Psychological Methods
1996, Vol. 1, No. 2, 115-129

1082-989X/96/$3.00

Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers

Frank L. Schmidt
University of Iowa

Data analysis methods in psychology still emphasize statistical significance testing, despite numerous articles demonstrating its severe deficiencies. It is now possible to use meta-analysis to show that reliance on significance testing retards the development of cumulative knowledge. But reform of teaching and practice will also require that researchers learn that the benefits that they believe flow from use of significance testing are illusory. Teachers must revamp their courses to bring students to understand that (a) reliance on significance testing retards the growth of cumulative research knowledge; (b) benefits widely believed to flow from significance testing do not in fact exist; and (c) significance testing methods must be replaced with point estimates and confidence intervals in individual studies and with meta-analyses in the integration of multiple studies. This reform is essential to the future progress of cumulative knowledge in psychological research.

In 1990, Aiken, West, Sechrest, and Reno published an important article surveying the teaching of quantitative methods in graduate psychology programs. They were concerned about what was not being taught or was being inadequately taught to future researchers and the harm this might cause to research progress in psychology. For example, they found that new and important quantitative methods such as causal modeling, confirmatory factor analysis, and meta-analysis were not being taught in the majority of graduate programs. This is indeed a legitimate cause for concern. But in this article, I am concerned about the opposite: what is being taught and the harm that this is doing. Aiken et al. found that the vast majority of programs were teaching, on a rather thorough basis, what they referred to as "the old standards of statistics": traditional inferential statistics. This includes the t test, the F test, the chi-square test, analysis of variance (ANOVA), and other methods of statistical significance testing. Hypothesis testing based on the statistical significance test has been the main feature of graduate training in statistics in psychology for over 40 years, and the Aiken et al. study showed that it still is.

Methods of data analysis and interpretation have a major effect on the development of cumulative knowledge. I demonstrate in this article that reliance on statistical significance testing in the analysis and interpretation of research data has systematically retarded the growth of cumulative knowledge in psychology (Hunter & Schmidt, 1990b; Schmidt, 1992). This conclusion is not new. It has been articulated in different ways by Rozeboom (1960), Meehl (1967), Carver (1978), Guttman (1985), Oakes (1986), Loftus (1991, 1994), and others, and most recently by Cohen (1994). Jack Hunter and I have used meta-analysis methods to show that these traditional data analysis methods militate against the discovery of the underlying regularities and relationships that are the

An earlier version of this article was presented as the presidential address to the Division of Evaluation, Measurement and Statistics (Division 5 of the American Psychological Association) at the 102nd Annual Convention of the American Psychological Association, August 13, 1994, Los Angeles, California. This article is largely drawn from work that John Hunter and I have done on meta-analysis over the years, and I would like to thank John Hunter for his comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Frank L. Schmidt, Department of Management and Organizations, College of Business, University of Iowa, Iowa City, Iowa 52242. Electronic mail may be sent via Internet to frank-schmidt@uiowa.edu.

foundation for scientific progress (Hunter & Schmidt, 1990b). Those of us who are the keepers of the methodological and quantitative flame for the field of psychology bear the major responsibility for this failure because we have continued to emphasize significance testing in the training of graduate students despite clear demonstrations of the deficiencies of this approach to data analysis. We correctly decry the fact that quantitative methods are given inadequate attention in graduate programs, and we worry that this signals a future decline in research quality. Yet it was our excessive emphasis on so-called inferential statistical methods that caused a much more serious problem. And we ignore this fact.

My conclusion is 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. We must teach that for analysis of data from multiple studies, the appropriate method is meta-analysis. I am not the first to reach the conclusion that significance testing should be replaced by point estimates and confidence intervals. Jones stated this conclusion as early as 1955, and Kish in 1959. Rozeboom reached this conclusion in 1960. Carver stated this conclusion in 1978, as did Hunter in 1979 in an invited American Psychological Association (APA) address, and Oakes in his excellent 1986 book. So far, these individuals (and others) have all been voices crying in the wilderness.

Why then is the situation any different today? If the closely reasoned and logically flawless arguments of Kish, Rozeboom, Carver, and Hunter have been ignored all these years—and they have—what reason is there to believe that this will not continue to be the case? There is in fact a reason to be optimistic that in the future we will see reform of data analysis methods in psychology. That reason is the development and widespread use of meta-analysis methods. These methods have revealed more clearly than ever before the extent to which reliance on significance testing has retarded the growth of cumulative knowledge in psychology. These demonstrations based on meta-analysis methods are what is new. As conclusions from research literature come more and more to be based on findings from meta-analysis (Cooper & Hedges, 1994; Lipsey & Wilson, 1993; Schmidt, 1992), the significance test necessarily becomes less and less important. At worst, significance tests will become progressively deemphasized. At best, their use will be discontinued and replaced in individual studies by point estimation of effect sizes and confidence intervals.

