will eventually overwhelm the subjective/personalistic element in these inferences and converge to common solutions. The personal/subjective element enters in primarily at the outset of a series of updated Bayesian inferences. Nevertheless, at the outset, subjective/personal Bayesian inference based on these subjective/personal probabilities does not give what is just the evidentiary reasons to believe in something and is unable to separate in its inference what is subjective from what is objective and evidentiary (Pollock, 1986).

These criticisms of subjective Bayesian inference are not designed to refute the legitimate uses of Bayes' theorem with objectively determinable prior probabilities defined on explicitly defined sample spaces. But the proponents of objective Bayesian inference have not been inclined to regard their method as a universal prescription for inference, but rather as a method limited to situations where objectively determinable prior probabilities are possible. In the meantime that leaves significance testing as another route to probabilistic inference where knowledge of prior probabilities is not available or an incoherent idea.

The point to be made with respect to significance testing, is that significance testing is a procedure contributing to the (provisional) prima facie judgment about the objective, evidentiary validity of a substantive proposition. Subjective opinions and descriptions of subjective states or psychological states are not relevant to such judgments.

APPENDIX

There have been two major schools of thought advocating forms of significance testing, and the distinctions between these schools have not always been recognized. In fact, because of similarities in some of their positions and methods, they generally have been confused by textbook writers on psychological statis-

tics and in the teaching of psychological and social statistics (Gigerenzer, 1989, 1993; Gigerenzer & Murray, 1987). The first, and older of these two schools is due to the eminent statistician R. A. Fisher. The other is due to a successor generation of statisticians, Jerzy Neyman and Egon S. Pearson (the son of Karl Pearson who worked out the formula for the product-moment correlation coefficient that bears his name). Although there are similarities between these two schools, and the later school is in part a logical development from the earlier, a bitter debate between them lasted from the mid-1930's until Fisher died in 1962. Cowles (1989) provided an excellent summary of the development of these two schools and how the debate reflected conflicts between strong-willed, defensive personalities, although there are meaningful differences in emphasis.

Fisherian Significance Testing. R. A. Fisher, known both as a geneticist and a statistician, thought of himself as a research scientist first and statistician second. Yet his mathematical skills were formidable and these allowed him to make numerous major contributions to the mathematics of modern statistics during its formative years. Fisher's approach to significance testing grew out of his rejection of inductive inference based on the concept of inverse probability using prior probabilities (subjective Bayesian inference), an approach advocated by Laplace over 100 years before and still popular. On the one hand, Fisher regarded judgments of prior probabilities for scientific hypotheses to be either too subjective or impossible to formulate to the rigor required of a scientific method, while on the other hand he found the Bayesian argument—that in the absence of any prior knowledge, all hypotheses are to be regarded as being equally probable—to be unconvincing or to lead to mathematical contradictions. Furthermore he did not regard the mathematically well-defined concept of probability to be appropriate for expressing all forms of uncertainty or degrees of belief, which are often based on vague and uncircumscribed grounds. Thus he regarded mathematical probabilities as best limited to objective quantities that could be measured by observed frequencies. Consequently, he sought methods of inductive inference that dealt only with objective quantities and phenomena. In his first position as an agricultural statistician he formulated a system of how to design and draw inferences from experiments (Fisher, 1935). He argued that researchers should always include control conditions among their experimental treatments and should assign experimental treatments at random to experimental units. This would allow one to treat extraneous variable influences on the dependent variable, introduced via the experimental units, as randomized and unrelated to the experimental treatments. The effects of randomized extraneous variation would then in theory cancel one another in deriving expected mean outcomes.

Now a natural hypothesis to be tested, Fisher (1935) held, was that there is no effect from the experimental treatments. This would imply no difference be-

tween the expected means of any of the experimental treatments. This value reflects what you know you have put into the experiment by the process of randomization. You do not know whether there will be an effect or not. Fisher called this natural hypothesis the *null hypothesis*. To those who wondered why one did not test the opposite hypothesis, that there is an effect, he argued *that* is not an exact hypothesis, because no specific value for the effect is set forth by simply saying there will be an effect. By implication it would seem that for Fisher, if one had some specific value for the expected effect, one could use that as one's null hypothesis. But in most experimental situations where the researcher has no specific expected effect in mind, and is uncertain whether there will be an effect at all, the natural exact hypothesis to test is that there is no effect because that is what you would expect from randomization alone.

