What if there were no significance tests?

Eight Common But False Objections to the Discontinuation of Significance Testing in the Analysis of Research Data

Frank L. Schmidt
University of Iowa

John Hunter
Michigan State University

ABSTRACT

Logically and conceptually, the use of statistical significance testing in the analysis of research data has been thoroughly discredited. However, reliance on significance testing is strongly embedded in the minds and habits of researchers, and therefore proposals to replace significance testing with point estimates and confidence intervals often encounter strong resistance. This chapter examines eight of the most commonly voiced objections to reform of data analysis practices and shows each of them to be erroneous. The objections are: (a) Without significance tests we would not know whether a finding is real or just due to chance; (b) hypothesis testing would not be possible without significance tests; (c) the problem is not significance tests but failure to develop a tradition of replicating studies; (d) when studies have a large number of relationships, we need significance tests to identify those that are real and not just due to chance; (e) confidence intervals are themselves significance tests; (f) significance testing ensures objectivity in the interpretation of research data; (g) it is the misuse, not the use, of significance testing that is the problem; and (h) it is futile to try to reform data analysis methods, so why try?

Each of these objections is intuitively appealing and plausible but is easily shown to be logically and intellectually bankrupt. The same is true of the almost 80 other objections we have collected. Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution. After decades of unsuccessful efforts, it now appears possible that reform of data analysis procedures will finally succeed. If so, a major impediment to the advance of scientific knowledge will have been removed.

Although the use of statistical significance testing in the analysis of research data is still almost universal, it has now become a controversial practice. Two recent articles (Cohen, 1994; Schmidt, 1996) have argued strongly that the long -standing practice of reliance on significance testing is logically indefensible and retards the research enterprise by making it difficult to develop cumulative knowledge. These articles call for replacing significance tests with point estimates and confidence intervals. Articles of this sort have appeared from time to time since the mid-1950s, but the present situation appears to be different in two important respects. For the first time, the American Psychological Association (APA) Board of Scientific Affairs is looking into the possibility of taking the lead in the effort to reform data analysis procedures. In March of 1996, the Board appointed a task force to study the issue and make recommendations, recommendations that presumably will influence journal editors and reviewers. This development has the potential to effect major changes in data analysis practices.

A second development that makes the current situation different concerns meta-analysis. Over the last 20 years, the use of meta-analysis methods to integrate the findings of research literatures has become increasingly common (Schmidt, 1992). Meta-analysis methods also can be used to present dramatic demonstrations of how reliance on significance testing in data analysis makes it virtually impossible to discern the real meaning of research literatures. Previous arguments against significance testing have been based mostly on logical deficiencies in significance-testing procedures; these abstract logic-based arguments seem to have had only a limited impact. Meta-analysis–based demonstrations seem to be more effective in driving home the concrete distortions created by reliance on significance testing.

Significance testing has been the dominant mode of data analysis in psychology and many other disciplines since the 1930s (Cohen, 1994; Schmidt, 1996). As a result, no practice is more firmly embedded in the minds and habits of researchers than the practice of reliance on significance testing. Not surprisingly, then, the current proposals to replace significance testing with point estimates and confidence intervals have been met with reactions of surprise and shock from some researchers. For some, this proposal seems too radical to be acceptable, and they have advanced specific objections stating why it should not be implemented. For over 2 years, we have been systematically soliciting and collecting such objections. The purpose of this chapter is to present and examine some of these objections. However, only a small number of such objections can be examined here. We have collected more than 80 such objections; those presented in this chapter are only a selected subset of these.

All of the objections examined here are logically deficient and can be shown to be false for that reason. However, it is not possible or desirable to fully explicate all of their deficiencies in this chapter. This chapter is meant to be read in conjunction with Cohen (1994), Schmidt (1996), Carver (1978), Guttman (1985), Rozeboom (1960) and other articles that have critiqued significance testing.

OBJECTION 1

Perhaps the most common objection–and the most deeply rooted psychologically–is the one that states that significance tests are essential because without them we would not know whether a finding is real or just due to chance. The following are three statements of this objection:

> To my way of thinking, statistical significance should precede any discussion of effect size. To talk of effect sizes in the face of results that are not statistically significant does not make sense. In a nutshell, if it is not real, call it zero.
>
> Null hypothesis significance testing does serve a useful purpose. In conducting experiments, the first hypothesis we need to rule out is that the difference or effect is zero. That is, we must, as the first step, rule out the possibility that the result is something that could commonly occur by chance when the null hypothesis that the difference is zero is true.
>
> We need to be able to separate findings that are real from those that are just chance events. How can a researcher know whether two means are (or two correlations) are really different if he or she does not test to see if they are significantly different?

ANSWER TO OBJECTION 1

Of all the objections against discontinuing significance testing that we have collected, this one is the most fascinating to us. It is fascinating not for any technical statistical reasons, but because of what it reveals about the psychology of researchers. Of special significance here is the fact that we have seen this objection raised by researchers who had read (and seemingly understood) Cohen (1994), Schmidt (1996), and other similar articles. The content of these papers demonstrates that significance tests cannot separate real findings from chance findings in research studies. Yet because this objection, worded specifically as it is here, is not explicitly addressed in those articles, this point apparently escapes many readers. We attribute this psychologically interesting fact to the virtual brainwashing in significance testing that all of us have undergone, beginning with our first undergraduate course in psychological statistics.

It would indeed be desirable to have a simple data analysis technique that could reveal whether observed differences or relations in a data set are real or "just due to chance." This objection assumes that null hypothesis significance testing can perform that feat. Unfortunately, no such method exists--or is even possible. Certainly, the null hypothesis statistical significance test cannot achieve this purpose. The facts that have long been known about statistical power make this abundantly clear. 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, 1992; Schmidt, 1996; Schmidt, Hunter, & Urry, 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. The objection under consideration here assumes that if it is not significant, it is zero. (E.g., "In a nutshell, if it is not real, call it zero.") This assumption is not an aberration; it reflects the nearly universal decision rule among researchers that a nonsignificant result is probably just due to chance and should be considered to be zero (Oakes, 1986; Schmidt, 1996). So the conclusion in half of all studies will be that there is no relationship. Every one of these conclusions will be false. That is, in a research area in which there really is a difference or 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. This level of accuracy is so low that it could be achieved just by flipping a (unbiased) coin!

In fact, coin flipping would in many cases provide a higher level of accuracy than the significance test. With coin flipping, the expected error rate is never less than 50%. But in many research domains, the average statistical power is less than .50. One example is the literature on the relation between job satisfaction and absenteeism; there the power is approximately .15. Another example is the literature on the relation between job performance and job satisfaction, which has an average power of about .20. A third example is the literature on the relation between the personality trait of conscientiousness and job performance (Barrick & Mount, 1991), where mean power is less than .30. In these and other research areas, meta-analysis has shown that the relation was indeed always present and was never zero. In all such research areas, significance tests have an error rate (in distinguishing between real effects and chance effects) than is greater than that of a coin flip.