The reader's reaction to this might be that this is just one opinion and that there are defenses of statistical significance testing that are as convincing as the arguments and demonstrations I present in this article. This is not true. As Oakes (1986) stated, it is "extraordinarily difficult to find a statistician who argues explicitly in favor of the retention of significance tests" (p. 71). A few psychologists have so argued. But Oakes (1986) and Carver (1978) have carefully considered all such arguments and shown them to be logically flawed and hence false. Also, even these few defenders of significance testing (e.g., Winch & 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. As you read this article, I want you to consider this challenge: 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.

Traditional Methods Versus Meta-Analysis

Psychology and the other social sciences have traditionally relied heavily on the statistical significance test in interpreting the meaning of data, both in individual studies and in research literature. Following the fateful lead of Fisher (1932), null hypothesis significance testing has been the dominant data analysis procedure. The prevailing decision rule, as Oakes (1986) has demonstrated empirically, has been this: If the statistic (t, F, etc.) is significant, there is an effect (or a relation); if it is not significant, then there is no effect (or relation). These prevailing interpretational procedures have focused heavily on the control of Type I errors, with little attention being paid to the control of Type II errors. A Type I error (alpha error) consists of concluding that there is a relation or an effect when there is not. A Type II error (beta error) consists of the opposite, concluding

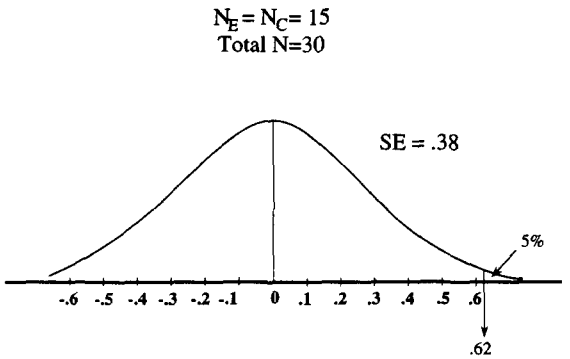

Figure 1. Null distribution of d values in a series of experiments. Required for significance: $d_c = 0.62$; $d_c = [1.645(0.38)] = 0.62$ (one-tailed test, $\alpha = .05$).

that there is no relation or effect when there is. Alpha levels have been controlled at the .05 or .01 levels, but beta levels have by default been allowed to climb to high levels, often in the 50% to 80% range (Cohen, 1962, 1988, 1990, 1994; Schmidt, Hunter, & Urry, 1976). To illustrate this, let us look at an example from a hypothetical but statistically typical area of experimental psychology.

Suppose the research question is the effect of a certain drug on learning, and suppose the actual effect of a particular dosage is an increase of one half of a standard deviation in the amount learned. An effect size of .50 is considered medium-sized by Cohen (1988) and corresponds to the difference between the 50th and 69th percentiles in a normal distribution. With an effect size of this magnitude, 69% of the experimental group would exceed the mean of the control group, if both were normally distributed. Many reviews of various literatures have found relations of this general magnitude (Hunter & Schmidt, 1990b). Now suppose that a large number of studies are conducted on this dosage, each with 15 rats in the experimental group and 15 in the control group.

Figure 1 shows the distribution of effect sizes (d values) expected under the null hypothesis. All variability around the mean value of zero is due to sampling error. To be significant at the .05 level (with a one-tailed test), the effect size must be .62 or larger. If the null hypothesis is true, only 5% will be that large or larger. In analyzing their data, researchers in psychology typically focus only on the information in Figure 1. Most believe that their significance test limits the probability of an error to 5%.

Actually, in this example the probability of a Type I error is zero, not 5%. Because the actual effect size is always .50, the null hypothesis is always false, and therefore there is no possibility of a Type I error. One cannot falsely conclude that there is an effect when in fact there is an effect. When the null hypothesis is false, the only kind of error that can occur is a Type II error: failure to detect the effect that is present (and the total error rate for the study is therefore the Type II error rate). The only type of error that can occur is the type that is not controlled.

Figure 2 shows not only the irrelevant null distribution but also the actual distribution of effect sizes across these studies. The mean of this distribution is the true value of .50, but because of sampling error, there is substantial variation in observed effect sizes. Again, to be significant, the effect size must be .62 or larger. Only 37% of studies conducted will obtain a significant effect size; thus statistical power for each of these studies is only .37. That is, the true (population) effect size of the drug is always .50; it is never zero. Yet it is only detected as significant in 37% of the studies. The error rate in this research literature is 63%, not 5%, as many would mistakenly believe.