Fisher (1935) then held that grounds for not believing the null hypothesis would consist in experimental results in the form of statistical values so improbable and so extreme from expected values according to the hypothesized distribution for the statistic under the presumption of the truth of the null hypothesis that to believe these results are consistent with the null hypothesis would strain belief. This gives rise to a significance test. He noted that researchers frequently regard as "significant" extreme results that under the hypothetical distribution of the null hypothesis would be that extreme or more in only 5% of cases. But deciding in this way that a result is significant, implying a lack of support for the null hypothesis is not an irreversible decision. "If we use the term rejection for our attitude to such a hypothesis," he said, it should be clearly understood that no irreversible decision has been taken; that as rational beings, we are prepared to be convinced by future evidence that appearances were deceptive, and that in fact a very remarkable and exceptional coincidence had taken place" (Fisher 1959, p. 35). On the other hand, Fisher held that if one does not obtain a significant result this does not mean the null hypothesis is necessarily true. As he put it, ". . . it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of the experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (Fisher, 1935, p. 19) .

In some respects Fisher's views on hypothesis testing anticipated the views of Karl Popper (1959/1935), the Viennese philosopher of science who argued that science cannot prove hypotheses from evidence but only falsify them. Popper argued that science may proceed with logic and deductive reasoning only insofar as it deduces hypotheses from certain premises and seeks to falsify them (see Mulaik and James, 1995). Fisher's approach to hypothesis testing with significance tests followed this recommendation, but with the realization that Popper overlooked, that even finding a deduced consequence to be false is not sufficient to prove a scientific hypothesis to be false. In this respect Fisher even anticipated

more recent views in epistemology that all inferences from experience are defeasible and reversible with additional evidence. His attitude that there are no final decisions or irreversible conclusions reached in science explains his often negative reactions to other approaches to statistical inference that he perceived (sometimes wrongly) to automate the process of reasoning from evidence and thereby to force the researcher to abide by some final decision imposed by the algorithm.

The other important point about Fisher's views of significance testing was that the significance test does not provide an actual probability for the truth of the hypothesis. As he put it: "In general, tests of significance are based on *hypothetical* probabilities calculated from their null hypothesis. They do not generally lead to any probability statements about the real world, but to a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test" (Fisher, 1959, p. 44). In other words, one can imagine what the probability distribution would be like for a test statistic if the null hypothesis were true. One can then imagine in connection with this hypothetical distribution what the probability would be of obtaining a deviation from the distribution's mean as extreme or more extreme than a certain value. If one selects first a small value for this probability and then finds the corresponding value of the test statistic that would be this extreme or more with this probability, then this value could serve as the critical value of the test statistic for rejecting the null hypothesis. The probabilities involved have nothing to do with real-world probabilities. They are all probabilities in an argument involving counterfactual or subjunctive conditionals (as the logicians would say) as to what sort of things should count in a researcher's mind as evidence against a hypothesis. This point is important, because a frequent criticism of significance testing is that researchers usually believe the probabilities described in connection with a significance test are about *actual* probabilities of making an error when one regards something to be significant. What is criticized in these cases is not Fisher's view of hypothesis testing, but some researchers' misconceptions about it.

It is also important to realize that Fisher's view was that ". . . a significance test of a null hypothesis is only a 'weak' argument. That is, it is applicable only in those cases where we have very little knowledge or none at all. For Fisher significance testing was the most primitive type of argument in a hierarchy of possible statistical analyses and inferences (see Gigerenzer et al., 1989, chapter 3)" (Gigerenzer, 1993, p. 314).

Neyman-Pearson Decision-Theoretic Significance Testing. The other school of significance testing grew out of what its authors initially believed was simply an extension of Fisher's ideas. Beginning in 1928 Jerzy Neyman, a young Polish statistician studying at the University of London, and Egon Pearson, Karl Pearson's son, published a series of articles (Neyman & Pearson 1928, 1933) that had

a major impact on hypothesis testing in the years afterward (Cowles, 1989; Kendall & Stuart, 1979). In their articles Neyman and Pearson argued that the outcome of a significance test should be behavioral, *accepting* or *rejecting* some hypothesis and acting accordingly. Hypotheses, furthermore, are of two kinds, simple and composite. Simple hypotheses specify a unique point in the sample space of the statistic, which represents the set of all possible values that the statistic can take. Composite hypotheses specify a region of points of the sample space. The hypothesis you are to test has to be a well-defined hypothesis. It may be a point hypothesis or a composite hypothesis. To test any hypothesis one first has to divide the sample space into two regions. If the test statistic z falls in one of these regions, one accepts the hypothesis. If it falls in the other region, one rejects the hypothesis. (*Acceptance* and *rejection* of the hypothesis are only provisional actions the researcher takes and do not imply final, irreversible decisions nor determinations that the hypothesis is incorrigibly true or false [Kendall & Stuart 1979, p. 177].)