But, one might object, this is not a defect of the significance test per se. This is just a problem of low statistical power. The problem of low power could be solved by requiring all researchers to use large enough sample sizes to ensure high power. Schmidt (1996) showed that this "solution" will not solve the problem. Such a requirement will make it impossible for most studies ever to be conducted. The effect sizes and relations examined in most research are small enough that power of even .80 (which still produces an error rate of 20%) requires more subjects than are often feasible to obtain. They may be unavailable at any cost or be beyond the resources of the researcher to obtain. Furthermore, as the theories tested and compared become more sophisticated over time, the effects sizes studied tend to become smaller (Schmidt, 1996), making the problem worse. So with this requirement, most of the studies in our current research literatures could never have been conducted. Would this be a loss? It certainly would. Such studies contain valuable information when combined with others like them in a meta-analysis. Precise meta-analysis results can be obtained based on studies that all have low statistical power individually. In fact, many of the meta-analyses in our current literature–meta-analyses that are the foundation of our accepted research conclusions–are of this type. All valuable information of this sort would be needlessly lost if a statistical power requirement were imposed on researchers in an ill-advised attempt to salvage the current practice of reliance on significance testing. So imposing a statistical power requirement on researchers will not work as a solution.

The preceding analysis of errors made by the significance test applies only when there really is an effect or relation. What about the case in which the population effect really is zero? If the population effect is zero, then the error rate of the significance test is always the alpha level (typically 5% or 1%). For example, if the true difference between two means in a study using analysis of variance (ANOVA) is zero, and we use an alpha level of .05, the error rate is only 5%. The problem with this objection is that it is rarely if ever the case that the population effect is zero (Carver, 1978; Cohen, 1994; Rozeboom, 1960; Schmidt, 1996). First, there is the fact that, even in the absence of substantive considerations, it is virtually impossible for the null hypothesis to be exactly true (Carver, 1978; Cohen, 1994; Rozeboom, 1960). Hence, it is virtually never true that the error rate is equal to the alpha level. Instead, it is equal to one minus the statistical power to detect the existing deviation of the population effect size from zero.

But there are more substantive considerations that are important here, too. In fact, they are more important. In most research areas, as researchers gain more experience with the phenomenon they are studying, they become more sophisticated in the hypotheses they advance. When research in a given area is first begun, hypotheses may be advanced that are truly false (or as nearly false as they can be)–for example, the hypothesis that raising self-esteem raises IQ. In these cases, the actual population effect is in fact zero–or very nearly so. But as time goes by the hypotheses proposed become more likely to be correct, and hence the corresponding null hypotheses become more and more likely to be false. After a time, virtually all null hypotheses tested are false–not only in a philosophical or logical sense, but (more importantly) in a substantive sense. When the null hypothesis is in fact false, the overall error rate is not the alpha level, but the Type II error rate. That is, when the null is false, the alpha level becomes irrelevant to the error rate of the significance test; it is impossible to falsely conclude that there is a relation when in fact there is a relation. The overall error rate is then one minus the statistical power. There is reason to believe that most research areas today have reached this point (Schmidt, 1996). Hence, it is typically the case that the error rate of the significance test is not 5% but somewhere in the 40% to 60% range, and often higher.

So in light of these considerations, let us reconsider Objection 1. Psychologically, it is easy to understand the desire for a technique that would perform the desirable function of distinguishing in our data sets between relations, differences, and effects that are real and those that are just chance fluctuations. This objection, we believe, is motivated psychologically by precisely this strong desire. But wanting to believe something is true does not make it true. Significance testing cannot perform this function.

OBJECTION 2

This objection holds that without significance testing, researchers would no longer be able to test hypotheses, and we would therefore no longer have a science. This objection has been phrased as follows:

> Discontinuing significance testing means the elimination of hypothesis testing. After all, the purpose of significance testing is to test scientific hypotheses. Nothing is more basic and fundamental to science than the testing of hypotheses; this is the process by which useful theories are developed. You cannot have a science without hypothesis testing. So it makes no sense to call for the discontinuation of significance testing.

ANSWER TO OBJECTION 2

The first aspect of this objection that should be noted is that it equates significance test-based hypothesis testing with <u>scientific hypothesis testing in general</u>. It assumes that there are no other ways to test hypothesis. If these two are one and the same thing, then it follows that discontinuing significance testing means the elimination of hypothesis testing. The fact that many researchers believe that null hypothesis significance testing and hypothesis testing in science in general are one and the same thing is a tribute to the persuasive impact of Fisher's writings (Fisher, 1932, 1935, 1959, 1973). In his writings, Fisher equated null hypothesis significance testing with scientific hypothesis testing.

Let us explore this objection. The physical sciences, such as physics and chemistry, do not use statistical significance testing to test hypotheses or interpret data. In fact, most researchers in the physical sciences regard reliance on significance testing as unscientific. So it is ironic that some psychologists and other behavioral scientists defend the use of significance tests on grounds that such tests are necessary for rigorous tests of scientific hypotheses!

How do researchers in the physical sciences analyze data and test hypotheses? Hedges (1987) found that in individual studies they use procedures that are equivalent to point estimates and confidence intervals. That is, in each individual study an estimate of the quantity of theoretical interest is computed and an "error band" (confidence interval) is placed around this estimate. Hypotheses are tested by comparing the estimated value to the theoretically predicted value; this comparison is not based on a significance test. Hedges also found that in combining findings across studies, researchers in the physical sciences employ methods that are "essentially identical" to meta-analysis. 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 computed and used to place a confidence interval around the mean estimate. (The confidence interval is not interpreted as a significance test.) These are essentially the procedures advocated by Hunter and Schmidt (1990) and Schmidt (1992, 1996). Hence, it is no accident that the physical sciences have not experienced the debilitating problems described, for example, in Schmidt (1996) that are the inevitable consequence of reliance on significance tests.

Let's consider a couple of examples from physics. Einstein's general theory of relativity produced the hypothesis that as light passes a massive body, it would be bent—and it predicted the amount by which it would be bent. That 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.

Consider a second example. Einstein's special theory of relativity predicts that a clock (of any kind—mechanical, electrical, or atomic) that is in an airplane traveling around the world at a certain speed (say, 500 miles per hour) will run slower than exactly the same clock on the ground. Studies were conducted to test this hypothesis derived from the theory, using various kinds of clocks and airplanes. Measurements were averaged across studies. The hypothesis was confirmed. The amount by which the clocks in the airplanes ran slower corresponded to the amount predicted by the theory. Again, no significance tests were used.