In actuality, the error rate would be even higher. Most researchers in experimental psychology would traditionally have used F tests from an ANOVA to analyze these data. This means the significance test would be two-tailed rather than one-tailed as in our example. With a two-tailed

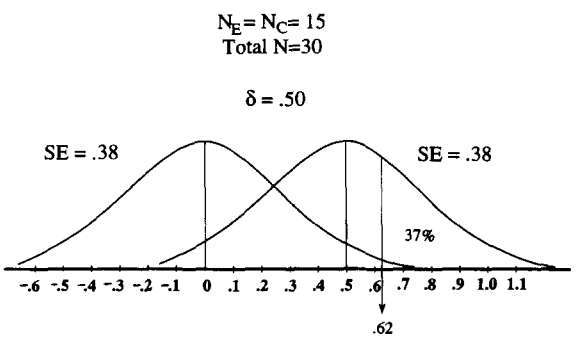

Figure 2. Statistical power in a series of experiments. Required for significance: $d_c = 0.62$ (one-tailed test, $\alpha = .05$); statistical power = 0.37; Type II error rate = 63%; Type I error rate = 0%.

test (i.e., one-way ANOVA), statistical power is even lower: .26 instead of .37. The Type II error rate (and hence the overall error rate) would be 74%. Also, this example assumes use of a z test; any researchers not using an ANOVA would probably use a t test. For a one-tailed t test with $\alpha = .05$ and $df = 28$, the effect size (d value) must be .65 to be significant. (The t value must be at least 1.70, instead of the 1.645 required for the z test.) With the t test, statistical power would also be lower: .35 instead of .37. Thus both commonly used alternative significance tests would yield even lower statistical power and produce even higher error rates.

Also, note in Figure 2 that the studies that are significant yield distorted estimates of effect sizes. The true effect size is always .50; all departures from .50 are due solely to sampling error. But the minimum value required for significance is .62. The obtained d value must be .12 above its true value—24% larger than its real value—to be significant. The average of the significant d values is .89, which is 78% larger than the true value of .50.

In any study in this research literature that by chance yields the correct value of .50, the conclusion under the prevailing decision rule is that there is no relationship. That is, it is only the studies that by chance are quite inaccurate that lead to the correct conclusion that a relationship exists.

How would this body of studies be interpreted as a research literature? There are two interpretations that would have traditionally been accepted. The first is based on the traditional voting method (critiqued by Light & Smith, 1971, and by Hedges & Olkin, 1980). Using this method, one would note that 63% of the studies found "no relationship." Since this is a majority of the studies, the conclusion would be that no relation exists. This conclusion is completely false, yet many reviews in the past have been conducted in just this manner (Hedges & Olkin, 1980).

The second interpretation is as follows: In 63% of the studies, the drug had no effect. However, in 37% of the studies, the drug did have an effect. (Moreover, when the drug did have an effect, the effect was quite large, averaging .89.) Research is needed to identify the moderator variables (interactions) that cause the drug to have an effect in some studies but not in others. For example, perhaps the strain of rat used or the mode of injecting the drug affects study outcomes. This interpretation is also completely erroneous. In addition, it leads to wasted research efforts to identify nonexistent moderator variables.

I. Compute Actual Variance of Effect Sizes

1. $S_d^2 = .1444$ (Observed Variance of d Values)
2. $S_e^2 = .1444$ (Variance Predicted from Sampling Error)
3. $S_\delta^2 = S_d^2 - S_e^2$
4. $S_\delta^2 = .1444 - .1444 = 0$ (True Variance of δ Values)

II. Compute Mean Effect Size

1. $\overline{d} = .50$ (Mean Observed d Value)
2. $\delta = .50$
3. $SD_\delta = 0$

III. Conclusion: There is only one effect size, and its value is .50 standard deviation.

Figure 3. Meta-analysis of drug studies.

Both traditional interpretations fail to reveal the true meaning of the studies and hence fail to lead to cumulative knowledge. In fact, the traditional methods based on significance testing make it impossible to reach correct conclusions about the meaning of these studies. This is what is meant by the statement that traditional data analysis methods militate against the development of cumulative knowledge.

How would meta-analysis interpret these studies? Different approaches to meta-analysis use somewhat different quantitative procedures (Bangert-Drowns, 1986; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Hunter & Schmidt, 1990b; Rosenthal, 1984, 1991). I illustrate this example using the methods presented by Hunter, Schmidt, and Jackson (1982) and Hunter and Schmidt (1990b). Figure 3 shows that meta-analysis reaches the correct conclusion. Meta-analysis first computes the variance of the observed d values (using the ordinary formula for the variance of any set of numbers). Next, it uses the standard formula for the sampling error variance of d values (e.g., see

Hunter & Schmidt, 1990b, chap. 7) to determine how much variance would be expected in observed d values from sampling error alone. The amount of real variance in population d values (δ values) is estimated as the difference between the two. In our example, this difference is zero, indicating correctly that there is only one population value. This single population value is estimated as the average observed value, which is .50 here, the correct value. If the number of studies is large, the average d value will be close to the true (population) value, because sampling errors are random and hence average out to zero.[1]

Note that these meta-analysis methods do not rely on statistical significance tests. Only effect sizes are used, and significance tests are not used in analyzing the effect sizes. Unlike traditional methods based on significance tests, meta-analysis leads to correct conclusions and hence leads to cumulative knowledge.