But to get the best critical region of rejection, you first have to specify a probability α—determined hypothetically according to the distribution you presume the test statistic will have if the hypothesis is true—that you will reject the hypothesis if the test statistic z falls in the critical region of rejection. To determine this critical region, you will also need to specify further what alternative hypothesis is to be considered. This too can be either a simple or a composite hypothesis. Neyman and Pearson in the works cited referred to the initial hypothesis as H_0 and the alternative hypothesis as H_1. (They did not call H_0 the "null hypothesis", but because of the "null" subscript on H_0 and its correspondence to the null hypothesis of Fisherian significance testing, most statisticians continue to call it that. But it is important to realize that "null hypothesis" in this context does not mean the hypothesized parameter value is zero. It can be *any* specific value. It is the hypothesis to be "nullified" (Gigerenzer 1993)). Once the alternative hypothesis has been specified, one can then seek the best critical region (BCR) for a test of the hypothesis. The best critical region is that region of rejection with size α that also would have the largest possible *power* of rejecting the hypothesis if the alternative hypothesis is true.

Power was a new concept introduced by Neyman and Pearson. It refers to the probability that one would accept the alternative hypothesis if it were true given the critical region for rejecting the null hypothesis. Again this is not an actual probability that one will accept the alternative hypothesis, but a hypothetical probability referred to a hypothetical probability distribution set up under the assumption that the alternative hypothesis is true. Power is related to the probability of making one of the two kinds of errors when testing a hypothesis:

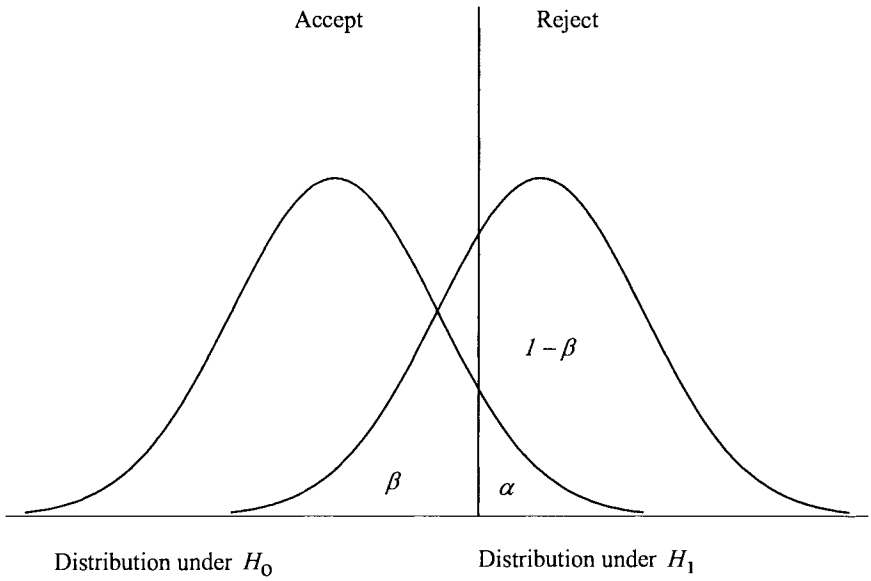

FIGURE 4.1 Power as probability under alternative hypothesis H_1 of rejecting H_0

1. *Type I Error.* Rejecting the null hypothesis when it is true.
2. *Type II Error.* Accepting the null hypothesis when the alternative hypothesis is true.

The conditional probability of making a Type I error when hypothesis H_0 is true is given by α and refers in the case of continuous statistics to the conditional probability of the test statistic's taking a value in the critical region of the statistic calculated according to the sampling distribution of the test statistic under the presumption that the hypothesis H_0 is true. This is also known as the a priori significance level of the test. The value of α is also regarded by Neyman and Pearson as a long-run relative frequency by which you would make a Type I error in repeated samples from the same population under the same significance-testing setup when H_0 is true.