The objection we are evaluating here would hold that these studies—and countless others in physics, chemistry, and the other physical sciences—are not really scientific because no significance tests were run on the data. For example, 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). Most people do not find it credible that these studies in physics are not scientific.

Null hypothesis significance testing was popularized by Fisher in the 1930s (e.g., Fisher, 1932). The Neyman—Pearson approach to significance testing was introduced about 10 years later (Neyman, 1962). Hence, this objection implies that prior to the 1930s, no legitimate scientific research was possible. If hypothesis testing is essential to science, and if the only way to test scientific hypotheses is by means of statistical significance tests, then all the physics and chemistry of the 19th century and earlier—including the work of Newton and the development of the periodic table of the elements—was pseudo science, not real science. How credible is such a position?

OBJECTION 3

This objection holds that the problems that researchers have experienced in developing cumulative knowledge stem not from reliance on significance tests, but from failure to develop a tradition of replication of findings:

> Cohen (1994) and Schmidt (1996) have called for abandonment of null hypothesis significance testing, as advanced by R. A. Fisher. But they have both very much misunderstood Fisher's position. Fisher's standard for establishing firm knowledge was not one of statistical significance in a single study but the ability to repeatedly produce results significant at (at least) the .05 level. Fisher held that replication is essential to confidence in the reliability (reproducibility) of a result, as well as to generalizability (external validity). The problems that Cohen and Schmidt point to result not from use of significance testing, but from the our failure to develop a tradition of replication of findings. Our problem is that we have placed little value on replication studies.

ANSWER TO OBJECTION 3

Of all the objections considered in this chapter, this one is perhaps the most ironic. It is not that this objection misstates Fisher's position. Fisher did hold that a single statistically significant rejection of the null hypothesis was not adequate to establish a scientific fact. And Fisher did state that the appropriate foundation for scientific knowledge was replication, specifically the ability to repeatedly produce statistically significant results (Oakes, 1986).

The irony is this. Reproducibility requires high statistical power. Even if all other aspects of a study are carried out in a scientifically impeccable manner, the finding of statistical significance in the original study will not replicate consistently if statistical power is low. As discussed elsewhere in this chapter, statistical power in most psychology literatures averages about .50. Under these circumstances, only 50% of all replications will be successful according to Fisher's criterion of success–rejection of the null hypothesis at the .05 level or less.

Furthermore, the probability of a successful series of replications is the product of statistical power across studies. For example, suppose the requirement is that any statistically signficant finding must be replicated in two additional studies before it is accepted as real. Then, if power in each study is .50, the probability of meeting this replication requirement is (.50)(.50) = .25. So the error rate is now 75%, instead of 50%. Suppose the requirement is five replications. Then the probability of successful replication in all five studies is .50 raised to the 5th power, which is .03. So there is only a 3% chance that an effect that really exists would be concluded to be real, given this replication requirement. And the error rate is 97%; that is, 97% of the time when this replication rule is applied, it will lead to the false conclusion that no effect exits! So much for the value of requiring replication.

We note in passing a curious thing about this objection: It contains a double standard for statistically significant and statistically nonsignificant findings. It requires significant findings to be replicated, but does not require this of nonsignificant findings. That is, in the case of nonsignificant findings, it implicitly holds that we can accept these without need for replication. It assumes that nonsignificant finding will replicate perfectly. But in fact in most real literatures nonsignificant findings do not replicate any better than significant findings. The probability of replication of a nonsignificant finding in a new study is one minus the statistical power in that study. If the statistical power assumes the typical value of .50, the one minus the power is also .50. So the probability of replicating a nonsignificant finding in a second study is .50. The probability of replicating in each of two new studies is (.50)(.50) = .25. And the probability of successful replication of each of five new studies is .03. So in the typical research literature, nonsignificant findings do not replicate any better than significant findings.

Three factors combine to determine statistical power: the size of the underlying population effect, the sample size, and the alpha level. Sample size and alpha level may sometimes be subject to control by the researcher; the population effect usually is not. But long before these factors are considered there is another requirement: One must acknowledge the fact that statistical power is a critical determinant of the success of any effort at replication. Fisher did not take this essential first step. In fact, Fisher rejected the very concept of statistical power and tried to discredit both the concept and the methods for its analysis when these were introduced by Neyman and Pearson in the 1930s and 1940s (Neyman & Pearson, 1933; Oakes, 1986). Fisher based his approach to statistical significance testing solely on the null hypothesis and its rejection. He rejected the concept of any alternative hypothesis that could be the basis for computing power. Hence in Fisherian statistics, there is no concept of statistical power. Therefore, there is no method for determining whether statistical power is high or low, either in a single study or in a series of replication studies. So there is nothing in Fisherian statistics and research methodology to hold power to adequate levels or to prevent statistical power from being quite low. As a result, there is nothing to prevent failures of replication due solely to low statistical power. For example, if replication fails in, say, 50% of the replication attempts, there is no way to discern whether this failure is due to low statistical power or to some other problem.

Hence the irony is that although Fisher made replication the foundation of scientific knowledge, his methods made it impossible to determine whether it was statistically logical or not to expect research findings to replicate. Remember that this objection is specifically a defense of Fisher against Cohen (1994) and Schmidt (1996). Problems of statistical power exist also, of course, with the Neyman–Pearson approach to significance testing. But with that approach, the researcher can at least estimate levels of statistical power. These estimates then allow one to determine whether it is logical from a statistical standpoint to expect the findings to replicate.

Internal contradictions aside, what would be the practical result of accepting this objection at face value? This objection holds that findings that do not repeatedly and consistently replicate must be rejected as not demonstrating any scientific knowledge. Looking at typical research literatures in psychology and other social sciences, we see significance rates for individual hypotheses of around 50%. That is, after the appearance of the first study obtaining significant results, about half of the subsequent replication attempts fail to get significant results. According to the rule advanced in this objection, all such research literatures are completely inconclusive. They demonstrate nothing and establish no scientifically reliable findings.

Is this really the case? Is this conclusion realistic? First, a 50% significance rate is much higher than the 5% rate one would expect if the null hypothesis were true. This fact would appear to cast serious doubt on the Fisherian conclusion. But more revealing are the results of two other analyses that can be performed on such research literatures. A statistical power analysis would reveal an average power of 50% and therefore indicate that, if the effect or relation were present in every study, we should nevertheless expect only 50% of these studies to attain statistical significance—just what we in fact observe. Second, meta-analysis applied to this literature would likely indicate a substantial effect size (or relation) underlying the research literature—and perhaps little real variation around this mean value after controlling for variation produced by sampling error and other artifacts. This is in fact what many meta-analyses to date of such literatures have shown to be the case (Hunter & Schmidt, 1990; Schmidt, 1992).