The data in this example are hypothetical. However, if one accepts the validity of basic statistical formulas for sampling error, one will have no reservations about this example. But the same principles do apply to real data, as shown next by an example from research in personnel selection. Table 1 shows observed validity coefficients (correlations) from 21 studies of a single clerical test and a single measure of job performance. Each study has $n = 68$ (the median n in the literature in personnel psychology), and every study is a random draw (without replacement) from a single larger validity study with 1,428 subjects. The correlation in the large study (uncorrected for measurement error, range restriction, or other artifacts) is .22 (Schmidt, Ocasio, Hillery, & Hunter, 1985).

The validity is significant in 8 (or 38%) of these studies, for an error rate of 62%. The traditional conclusion would be that this test is valid in 38% of the organizations, and invalid in the rest, and that in organizations in which it is valid, its mean observed validity is .33 (which is 50% larger than its real value). Meta-analysis of these validities indicates that the mean is .22 and that all variance in the coefficients is due solely to sampling error. The meta-analysis conclusions are correct; the traditional conclusions are false.

In these examples, the only type of error that is controlled—Type I error—is the type that cannot occur. In most areas of research, as time goes by, researchers gain a better and better understanding of the processes they are studying; as a result, it is less and less frequently the case that the null hypothesis is "true" and more and more likely that the null hypothesis is false. Thus Type I error decreases in importance, and Type II error increases in importance. This means that as time goes by, researchers should be paying increasing attention to Type II error and to statistical power and increasingly less attention to Type I error. However, a recent review in *Psychological Bulletin* (Sedlmeier & Gigerenzer, 1989) concluded that the average statistical power of studies in one APA journal had declined from 46% to 37% over a 22-

Table 1
21 Validity Studies, N = 68 for Each

Study	Observed validity correlation
1	.04
2	.14
3	.31*
4	.12
5	.38*
6	.27*
7	.15
8	.36*
9	.20
10	.02
11	.23
12	.11
13	.21
14	.37*
15	.14
16	.29*
17	.26*
18	.17
19	.39*
20	.22
21	.21

* $p < .05$ (two tailed).

[1] Actually the d statistic has a slight positive (upward) bias as the estimator of δ, the population value. Formulas are available to correct observed d values for this bias and are given in Hedges and Olkin (1985, p. 81) and Hunter and Schmidt (1990b, p. 262). This example assumes that this correction has been made. This bias is trivial if the sample size is 10 or greater in both the experimental and control groups.

year period (despite the earlier appeal in that journal by Cohen in 1962 for attention to statistical power). Only 2 of the 64 experiments reviewed even mentioned statistical power, and none computed estimates of power. The review concluded that the decline in power was due to increased use of alpha-adjusted procedures (such as the Newman-Keuls, Duncan, and Scheffe procedures). That is, instead of attempting to reduce the Type II error rate, researchers had been imposing increasingly stringent controls on Type I errors, which probably cannot occur in most studies. The result is a further increase in the Type II error rate, an average increase of 17%. This trend illustrates the deep illogic embedded in the use of significance tests.

These examples have examined only the effects of sampling error. There are other statistical and measurement artifacts that cause artifactual variation in effect sizes and correlations across studies, for example, differences between studies in amount of measurement error, in degree of range restriction, and in dichotomization of measures. Also, in meta-analysis, *d* values and correlations must be corrected for downward bias due to such research artifacts as measurement error and dichotomization of measures. These artifacts are beyond the scope of this presentation but are covered in detail elsewhere (Hunter & Schmidt, 1990a, 1990b; Schmidt & Hunter, 1996). My purpose here is to demonstrate only that traditional data analysis and interpretation methods logically lead to erroneous conclusions and to demonstrate that meta-analysis solves these problems at the level of aggregate research literatures.

For almost 50 years, reliance on statistical significance testing in psychology and the other social sciences has led to frequent serious errors in interpreting the meaning of data (Hunter & Schmidt, 1990b, pp. 29–42 and 483–484), errors that have systematically retarded the growth of cumulative knowledge. Despite the best efforts of such individuals as Kish (1959), Rozeboom (1960), Meehl (1967), Carver (1978), Hunter (1979), Guttman (1985), and Oakes (1986), it has not been possible to wean researchers away from their entrancement with significance testing. Can we now at least hope that the lessons from meta-analysis will finally stimulate change? I would like to answer in the affirmative, but later in this article I present reasons why I do not believe these demonstrations alone will be sufficient to bring about reform. Other steps are also needed.

In my introduction, I state that the appropriate method for analyzing data from multiple studies is meta-analysis. These two examples illustrate that point dramatically. I also state that the appropriate way to analyze data in a single study is by means of point estimation of the effect size and use of a confidence interval around this point estimate. If this had been done in the studies in these two examples, what would these two research literatures have looked like prior to application of meta-analysis?