The conditional probability of making a Type II error when the alternative hypothesis H_1 is true is given by β and is the hypothetical probability of the test statistic's falling in the region of acceptance derived from the hypothetical sampling distribution of the test statistic under the assumption that the alternative hypothesis H_1 is true.

Power is the conditional probability of accepting the alternative hypothesis when it is true. Power is given by $1 - \beta$, and refers to the area in the region of re-

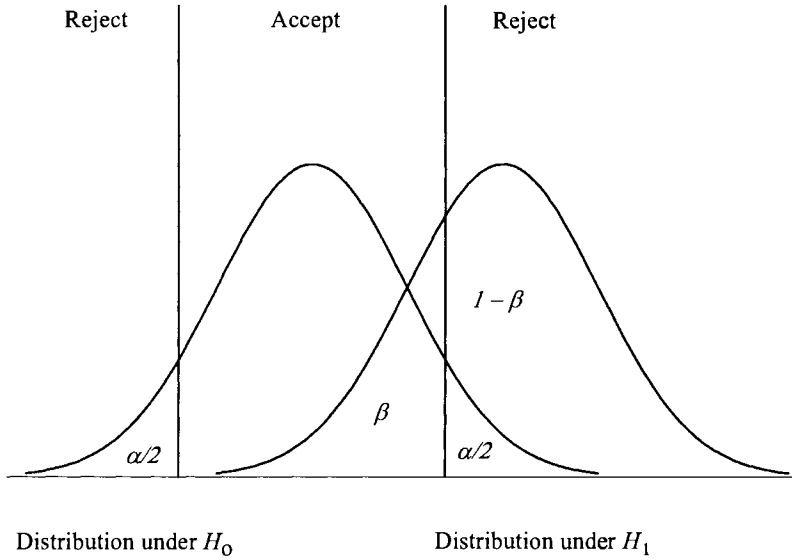

FIGURE 4.2 Power under a two-tailed test.

jection under the hypothetical curve of the distribution of the test statistic considered as if the alternative hypothesis is true.

Finding best critical regions for a given α when both the hypothesis H_0 and H_1 are simple hypotheses is relatively straightforward. The situation is illustrated in Figure 4.1, where H_0 is that the population mean $\mu = \mu_0$ and the alternative hypothesis H_1 is that $\mu = \mu_1$. The test statistic is $z = (\bar{x} - \mu_0)/\sigma$. In this case a one-tailed test is appropriate.

On the other hand, when one tests a simple hypothesis against a composite hypothesis as in the case where one hypothesizes $H_0: \mu = \mu_0$ against a composite, $H_1: \mu \neq \mu_0$, finding the region that maximizes power varies with the specific value of the parameter of the composite set chosen to consider. If one searches for a unique best critical region which is the best in the sense of optimizing power for all values of the parameter under H_1, then one will not find such a region. But a reasonable compromise is to split the value of α in half and locate the region of rejection in the two tails of the distribution, that is, to perform a two-tailed test. This is illustrated in Figure 4.2.

Neyman and Pearson's approach to significance testing focused on what affects power and how the researcher could optimize it when testing a hypothesis against specific alternatives. They noted that power depends on α, and power can be increased by increasing α, or by increasing sample size, or by redefining the critical region of rejection.

Conflicts Between Fisher and Neyman and Pearson

Fisher's (1959) rejection of the Neyman-Pearson approach to significance testing grew out of his perception that the method prescribed a mechanical, automated decision to accept or reject the null hypothesis in a final, irreversible way. Neyman and Pearson used language that equated significance testing with testing for acceptance in manufacturing. This comparison was not appropriate, Fisher held, for significance testing in science. To behave toward a scientific hypothesis the way you behave toward acceptance or rejection of manufactured goods would take away from the researcher the obligation to use his/her independent judgment in making scientific inferences. Acceptance testing involves finite, well-defined populations from which samples can be repeatedly drawn. Acceptance testing also takes into account cost functions. Finally acceptance testing involves making final and irrevocable decisions. One decides irrevocably and finally to accept or reject a lot of manufactured goods on the basis of a sample of them. In science the hypothesized populations have no objective reality but are simply the products of a statistician's imagination. There is no well-defined population from which repeated samples can be drawn. There are also no well-defined cost functions. And no decision about a hypothesis is irrevocable or final.