So we can now contrast the two conclusions. The Fisherian conclusion is that this research literature is inconclusive; it can support no scientific conclusions. The meta-analysis—based conclusion is that the hypothesis tested in the studies making up this literature is strongly supported, and the best estimate of this effect size is, say, $\underline{d}$ = .65 (that is, the treatment effect is 65% of a standard deviation). The Fisherian conclusion is false; the meta-analysis conclusion is correct.

So in conclusion, this attempted defense of Fisherian null hypothesis signficance testing fails on all counts.

OBJECTION 4

This position holds that significance testing has the practical value of making it easier to interpret research findings—one's own and others. It does this by allowing one to eliminate from further consideration all findings that are not statistically significant. The following are three statements of this objection:

> I work in survey research. In that area, there are often hundreds of relationships that can be examined between the various responses and scales. Significance testing is very useful in sifting through these and separating those that are real from those that are just due to chance. Being able to do this has important practical value to me as a researcher. For one thing, it saves a tremendous amount of time.

> I am a great fan of confidence intervals, but for some purposes significance tests are quite useful from a practical point of view. For example, suppose you are doing an ANOVA on a full 2 by 8 by 4 factorial design with all the interactions. While it might be virtuous to look at all the confidence intervals, using significance tests ($\underline{p}$ values) to pick out the interesting effects saves you a lot of time.

> Significance tests are needed to eliminate whole classes of variables that have no effect on, or relation to, dependent variables. Otherwise, it becomes

> impossible to see what is important and what is not, and interpretation of the data becomes an almost impossibly complex task. I know that when I read research reports in journals, I just look for the variables that are significant. If I had to think about all the relationships that were considered, significant or not, I would be driven to distraction.

ANSWER TO OBJECTION 4

We suspect that this objection is one that all of us can identify with. How many of us, when faced in our reading of a research study with several regression equations each containing 20 or more independent variables, have not yielded to the temptation to simplify the task by just singling out for our attention only those variables that are significant? Most of us do this, yet it is not a valid argument for the significance test.

Although it is perhaps not readily apparent, Objection 4 is essentially Objection 1 with a practical, rather than a scientific, rationale. Objection 1 holds that we need to use significance tests to distinguish between chance findings and real findings for scientific reasons. Objection 4 holds that we need to do the same thing, but for practical reasons. That is, Objection 4 holds that significance testing should be used to separate real from chance findings because doing so reduces the amount of time and effort required in interpreting data, both in one's own studies and in studies that one reads in the research literature.

Both objections falsely assume that significance testing can separate real from chance findings. And in both cases, this assumption is based on the further false assumption that findings that are nonsignificant are zero, whereas findings that are significant are real. These false assumptions are explored and refuted in the response to Objection 1; there is no need to repeat this material here. Instead, we want to point out certain features unique to this objection.

First, the "convenience rationale" underlying Objection 4 is the product of an undesirable atheoretical, rawly empirical approach to science; that is, the approach that empirically "mines" scientific data rather than focusing on the testing of specific hypothesis and theories. The practice of scanning a large number of relationships looking for those with low p values has never been successful in contributing to the development of cumulative knowledge. Successful science requires the hard work of theory development followed by focused empirical tests.

But one might object that, although this is true, researchers are sometimes going to use this rawly empirical approach anyway, if for no other reason than the fact that they are often not going to be willing to supply the needed theoretical framework if the author of the study did not provide it to begin with. This leads to our second point: Even under these circumstances, use of significance test results is an inferior way to proceed. That is, even if one deliberately decides to "scan" data in this unprofitable rawly empirical manner, this can better be accomplished by focusing not on *p* values, but on point estimates of effect sizes. Even under these circumstances, it is better to pick out for more detailed attention the largest effect sizes (largest *d* values, correlations, or etas) than the smallest *p* values. It is true that the same problems of capitalization on chance exist in both cases. It is also true that when sample sizes are the same, there is a perfect relation between any index of effect size and its *p* value. However, sample sizes are not always the same, even within the same study; so the largest effect sizes may not be the relations with the smallest *p* values. And in comparing across studies, sample sizes are virtually never the same. But more importantly, focusing on the effect size keeps the actual magnitude of the effects and relations foremost in one's mind. This is important. To the extent that this empirical data-mining procedure ever works at all, it works by singling out unexpected large effects or relations—which can perhaps provide the basis for subsequent hypothesis construction. On the other hand, findings with very low *p* values can reflect underlying effects that are so small as to be of no theoretical or practical interest. If so, this fact is concealed by the *p* value.

So in conclusion, there is never a legitimate role for significance testing, not even in data-mining.

Finally, we would like to use Objections 1 and 4 to point out another fascinating aspect of the psychology of addiction to significance testing. As noted earlier, Objections 1 and 4 are conceptually identical; they are objectively and logically the same objection. Yet to most researchers they do not appear to be identical; in fact, to most researchers in our experience they appear to be not only different objections, but *unrelated* ones as well. We have found this to be a general phenomenon: If they appear under different guises, defenses of significance testing that are conceptually identical appear to most researchers to be different arguments. Thus a researcher who has accepted a telling refutation of one defense will continue to cling to another that is conceptually identical. This creates the apparent necessity on the part of proponents of data analysis reform to address separately each and every possible wording in which defenses of significance testing appear—an impossible task, because the number of different wordings is seemingly limitless. For example, for every objection quoted in this chapter, we have alternative wordings of the same objection that have not been included because of space limitations. (In addition to the space required to list them, it requires space to explain why they are conceptually identical to others already listed and discussed.) Although fully addressed in this chapter, many of these alternately worded statements would probably appear plausible to researchers who have read and accepted the contents of this chapter.

Psychologically, why is this the case? We suspect that the explanation is something like the following. Accepting the proposition that significance testing should be discontinued and replaced by point estimates and confidence intervals entails the difficult effort of changing the beliefs and practices of a lifetime. Naturally such a prospect provokes resistance. Researchers would like to believe it is not so; they would like to believe there is a legitimate rationale for refusing to make such a change. This desire makes it hard for them to see that new defenses of significance testing are conceptually identical to those they have already acknowledged are vacuous. So they accept these arguments as new challenges that have not yet been met by the critics of significance testing. In the meantime, they are relieved of the feeling of any obligation to make changes in their data analysis practices. We and others promoting reform of data analysis methods can only hope that at some point this process will come to an end.

OBJECTION 5

This objection holds that replacing significance testing with point estimates and confidence intervals will <u>not</u> mean discontinuation of significance testing:

> There is less than meets the eye to the change that you are advocating. You and others call for discontinuing use of significance testing, but the confidence intervals that you say should replace the significance test are in fact significance tests! If the lower bound of the confidence interval does not include zero, then the result is statistically significant. If it does, then the result is not significant. So the confidence interval <u>is</u> a significance test. You are throwing significance tests out the front door and then sneaking them back in through the back door!