In the first example, from the experimental psychology literature, the traditional practice would have been to report only the *F* statistic values and their associated significance levels. Anyone looking at this literature would see that 26% of these *F* ratios are significant and 74% are nonsignificant. This would create at best the impression of a contradictory set of studies. With appropriate data analysis methods, the observed *d* value is computed in each study; this is the point estimate of the effect size. Anyone looking at this literature would quickly see that the vast majority of these effect sizes—91%—are positive. This gives a very different and much more accurate impression than does the observation that 74% of the effects are nonsignificant. Next, the confidence interval around each effect size is computed and presented. A glance at these confidence intervals would reveal that almost all of them overlap with almost all of the other confidence intervals. This again correctly suggests that the studies are in substantial agreement, contrary to the false impression given by the traditional information that 26% are significant and 74% are nonsignificant. (These studies use simple one-way ANOVA designs; however, *d* values and confidence intervals can also be computed when factorial ANOVA designs or repeated measures designs are used.)

To see this point more clearly, let us consider again the observed correlations in Table 1. The observed correlation is an index of effect size, and therefore in a truly traditional analysis it would not be reported; only significance levels would be reported. So all we would know is that in 62% of the studies there was "no significant relationship," and in 38% of the studies there

Table 2
95% Confidence Intervals for Correlations From Table 1, N = 68 for Each

Study	Observed correlation	95% confidence interval	
		Lower	Upper
1	.39	.19	.59
2	.38	.18	.58
3	.37	.16	.58
4	.36	.15	.57
5	.31	.09	.53
6	.29	.07	.51
7	.27	.05	.49
8	.26	.04	.48
9	.23	.00	.46
10	.22	−.01	.45
11	.21	−.02	.44
12	.21	−.02	.44
13	.20	−.03	.43
14	.17	−.06	.40
15	.15	−.08	.38
16	.14	−.09	.37
17	.14	−.09	.37
18	.12	−.12	.36
19	.11	−.13	.35
20	.04	−.20	.28
21	.02	−.22	.26

was a significant relationship. Table 2 shows the information that would be provided by use of point estimates of effect size and confidence intervals. In Table 2, the observed correlations are arranged in order of size with their 95% confidence intervals.

The first thing that is obvious is that all the correlations are positive. It can also be seen that every confidence interval overlaps every other confidence interval, indicating that these studies could all be estimating the same population parameter, which indeed they are. This is true for even the largest and smallest correlations. The confidence interval for the largest correlation (.39) is .19 to .59. The confidence interval for the smallest correlation (.02) is −.22 to 26. These confidence intervals have an overlap of .07. Thus in contrast to the picture provided by null hypothesis significance testing, point estimates and confidence intervals provide a much more correct picture, a picture that correctly indicates substantial agreement among the studies.[2]

There are also other reasons for preferring confidence intervals (see Carver, 1978; Hunter & Schmidt, 1990b, pp. 29–33; Kish, 1959; Oakes, 1986, p. 67; and Rozeboom, 1960). One important reason is that, unlike the significance test, the confidence interval does hold the overall error rate to the desired level. In the example from experimental psychology, we saw that many researchers believed that the error rate for the significance test was held to 5% because the alpha level used was .05, when in fact the error rate was really 63% (74% if F tests from an ANOVA are used). However, if the 95% confidence interval is used, the overall error rate is in fact held to 5%. Only 5% of such computed confidence intervals will be expected to not include the population (true) effect size and 95% will.

To many researchers today, the idea of substituting point estimates and confidence intervals for significance tests might seem radical. Therefore it is important to remember that prior to the appearance of Fisher's 1932 and 1935 texts, data analysis in individual studies was typically conducted using point estimates and confidence intervals (Oakes, 1986). The point estimates were usually accompanied by estimates of the "probable error," the 50% confidence interval. Significance tests were rarely used (and confidence intervals were not interpreted in terms of statistical significance). Most of us rarely look at the psychological journals of the 1920s and early 1930s, but if we did, this is what we would see. As can be seen both in the psychology research journals and in psychology statistics textbooks, during the latter half of the 1930s and during the 1940s, under the influence

[2] The confidence intervals in Table 2 have been computed using the usual formula for the standard error of the sample correlation: $SE = (1 - r^2)/\sqrt{N - 1}$. Hence these confidence intervals are symmetrical. Some would advocate the use of Fisher's Z transformation of r in computing confidence intervals for r. This position is typically based on the belief that the sampling distribution of Fisher's Z transformation of r is normally distributed, while r itself is skewed. Actually, both are skewed and both approach normality as N increases, and the Fisher's Z transformation approaches normality only marginally faster than r as N increases. For a population correlation of .22 and sample sizes in the ranges considered here, the differences are trivial.

of Fisher, psychological researchers adopted en masse Fisher's null hypothesis significance testing approach to analysis of data in individual studies (Huberty, 1993). This was a major mistake. It was Sir Ronald Fisher who led psychological researchers down the primrose path of significance testing. All the other social sciences were similarly deceived, as were researchers in medicine, finance, marketing, and other areas.