The one-tailed test of a simple hypothesis against a simple alternative also leads to paradoxes when an observed value of the statistic falls very far to the extreme in the tail opposite the tail where the critical region is located. It seems to strain credibility that something is not wrong with the initial hypothesis H_0, but the observed value does not fall in a region of rejection. Furthermore, in any given application, H_0 may be rejected for some other reason than that H_1 is true. Being forced to chose between two alternatives is unrealistic.

Fisher, it would also seem, regarded a test of a hypothesis to be based on the ordinary view of what in experience invalidates a hypothesis about the value of a parameter in cases where random error of measurement is absent and one has perfect precision of measurement: the observed value differs from the hypothesized value. This difference can either be positive or negative. Furthermore, the larger the difference in absolute magnitude, the more incredible the hypothesis. That random error is added to a true value in obtaining an observed value only changes things insofar as one then has uncertainty as to the true value of the observed parameter. One's uncertainty is less as the observed value becomes more extreme in differing from the hypothesized value, and values that or more extreme when the hypothesized value is correct have quite low probability. This reasoning implies that a two-tailed test of a point hypothesis is the only appropriate test to apply in theoretical scientific work. It corresponds to a test of a hypothesis of the form $H_0: \theta = \theta_0$ against the composite alternative $H_1: \theta \neq \theta_0$. Thus

Fisher could regard Neyman and Pearson's focus on alternate hypotheses and power, to determine an optimal critical region, as irrelevant.

Perhaps Fisher's greatest error was that he refused to consider the import of other uses for the concept of power than determining a region of rejection, even though he had recognized the rudiments of such a concept in his discussion of the effects of sample size on the sensitivity of a test of a person's ability to detect how a batch of tea has been formed when illustrating significance testing (Fisher, 1935, pp. 17–18). Power analysis, after all, is simply an extension of the conditional reasoning implicit in the reasoning used to choose a significance level for a significance test, which Fisher indulged in every time he considered a significance test. If one uses subjunctive reasoning to establish a critical region of acceptance of the null hypothesis with respect to a probability distribution of the test statistic under the assumption that the null hypothesis is true, one uses counterfactual reasoning to consider the probability distribution of the test statistic under the assumption that the true value of the parameter is some other value. Power is the conditional probability of rejecting the null hypothesis given that the population parameter is some specific value other than the value under the null hypothesis. Fisher seems to have been prejudiced against considering the concept of power because it was framed in terms of a decision to be made between two hypotheses, and he was only concerned with evaluating a given statistical hypothesis in terms of the support given to it by data. But the value of the parameter considered counterfactually when evaluating power, need not be a "hypothesis" one is forced to consider along with the value of the null hypothesis and to accept when the original hypothesis is not supported by the data. The alternative value for the parameter merely represents a possible reality against which one evaluates the capacity of the significance test to detect the difference between that value and the value of the null hypothesis. When one obtains data so different from the hypothesized value and improbable under the distribution given the hypothesized value, it may be for any of an infinite number of possible values for the true value of the parameter. One is not thereby able to infer what specific value this is. All one can infer is that the hypothesis is likely not true. Power concerns the resolving power of a significance test. So the concept of power can be incorporated into the framework of Fisherian significance testing, while one rejects the paradigm of choosing between alternative point *hypotheses*.

We have set forth this discussion of these two schools of significance testing so that the reader can compare what is asserted by the critics of significance testing with the positions of those who developed these methods.

REFERENCES

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

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Budd, M. (1992). *Wittgenstein's philosophy of psychology.* London: Routledge.

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

Chihara, C. S. (1994). The Howson-Urbach proofs of Bayesian principles. In E. Eells & B. Skyrms (Eds.) *Probability and conditionals,* pp. 179–199.

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

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (ed.), *Handbook of clinical psychology,* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

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

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

Cowles, M. P. (1989). *Statistics in psychology: a historical perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Earman, J. (1992). *Bayes or bust?* Cambridge, MA: MIT Press.

Eddington, A. S. (1920/1987). The new law of gravitation and the old law. In J. H. Weaver (Ed. and commentator) *The world of physics. Volume II.* New York: Simon and Schuster.

Finnis, J. M. (1995). "defeasible". In T. Honderich (ed.) *The Oxford companion to philosophy* (p. 181). Oxford: Oxford University Press.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, B, 17,* 69–78.