ANSWER TO OBJECTION 5

This objection is incorrect, but we want to first point out that even if it were true, replacing significance tests with point estimates and confidence intervals would still be an improvement–although a limited one. That is, even if researchers <u>interpreted</u> confidence intervals as significance tests, there would still be benefits from using confidence intervals to conduct the significance tests, because the confidence intervals would reveal to researchers useful facts that are completely concealed by the significance test. First, researchers would see a point estimate of the effect size or relation, so the researcher would have knowledge of the magnitude of the effect. The researcher does not see this information when using only a significance test. Second, the confidence interval would reveal to the researcher the extent of uncertainty in his or her study. For typical studies, the confidence interval would be quite large, revealing that the study contains only very limited information, an important fact concealed by the significance test. Likewise, the researcher would see that a wide range of estimates of the effect size are plausible--many of which if realized in a study would be statistically significant. Finally, consideration of this information revealed by the confidence interval would likely temper and reduce the emphasis placed by the researcher on the outcome of the significance test. That is, researchers might learn to take the significance test less seriously. So there would be benefits even if researchers interpreted confidence intervals as significance tests. But they need not and should not.

The assumption underlying this objection is that because confidence intervals can be interpreted as significance tests, they must be so interpreted. But this is a false assumption. Long before the significance tests was ever proposed, confidence intervals were being used and were interpreted as "error bands" around point estimates. In fact, confidence intervals have a very long history, much longer than that of significance tests. Significance tests were advanced (mostly by Ronald Fisher) in the 1930s, but confidence intervals were advanced, advocated, and applied by some of the earliest contributors to statistics, including Bernoulli, Poisson, and others in the 1700s. In the 20th century, prior to the appearance of Fisher's 1932 and 1935 texts, data analysis in individual studies typically was conducted using point estimates and confidence intervals (Oakes, 1986). The confidence interval that was usually presented along with the point estimate was the "probable error"--the 50% confidence interval. Significance testing generally was not employed, and the confidence intervals presented were not interpreted as significance tests.

In fact, the probable error confidence interval is perhaps the best reply to the proposition that confidence intervals must be interpreted as significance tests. Individuals maintaining that confidence intervals must be interpreted as significance tests have in mind 95% or 99% confidence intervals. These are indeed the two most frequently used confidence intervals today. But confidence intervals of any width can be used. For example, many statistics books today still discuss and present the 68% confidence interval--which extends one standard error above and below the point estimate. The 68% confidence interval is just as correct and legitimate as the 99% confidence interval. So is the 50% probable error confidence interval. It seems clear that those who advance Objection 5 would not be willing to interpret these confidence intervals as significance tests. Perhaps they would revise their objection to say that not all confidence intervals must be interpreted as significance tests, but only 95% and wider confidence intervals. Obviously, this is a purely arbitrary position and could not salvage the original objection. Hence it seems clear that no confidence interval must be interpreted as a significance test.

In responding to this objection, it is also revealing to look at the use of confidence intervals in the physical sciences, such as physics and chemistry. In the physical sciences, confidence intervals are used but not significance tests. (See the discussion in the response to Objection 2.) So the physical sciences provide us with a model of successful research endeavors that have employed confidence intervals but have not interpreted those confidence intervals as significance tests. Would those who advance Objection 5 maintain that data analysis practices in the physical sciences are erroneous and unscientific? To be logically consistent, they would have to.

Although it is clear that confidence intervals need not (and, we argue, should not) be interpreted as significance tests, it does seem to be the case that many feel a strong compulsion to so interpret them. What is the basis for this compulsion? We believe it is a strongly felt need for an objective, mechanical procedure for making a dichotomous decision in analyzing study data. That is, it is based on the belief that the researcher must make a dichotomous decision; the researcher must conclude either that the study results support the hypothesis or that they do not. The traditional use of significance testing provides a procedure for doing this. Because this dichotomous decision must be made, this belief holds, an objective procedure for doing this is essential and must be retained—even if confidence intervals are used in place of significance tests.

But in fact no such dichotomous decision need be made in any individual study. Indeed, it is futile to do so, because no single individual study contains sufficient information to support a final conclusion about the truth or value of an hypothesis. Only by combining findings across multiple studies using meta-analysis can dependable scientific conclusions be reached (Hunter & Schmidt, 1990; Schmidt, 1992, 1996). From the point of view of the goal of optimally advancing the cumulation of scientific knowledge, it is best for individual researchers to present point estimates and confidence intervals and refrain from attempting to draw final conclusions about research hypotheses. These will emerge from later meta-analyses.

However, although this is the situation objectively, it is not one that many researchers feel comfortable with. In many cases, the motivation that moves the primary researcher to undertake and complete an arduous and time-consuming primary study is the belief that his or her study will answer the relevant scientific question. We suspect that it is usually not the case that the motivation is to contribute a study to a large group of studies that will be included in a meta-analysis. This feeling on the part of primary researchers is understandable (and we have experienced it ourselves), but the belief that one's individual study will answer a scientific question will not make it true, no matter how understandable the belief.

Objectively, this belief has never been true; and even in the routine practice of the research enterprise, it has never been accepted as true. Long before the advent and acceptance of meta-analysis, research literatures accumulated on many research hypotheses. When these literatures were interpreted using traditional narrative review methods, no single study was ever viewed by the reviewer as having itself alone resolved the scientific question. Hence it is not the case, as maintained by some, that the ascendance of meta-analysis has reduced the status and probative value of the individual primary study. That process took place much earlier; in every research area it occurred as soon as a substantial research literature accumulated on a particular scientific question.

One last note: Some have interpreted our position that single studies cannot be the basis for final conclusions as forbidding researchers from discussing and interpreting the findings in their individual studies. This is not our position. Results in individual studies–and in particular the results in the first study on a question–can be and should be taken as preliminary findings and can be considered and discussed as such and even used in preliminary and tentative evaluations of theories. In fact, one advantage of replacing significance testing with point estimates and confidence intervals is that these more desirable data analysis methods allow researchers to see that the findings in individual studies are clearly tentative and preliminary. Reform of data analysis procedures does not mean that researchers will no longer be allowed to discuss and interpret their findings in individual studies. In fact, it means they will be able to do so better.

OBJECTION 6

This objection states that significance testing serves the invaluable purpose of ensuring objectivity in the interpretation of research data:

> One of the important distinguishing features of science as a human activity is objectivity. Unlike art or literature, science strives for objectivity. Its conclusions are to be independent of the subjective judgments of any one individual. A major reason why significance testing has become dominant in data analysis is the need to have an objective procedure that does not depend on subjective judgments made by individual researchers. Significance testing provides this objectivity better than alternative procedures. Significance testing is objective in the important sense that the results obtained are (or should be) the same no matter who conducts the test.