Fisher's influence not only explains this unfortunate change, it also suggests one reason why psychologists for so long gave virtually no attention to the question of statistical power. The concept of statistical power does not exist in Fisherian statistics. In Fisherian statistics, the focus of attention is solely on the null hypothesis. No alternative hypothesis is introduced. Without an alternative hypothesis, there can be no concept of statistical power. When Neyman and Pearson (1932, 1933) later introduced the concepts of the alternate hypothesis and statistical power, Fisher argued that statistical power was irrelevant to statistical significance testing as used in scientific inference (Oakes, 1986). We have seen in our two examples how untrue that statement is.

Thus it is clear that even if meta-analysis had never been developed, use of point estimates of effect size and confidence intervals in interpreting data in individual studies would have made our research literatures far less confusing, far less apparently contradictory, and far more informative than those that have been produced by the dominant practice of reliance on significance tests. Indeed, the fact of almost universal reliance on significance tests in data analysis in individual studies is a major factor in making meta-analysis absolutely essential to making sense of research literatures (Hunter & Schmidt, 1990b, chap. 1).

However, it is important to understand that meta-analysis would still be useful even had researchers always relied only on point estimates and confidence intervals, because the very large numbers of studies characteristic of many of today's literatures create information overload even if each study has been appropriately analyzed. Indeed, we saw earlier that applying meta-analysis to the studies in Table 1 produces an even clearer and more accurate picture of the meaning of these studies than the application of point estimates and confidence intervals shown in Table 2. In our example, meta-analysis tells us that there is only one population correlation and that that value is .22. Confidence intervals tell us only that there may be only one population value; they do not specify what that value might be. In addition, in more complex applications, meta-analysis makes possible corrections for the effects of other artifacts—both systematic and unsystematic—that bias effect size estimates and cause false variation in such estimates across studies (Hunter & Schmidt, 1990b).

Consequences of Traditional Significance Testing

As we have seen, traditional reliance on statistical significance testing leads to the false appearance of conflicting and internally contradictory research literatures. This has a debilitating effect on the general research effort to develop cumulative theoretical knowledge and understanding. However, it is also important to note that it destroys the usefulness of psychological research as a means for solving practical problems in society.

The sequence of events has been much the same in one applied research area after another. First, there is initial optimism about using social science research to answer socially important questions that arise. Do government-sponsored job-training programs work? One will do studies to find out. Does integration increase the school achievement of Black children? Research will provide the answer. Next, several studies on the question are conducted, but the results are conflicting. There is some disappointment that the question has not been answered, but policymakers—and people in general—are still optimistic. They, along with the researchers, conclude that more research is needed to identify the interactions (moderators) that have caused the conflicting findings. For example, perhaps whether job training works depends on the age and education of the trainees. Maybe smaller classes in the schools are beneficial only for lower IQ children. Researchers may hypothesize that psychotherapy works for middle-class patients but not lower-class patients.

In the third phase, a large number of research studies are funded and conducted to test these moderator hypotheses. When they are completed, there is now a large body of studies, but instead

of being resolved, the number of conflicts increases. The moderator hypotheses from the initial studies are not borne out. No one can make much sense out of the conflicting findings. Researchers conclude that the phenomenon that was selected for study in this particular case has turned out to be hopelessly complex, and so they turn to the investigation of another question, hoping that this time the question will turn out to be more tractable. Research sponsors, government officials, and the public become disenchanted and cynical. Research funding agencies cut money for research in this area and in related areas. After this cycle has been repeated enough times, social and behavioral scientists themselves become cynical about the value of their own work, and they begin to express doubts about whether behavioral and social science research is capable in principle of developing cumulative knowledge and providing general answers to socially important questions (e.g., see Cronbach, 1975; Gergen, 1982; Meehl, 1978). Cronbach's (1975) article "The Two Disciplines of Scientific Psychology Revisited" is a clear statement of this sense of hopelessness.

Clearly, at this point the need is not for more primary research studies but for some means of making sense of the vast number of accumulated study findings. This is the purpose of meta-analysis. Applications of meta-analysis to accumulated research literatures have generally shown that research findings are not nearly as conflicting as we had thought and that useful general conclusions can be drawn from past research. I have summarized some of these findings in a recent article (Schmidt, 1992; see also Hunter & Schmidt, in press). Thus, socially important applied questions can be answered.

Even more important, it means that scientific progress is possible. It means that cumulative understanding and progress in theory development is possible after all. It means that the behavioral and social sciences can attain the status of true sciences; they are not doomed forever to the status of quasi-sciences or pseudosciences. One result of this is that the gloom, cynicism, and nihilism that have enveloped many in the behavioral and social sciences is lifting. Young people starting out in the behavioral and social sciences today can hope for a much brighter future.