Fisher, R. A. (1959). *Statistical methods and scientific inference.* Edinburgh: Oliver and Boyd.

Forster, M. R. (1995). Bayes and bust: Simplicity as a problem for a probabilistic approach to confirmation. *British Journal for the Philosophy of Science, 46,* 399–424.

Funk & Wagnalls New Encyclopedia. (1994/1995). *Infopedia.* San Diego, CA: Future Vision Multimedia, Inc.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* London: George Allen & Unwin Ltd.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Giere, R. N. (1988). *Explaining Science.* Chicago: University of Chicago Press.

Gigerenzer, G. (1989). *The empire of chance.* Cambridge, England: Cambridge University Press.

Gigerenzer, G. & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, N. J.: L. Erlbaum Associates.

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.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 311–339.

Gillies, D. A. (1973). *An objective theory of probability.* London: Methuen.

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

Glymour, C. (1981). *Theory and evidence.* Chicago: University of Chicago Press.

Glymour, C. (1996). Why I am not a Bayesian. In D. Papineau (ed.) *Philosophy of Science.* Oxford: Oxford University Press pp. 290–313. (Published originally as a part of Glymour, C. (1981). *Theory and evidence.* Chicago: University of Chicago Press).

Guttman, L. B. (1977). What is not what in statistics. *The Statistician, 26,* 81–107.

Guttman, L. B. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1,,* 3–10.

Hacking, I. (1965). *Logic of statistical inference.* Cambridge, England: Cambridge University Press.

Hacking, I. (1975). *The emergence of probability.* Cambridge, England: Cambridge University Press.

Hays, W. L. (1963). *Statistics.* New York: Holt, Rinehart, and Winston.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42,* 443–455.

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

Howson, C., & Urbach, P. M. (1989). *Scientific reasoning: the Bayesian approach.* La Salle: Illinois: Open Court.

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

Kendall, M. & Stuart, A. (1979). *The advanced theory of statistics. Vol. 2. Inference and relationship.* London: Charles Griffin & Co.

Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Maher, P. (1996). Subjective and objective confirmation. *Philosophy of Science, 63,* 149–174.

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

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Moyer, D. F. (1979). Revolution in science: the 1919 eclipse test of general relativity. In A. Perlmutter & L. F. Scott (Eds.) *On the path of Einstein.* New York: Plenum Press.

Mulaik, S. A. (1990, June). An analysis of the conditions under which the estimation of parameters inflates goodness of fit indices as measures of model validity. Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.

Mulaik, S. A. (1995). The metaphoric origins of objectivity, subjectivity, and consciousness in the direct perception of reality. *Philosophy of Science, 62,* 283–303.

Mulaik, S. A., & James, L. R. (1995). Objectivity and Reasoning in Science and Structural Equations Modeling. In R. H. Hoyle (Ed.) *Structural Equation Modeling: Issues and Applications* (pp. 118–137). Beverly Hills, CA: Sage.

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika, 32,* 128–150.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20a,* 175–240, 263–294.

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society, 29,* 492–510.

Nunnally, J. C. (1975). *Introduction to statistics for psychology and education.* New York: McGraw-Hill.

Pollard, P. (1993). How significant is "significance?" In G. Keren & C. Lewis (Eds.) *A handbook for data analysis in the behavioral sciences: methodological issues* (pp. 449–460). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pollock, J. L. (1986). *Contemporary theories of knowledge.* Totowa, NJ: Rowman & Littlefield.

Pollock, J. L. (1990). *Nomic probability and the foundations of induction.* New York: Oxford University Press.

Popper, K. R. (1959/1935). *The logic of scientific discovery.*. (K. R. Popper, Trans.). London: Hutchinson & Co. Ltd. (Original work published 1935).

Rosenthal, R. (1993). Cumulating Evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues* (pp. 519–559). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosenthal, R. & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504.

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

Schmidt, F. L. (1992). What do data really mean? *American Psychologist, 47,* 1173–1181.

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

Schmidt, F. L. & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.

Schulte, J. (1992). *Wittgenstein.* W. H. Brenner & J. F. Holley, translators. Albany, NY: State University of New York Press.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Serlin, R. C. & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues* (pp. 199–228). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wittgenstein, L. (1953). *Philosophical Investigations.* New York: MacMillan.