ANSWER TO OBJECTION 6

There is an implicit assumption in this objection that significance testing is objective but that confidence intervals are not. In fact, point estimates and confidence intervals are just as objective as significance tests. There is nothing less objective about the computation of confidence interval or the properties of the final confidence interval. Hence objectivity is a non issue.

Implicit in this objection is the assumption that objectivity requires in each study a dichotomous decision as to whether the hypothesis being studied is confirmed (supported) or disconfirmed (not supported). This objection views significance testing as providing an objective, mechanical set of rules for making the necessary binary decision whether to accept or reject the research hypothesis. Point estimates and confidence intervals, on the other hand, do not necessarily produce a dichotomous decision. They provide instead a point estimate and an error band around that estimate. But there is no scientific reason to have or want a binary decision to accept or reject an hypothesis in an individual study—because the final decision on whether to accept or reject an hypothesis cannot depend on the results of one study. Instead, decisions to accept or reject an hypothesis must be based on results integrated across all the studies that have tested that hypothesis. The usual way of doing this is through meta-analysis. (There may be several meta-analyses, because different clusters of studies may have tested the hypothesis in different ways or tested different aspects of the hypothesis.)

This objection appears to be based on a desire for something that is not possible: a procedure that will provide the answer to whether the hypothesis is correct or incorrect from a single study. We know from statistical power analyses that the significance tests cannot do this. The fact that power in typical research is usually between .30 and .60 shows that the individual study does not contain enough information to allow a conclusion about the truth or falseness of the hypothesis tested. The confidence interval reveals this lack of information more clearly and directly: Confidence intervals are typically quite wide. The significance test conceals the lack of information; only those rare researchers who go beyond computing the significance test itself and compute the statistical power of their study ever actually see how little information there is in their study. This concealment is very convenient for users of significance tests: It allows them to feel good about making accept—reject decisions about hypotheses even when the error rate is 50% or more.

To adopt a procedure because it has the virtue of being objective when it has typical error rates of 50%, 60%, or 70% is to make a fetish of objectivity. This is especially so given that an equally objective procedure is available that will hold the total error rate to 5%, 1%, or whatever we want it to be. That procedure is the confidence interval (Schmidt, 1996).

Flipping a coin to decide whether to accept or reject an hypotheses would also be objective. Furthermore, from a statistical power point of view, it would guarantee that the average error rate would not go above 50%—a guarantee the significance test cannot provide. And, of course, it would be more efficient: No study would have to be done. So from the perspective of objectivity as the overriding value, flipping a coin is often a better choice than the significance test.

But aren't there at least some studies with either large enough Ns, large enough effect sizes, or both to produce very high statistical power? If so, can't such a single study answer a scientific question (i.e., provide a sound basis for accepting or rejecting an hypothesis)? And in the case of such studies, why not use significance tests, because power is not a problem?

Taking the last question first: Low statistical power in typical research studies is only one of the severe problems associated with significance testing. There are many other reasons for not using significance testing (see Carver, 1978; Cohen, 1994; Rozeboom, 1960; Schmidt, 1996; and others; see also the responses to Objections 3 and 7 in this chapter).

Second, there are reasons beyond inadequate power why a single study is not adequate to settle a scientific question. That is, there are additional reasons why multiple studies are necessary in science. There may be, for example, measurement differences between studies (e.g., the dependent variable may be measured differently), treatment operationalization differences, study design differences, instructional differences, sample composition differences, and other differences. Once a body of studies has accumulated, meta-analysis can be used to determine whether any of these differences make any difference in research outcomes; a finding that they do not greatly strengthens the final conclusion. For example, the mean effect sizes can be compared for the cluster of studies that used operationalization A and the cluster using operationalization B. In this way, possible methodological moderators of results can be disconfirmed or identified.

In addition to methodological moderators, substantive moderator hypotheses can also be tested in meta-analysis. For example, a researcher may hypothesize that the increased use of computers over time will reduce the correlation between constructs X and Y among high school students. If multiple studies are available across time and are examined in a meta-analysis, this hypothesis can be evaluated. A single study, no matter how large, is unlikely to permit such an analysis. Hence even if statistical power is very high, a single study still cannot answer a scientific question. And it is clear that there is potentially much more useful information in a large number of smaller studies than in a single large study with the same total sample size.

Many primary researchers do not like these facts. They strongly want to believe that their individual primary study can answer an important scientific question. But wanting something to be true does not make it true.

OBJECTION 7

This objection holds that the problem is not use but misuse of significance testing:

> The problem is not the use of significance testing per se, but the misuse. It is true that misuse—and even abuse—of significance testing is quite common in our research literatures. This occurs when researchers give significance tests erroneous interpretations and when significance tests are used despite violations of their underlying assumptions. But the solution is to educate researchers to end their misuses, not to discontinue significance testing. Significance testing is just a tool. We just need to teach researchers to use that tool properly.

ANSWER TO OBJECTION 7

The response to this objection must address two questions. First, is it possible to educate researchers so that they will no longer misinterpret the meaning of significance tests and no longer use them inappropriately? Second, if we could successfully do this, would significance testing then be a useful research tool?

The evidence indicates that the answer to the first question is no. Significance testing has been the dominant mode of data analysis in psychology and many other disciplines for approximately 50 years. During that long period, there has been no apparent decrease in the prevalence of misinterpretation of the meaning of significance tests, despite the efforts of many methodologists. This is in part because some teachers of statistics courses implicitly and explicitly accept and endorse some of these misinterpretations and even incorporate them into their textbooks (Carver, 1978; Schmidt, 1996). Of all these misinterpretations, the one most devastating to the research enterprise is the belief that if a difference or relation is not statistically significant, then it is zero. This is the false belief underlying Objection 1; it was discussed in detail there. Another frequent misinterpretation is the belief that level of statistical significance (p value) indicates the importance or size of a difference or relation. A third is the false belief that statistical significance indicates that a finding is reliable and will replicate if a new study is conducted. These misinterpretations are discussed in some detail in Schmidt (1996) and Oakes (1986); these and still others are discussed by Carver (1978), in what is probably the most thorough published exploration of the misinterpretations researchers place on significance tests. Recently, Carver (1993) stated that there is no indication that his 1978 article has had any noticeable impact on the practices and interpretations of researchers. Perhaps a reasonable conclusion would be that it is unlikely that it will be possible to wean researchers away from their attachment to erroneous beliefs about, and interpretations of, significance tests.

By contrast, the story with respect to point estimates and confidence intervals is very different. Point estimates of effect sizes and their associated confidence intervals are much easier for students and researchers to understand and, as a result, are much less frequently misinterpreted. Any teacher of statistics knows it is much easier for students to understand point estimates and confidence intervals than significance testing with its strangely inverted logic. This is another plus for point estimation and confidence intervals.