These are among the considerable benefits of abandoning statistical significance testing in favor of point estimates of effect sizes and confidence intervals in individual studies and meta-analysis for combining findings across multiple studies.

Is Statistical Power the Solution?

So far in this article, the deficiencies of significance testing that I have emphasized are those stemming from low statistical power in typical studies. Significance testing has other important problems, and I discuss some of these later. However, in our work on meta-analysis methods, John Hunter and I have repeatedly been confronted by researchers who state that the only problem with significance testing is low power and that if this problem could be solved, there would be no problems with reliance on significance testing in data analysis and interpretation. Almost invariably, these individuals see the solution as larger sample sizes. They believe that the problem would be solved if every researcher before conducting each study would calculate the number of subjects needed for "adequate" power (usually taken as power of .80), given the expected effect size and the desired alpha level, and then use that sample size.

What this position overlooks is that this requirement would make it impossible for most studies ever to be conducted. At the inception of research in a given area, the questions are often of the form, "Does Treatment A have an effect?" If Treatment A indeed has a substantial effect, the sample size needed for adequate power may not be prohibitively large. But as research develops, subsequent questions tend to take the form, "Does Treatment A have a larger effect than does Treatment B?" The effect size then becomes the difference between the two effects. A similar progression occurs in correlational research. Such effect sizes will often be much smaller, and the required sample sizes are therefore often quite large, often 1,000 or more (Schmidt & Hunter, 1978). This is just to attain power of .80, which still allows a 20% Type II error rate when the null hypothesis is false. Many researchers cannot obtain that many subjects, no matter how hard they try; either it is beyond their resources or the subjects are just unavailable at any cost. Thus the upshot of this position would be that many—perhaps most—studies would not be conducted at all.

People advocating the position being critiqued here would say this would be no loss at all. They argue that a study with inadequate power contributes nothing and therefore should not be conducted. But such studies do contain valuable information when combined with others like them in a meta-analysis. In fact, very precise meta-analysis results can be obtained on the basis of studies that all have inadequate statistical power individually. The information in these studies is lost if these studies are never conducted.

The belief that such studies are worthless is based on two false assumptions: (a) the assumption that each individual study must be able to support and justify a conclusion, and (b) the assumption that every study should be analyzed with significance tests. In fact, meta-analysis has made clear that any single study is rarely adequate by itself to answer a scientific question. Therefore each study should be considered as a data point to be contributed to a later meta-analysis, and individual studies should be analyzed using not significance tests but point estimates of effect sizes and confidence intervals.

How, then, can we solve the problem of statistical power in individual studies? Actually, this problem is a pseudoproblem. It can be "solved" by discontinuing the significance test. As Oakes (1986, p. 68) noted, statistical power is a legitimate concept only within the context of statistical significance testing. If significance testing is no longer used, then the concept of statistical power has no place and is not meaningful. In particular, there need be no concern with statistical power when point estimates and confidence intervals are used to analyze data in studies and when meta-analysis is used to integrate findings across studies.[3] Thus when there is no significance testing, there are no statistical power problems.

Why Are Researchers Addicted to Significance Testing?

Time after time, even in recent years, I have seen researchers who have learned to understand the deceptiveness of significance testing sink back into the habit of reliance on significance testing. I have occasionally done it myself. Why is it so hard for us to break our addiction to significance testing? Methodologists such as Bakan (1966), Meehl (1967), Rozeboom (1960), Oakes (1986), Carver (1978), and others have explored the various possible reasons why researchers seem to be unable to give up significance testing.

Significance testing creates an illusion of objectivity, and objectivity is a critical value in science. But objectivity makes a negative contribution when it sabotages the research enterprise by making it impossible to reach correct conclusions about the meaning of data.

Researchers conform to the dominant practice of reliance on significance testing because they fear that failure to follow these conventional practices would cause their studies to be rejected by journal editors. But the solution to this problem is not conformity to counterproductive practices but education of editors and reviewers.

There is also a feeling that, as bad as significance testing is, there is no satisfactory alternative; just looking at the data and making interpretations will not do. But as we have seen, there is a good statistical alternative: point estimates and confidence intervals.

However, I do not believe that these and similar reasons are the whole story. An important part of the explanation is that researchers hold false beliefs about significance testing, beliefs that tell them that significance testing offers important benefits to researchers that it in fact does not. Three of these beliefs are particularly important.

The first is the false belief that the significance level of a study indicates the probability of successful replication of the study. Oakes (1986, pp. 79–82) empirically studied the beliefs about the meaning of significance tests of 70 research psychologists and advanced graduate students. They were presented with the following scenario:

> Suppose you have a treatment which you suspect may alter performance on a certain task. You compare the means of your control and experimental

[3] Some state that confidence intervals are the same as significance tests, because if the lower bound of the confidence interval does not include zero, that fact indicates that the effect size estimate is statistically significant. But the fact that the confidence interval can be interpreted as a significance test does not mean that it must be so interpreted. There is no necessity for such an interpretation, and as noted earlier, the probable errors (50% confidence intervals) popularly used in the literature up until the mid 1930s were never interpreted as significance tests.