But suppose it were possible to eliminate "misuse" of significance tests. Suppose researchers were somehow all educated to avoid misinterpretations of significance tests. Would use of significance tests then be a useful research tool? The answer is no. Even if misinterpretations were eliminated, it still remains that statistical significance testing is singularly unsuited to the task of advancing the development of cumulative scientific knowledge. This is true in interpreting both individual studies and research literatures. In individual studies, significance testing conceals rather than reveals the important information researchers need. Unlike point estimates, it provides no estimate of the size of the effect or relation; so there is no information to indicate whether the effect is minuscule or quite large. Unlike confidence intervals, significance testing provides no information on the degree of uncertainty in the study. Confidence intervals correctly reveal that individual small-sample studies contain very little information about the hypothesis being tested; this is revealed by the fact that the confidence interval is surprisingly wide. The significance test conceals this important piece of evidence from researchers, allowing them to continue to harbor false beliefs about the evidential value of their individual studies.

Another important fact is that even in the absence of misinterpretations of significance tests, the problem of low statistical power would remain. What would change is that in the absence of misinterpretations, researchers would no longer falsely conclude that nonsignificant differences are in fact zero. Instead, they would conclude that they are merely inconclusive, the correct conclusion according to Fisher (1932). However, this "proper" interpretation of significance tests would create major errors in the interpretation of research literatures. A typical research literature, with approximately 50% of studies finding significance and 50% not, would be incorrectly interpreted as inconclusive. Meta-analysis shows that such research literatures do provide the foundation for research conclusions, as explained in the response to Objection 3. In fact, the conclusions that can be reached are often precise conclusions about the magnitudes and variation of effects and relations in the relevant population.

Finally, in addressing the question of whether significance tests would be a useful research tool if only researchers did not misinterpret them, it is useful to remember the function that researchers feel significance tests perform: they provide an objective, mechanical procedure for making a dichotomous, yes–no decision about whether the hypothesis is confirmed or not. But as explained in more detail in the response to Objection 5, there is in fact no need in individual studies for such a binary decision, and it is futile to even attempt such a decision. Findings in any individual study are (at best) preliminary and tentative. Thus the primary purpose for which significance testing has been used is a purpose for which there is no need.

In conclusion, even if there were no misinterpretations of the meaning of significant and nonsignificant findings, significance testing would still not be a useful research tool. On the contrary, its use would still be strongly counterproductive in the interpretation of research literatures. Its use also would mean that significance testing would be displacing procedures that <u>are</u> useful in the research enterprise: point estimates and confidence intervals. This means that <u>any</u> use of significance tests, even in the complete absence of misinterpretations, is a <u>misuse</u> of significance tests.

Significance testing never makes a useful contribution to the development of cumulative knowledge.

OBJECTION 8

This position holds that the effort to reform data analysis methods should be dropped because it is futile. The following statement, from an older researcher, is the purest formulation we have found of this objection:

> The controversy over significance testing is an old debate which crops up repeatedly every 10 years or so. I remember discussing these same issues in a graduate class in the 1950s and reading articles in <u>Psychological Bulletin</u> on this debate in the 1950s and 1960s. This debate keeps repeating itself but never leads to any definitive conclusions or changes in practices. So it is basically just a waste of time. As researchers in the trenches, we must collect data and make decisions. Our tools and methods are imperfect and there are many philosophical problems with them. However, despite this we must work. That is the nature of the game. As a result of these problems and imperfections of our methods, we undoubtedly make errors in our conclusions, but science is an open enterprise, so someone will eventually point out and correct these errors. As

research accumulates, it will become apparent what the truth is. So in the long run, our present system probably does work.

ANSWER TO OBJECTION 8

Not everything in this statement is false. It is true that telling critiques of reliance on significance testing have appeared periodically and stimulated widespread debate and discussion. Schmidt (1996) reviewed the history of this process. The earliest such critique we have been able to find in the psychology literature was published in 1955, over 40 years ago. Jones (1955) pointed out the logical deficiencies of significance testing and called for replacing significance tests with point estimates and confidence intervals. The last such episode prior to the current debate appears to have been in the late 1970s (Carver, 1978; Hunter, 1979). It is also true that none of these attempts to reform data analysis procedures has produced any real change in data analysis practices. In fact, Hubbard (1996) and Hubbard, Parsa, and Luthy (1996) have shown that over the period from the 1950s to the present, the use of significance tests in published research has actually increased. This increase has been continuous and steady and has continued up to 1995.

However, at the beginning of this chapter, we pointed out two reasons why the present situation may be different. For the first time, the APA Board of Scientific Affairs is looking into the possibility of taking the lead in the reform effort. The Board has appointed a task force to study the question and make recommendations. The report produced by this group and its recommendations will for the first time give institutional support to the effort to reform data analysis procedures. We can expect this report to be read by, and to have influence on, journal editors and journal reviewers. We can also expect it to be cited by researchers in their research reports as justification for data analysis procedures that do not include significance tests. In the past, there has been no such "top-down" support for reform; instead, each episode of debate over significance testing has been stimulated by a lone individual who publishes a single critical article, usually without even a follow-up article by that same author, and with no support from prestigious centers within institutional science.

The second development that makes the current situation different is the impact of meta-analysis. Meta-analysis can be used to present dramatic demonstrations of how reliance on significance testing makes it virtually impossible to discern the true meaning of research literatures. Schmidt (1996) presented two such demonstrations; others are presented in Hunter and Schmidt (1990) and Hunter, Schmidt, and Jackson (1982). Most previous critiques of significance testing have focused on logical and philosophical deficiencies and contradictions in reliance on significance testing. These are the issues referred to in Objection 8 as "philosophical problems." An example is the fact that in testing and rejecting the null hypothesis researchers are focusing not on the actual scientific hypothesis of interest, but on a scientifically irrelevant hypothesis. Another is the fact that we always know in advance that the null hypothesis must be false to some degree, so that a finding of statistical significance is merely a reflection of whether the sample size was large or not. Criticisms of this sort should be compelling to researchers but often are not; they are apparently too abstract and seemingly removed from the actual experiences of researchers in doing data analysis to have much impact. On the other hand, the demonstrations that can be produced using meta-analysis are more concrete; they are based on the kinds of data analysis that researchers actually carry out. In our experience, they have a greater impact on audiences and readers.

So for both these reasons we believe there is a basis for the hope that this time the reform effort will actually bear fruit.

Objection 8 states that, because science is an open system, errors in data interpretation resulting from reliance on significance testing eventually will be corrected as more research is conducted on a given question. This statement is erroneous. The average statistical power in may research areas in less than .50. And the predominant interpretational rule is that if a difference or relation is not significant, then it is just a chance finding and can be considered to be zero. That is, the predominant decision rule is: If it is significant, it is a real effect; if it is nonsignificant, it is zero. Under these circumstances, as more and more research studies are conducted the probability of an erroneous conclusion about the meaning of the research literature increases (Hedges & Olkin, 1980; Schmidt, 1996). That is, as more and more studies become available, it becomes increasingly clear that the majority of studies found "no relationship," strengthening the false conclusion that no relationship exists. So it is clear that we cannot rely on the accumulation of additional significance test-based research studies to eliminate erroneous conclusions.

Suppose, on the other hand, that average statistical power is approximately .50. Then as more and more studies are conducted, researchers become increasingly likely to conclude that the research literature is conflicting and inconclusive. So it is again clear that we cannot rely on the accumulation of additional significance test-based research studies to lead us to correct conclusions.

In the last 15 years, application of meta-analysis to such research literatures has corrected such errors resulting from reliance on significance testing (Hunter & Schmidt, 1990; Hunter, Schmidt & Jackson, 1982; Schmidt, 1992, 1996). Meta-analysis can and has revealed (and calibrated with precision) underlying population effects that have been undetected in the majority of the studies going into the meta-analysis. So it is true that at the level of research literatures meta-analysis can detect and correct the errors resulting from reliance on significance testing in individual studies. However, two points are important here. First, this is not the correction process that this objection has in mind; this objection postulates that the accumulation of individual studies per se will correct such errors, a proposition that we see here is patently untrue.

Second, the use of meta-analysis to synthesize research literatures represents the elimination of significance tests at that level of analysis. The effort to reform data analysis procedures has two aspects or prescriptions. The first is that in data analysis in individual studies, point estimates and confidence intervals should replace significance testing. The second is that in analyzing and interpreting research literatures, meta-analysis should replace traditional narrative reviews. This second objective of the reform effort has now largely been realized in many research areas and disciplines. The first objective–reforming data analysis in individual studies–is really an extension downward to individual studies of the second reform. This objective has not been attained. It is the subject of the current reform effort.

The objection that attempts to reform data analysis methods are futile is sometimes expressed in forms that are more cynical than the statement quoted earlier. The following statement from a younger researcher (in his 30s) is an example of this more cynical version of this objection:

> You have ignored an important reason why researchers are not going to give up significance testing. It is easier to use significance testing because it is what reviewers expect and require and some researchers (myself included, I must reluctantly admit) feel they have more important things to do than educate reviewers and editors about significance testing. Such efforts are probably futile anyway; they have been in the past. So we just comply and use significance tests. To be sure, from an ethical standpoint, this response is problematic—we are complicators in an illegitimate process that impairs scientific progress. But most researchers comply with all sorts of illegitimate processes (e.g., worrying more about getting something published than about whether it really advances knowledge in the field, data mining, etc.). I have my moments of weakness like most, but at least I am willing to admit it!

This statement reflects the often expressed feeling that it is futile to fight against a powerful conventional establishment supporting significance tests. This established conventional wisdom may be obviously erroneous, but it cannot be successfully opposed. So one has no choice but to conform. We have heard this feeling expressed by many researchers, particularly younger researchers concerned about publication, tenure, and promotion. This resignation and conformity to error produces an intellectual cynicism that is corrosive of scientific values. It undermines the intellectual honesty and idealistic motivation to search for truth that are the driving engines of science. One important benefit that successful reform of data analysis procedures will bring is the reduction, and perhaps elimination, of such cynicism. This is not a minor benefit.

APPLICATIONS

Each chapter in this book is supposed to contain a section stating how the content of the chapter is to be applied by readers. The most important application of this chapter is this: Do not use significance testing in data analysis; instead, use confidence intervals and point estimates in individual studies and meta-analysis in integrating findings across multiple studies. In doing so, you will be helping to advance cumulative scientific knowledge. However, these applications are contained in earlier articles (e.g., Cohen, 1994; Schmidt, 1996) and are not new to this chapter. The main lesson specific to this chapter is this: Beware of plausible and intuitively appealing objections to discontinuing the use of significance testing. They are always false, no matter how convincing they may seem. Finally, try to show enough intellectual courage and honesty to reject the use of significance tests despite the pressures of social convention to the contrary. Scientists must have integrity.

CONCLUSION

In this chapter, we have examined eight objections to discontinuing significance testing and using instead point estimates and confidence intervals to analyze research data. We have seen that each of these objections, although appearing plausible and even convincing to many researchers, is logically and intellectually bankrupt. What has been shown for these 8 objections can also be shown for each of the additional 79 that we have collected but could not include in this chapter. It could also be shown for any "new" objections yet to be collected by us. Significance testing never makes a positive contribution.

For almost 3 years, we have challenged researchers to describe even one legitimate contribution that significance testing makes or has made to the research enterprise (i.e., to the development of cumulative research knowledge). This challenge has resulted in the large collection of objections from which the eight presented in this chapter are sampled. But it has produced no examples of contributions significance testing has made to research. The fight to reform data analysis methods so that those methods contribute to the development of knowledge, rather than detract from this effort, has been long and hard. It has been resisted at every turn by rock-hard psychological defenses. But there is now reason to hope that reform is finally a realistic possibility.

REFERENCES

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 41*, 1–26.

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

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287–292.

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

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

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*, 98–101.

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

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Edinburgh, Scotland: Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments* (subsequent editions 1937, 1942, 1947, 1949, 1951, 1966). London: Oliver and Boyd.

Fisher, R. A. (1959). *Statistical methods and scientific inference* (2nd ed.). Edinburgh, Scotland: Oliver and Boyd.

Fisher, R. A. (1973). *Statistical methods and scientific inference* (3nd ed.). Edinburgh, Scotland: Oliver and Boyd.

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

Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist, 42*, 443–455.

Hedges, L. V., & Olkin, I. (1980). Vote counting methods in research synthesis. Psychological Bulletin, 88, 359–369.

Hubbard, R. (1996). *Sanctifying significance and relegating replication: A misplaced emphasis.* Manuscript under review.

Hubbard, R., Parsa, A. R., & Luthy, M. R. (1996). *The diffusion of statistical significance testing in psychology: The case of the Journal of Applied Psychology*. Manuscript under review.

Hunter, J. E. (1979, September). *Cumulating results across studies: A critique of factor analysis, canonical correlation, MANOVA, and statistical significance testing*. Invited address presented at the 86th Annual Convention of the American Psychological Association, New York.

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Jones, L. V. (1955). Statisics and research design. *Annual Review of Psychology, 6,* 405–430. Stanford, CA: Annual Reviews, Inc.

Loftus, G. R. (1994, August). *Why psychology will never be a real science until we change the way we analyze data.* Address presented at the American Psychological Association 102nd annual convention, Los Angeles.

Neyman, J. (1962). Two breakthroughs in the theory of statistical decision making. *Review of the International Statistical Institute, 25,* 11–27.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, 231,* 289–337.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

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? Research findings, meta-analysis, and cumulative knowledge in psychology. *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., & Urry, V. E. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473–485.

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