> groups (20 subjects in each). Further, suppose you use a simple independent means t test and your result is $t = 2.7$, d.f. = 38, $p = .01$. (Oakes, 1986, p. 79)

He then asked them to indicate whether each of several statements were true or false. One of these statements was this:

> You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result in 99% of such studies. (Oakes, 1986, p. 79)

Sixty percent of the researchers indicated that this false statement is true. The significance level gives no information about the probability of replication. This statement confuses significance level with power. The probability of replication is the power of the study; the power of this study is not .99, but rather .43.[4] If this study is repeated many times, the best estimate is that less than half of all such studies will be significant at the chosen alpha level of .01. Yet 60% of the researchers endorsed the belief that 99% of such studies would be significant. This false belief may help to explain the traditional indifference to power among researchers. Many researchers believe a power analysis does not provide any information not already given by the significance level. Furthermore, this belief leads to the false conclusion that statistical power for every statistically significant finding is very high, at least .95.

That many researchers hold this false belief has been known for decades. Bakan criticized this error in 1966, and Lykken discussed it at some length in 1968. The following statement from an introductory statistics textbook by Nunnally (1975) is a clear statement of this belief:

> If the statistical significance is at the .05 level, it is more informative to talk about the *statistical confidence* as being at the .95 level. This means that the investigator can be confident with odds of 95 out of 100 that the observed difference will hold up in future investigations. (p. 195)

Most researchers, however, do not usually explicitly state this belief. The fact that they hold it is revealed by their description of statistically significant findings. Researchers obtaining a statistically significant result often refer to it as "a reliable difference," meaning one that is replicable. In fact, a false argument frequently heard in favor of significance testing is that we must have significance tests in order to know whether our findings are reliable or not. As Carver (1978) pointed out, the popularity of statistical significance testing would be greatly reduced if researchers could be made to realize that the statistical significance level does not indicate the replicability of research data. So it is critical that this false belief be eradicated from the minds of researchers.

A second false belief widely held by researchers is that statistical significance level provides an index of the importance or size of a difference or relation (Bakan, 1966). A difference significant at the .001 level is regarded as theoretically (or practically) more important or larger than a difference significant at only the .05 level. In research reports in the literature, one sees statements such as the following: "Moreover, this difference is highly significant ($p < .001$)," implying that the difference is therefore large or important. This belief ignores the fact that significance level depends on sample size; highly significant differences in large sample studies may be smaller than even nonsignificant differences in smaller sample studies. This belief also ignores the fact that even if sample sizes were equal across studies compared, the p values would still provide no index of the actual size of the difference or effect. Only effect size indices can do that.

Because of the influence of meta-analysis, the practice of computing effect sizes has become more frequent in some research literatures, thus mitigating the pernicious effects of this false belief. But in other areas, especially in many areas of experimental psychology, effect sizes are rarely computed, and it remains the practice to infer size or importance of obtained findings from statistical significance levels. In an empirical study, Oakes (1986, pp. 86–88) found that psychological researchers infer grossly overestimated effect sizes from significance levels. When the study p values

[4] The statistical power for future replications of this study is estimated as follows. The best estimate of the population effect size is the effect size (d value) observed in this study. This observed d value is $2t/\sqrt{N}$ (Hunter & Schmidt, 1990b, p. 272), which is .85 here. With 20 subjects each in the experimental and control groups, an alpha level of .01 (two tailed), and a population d value of .85, the power of the t test is .43.

were .01, they estimated effect sizes as five times as large as they actually were.

The size or importance of findings is information important to researchers. Researchers who continue to believe that statistical significance levels reveal the size or importance of differences or relations will continue to refuse to abandon significance testing in favor of point estimates and confidence intervals. So this is a second false belief that must be eradicated.

The third false belief held by many researchers is the most devastating of all to the research enterprise. This is the belief that if a difference or relation is not statistically significant, then it is zero, or at least so small that it can safely be considered to be zero. This is the belief that if the null hypothesis is not rejected, then it is to be accepted. This is the belief that a major benefit from significance tests is that they tell us whether a difference or effect is real or "probably just occurred by chance." If a difference is not significant, then we know that it is probably just due to chance. The two examples discussed earlier show how detrimental this false operational decision rule is to the attainment of cumulative knowledge in psychology. This belief makes it impossible to discern the real meaning of research literatures.

Although some of his writings are ambiguous on this point, Fisher himself probably did not advocate this decision rule. In his 1935 book he stated,

> It should be noted that this null hypothesis is never proved or established, but is possibly disproved in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (p. 19)

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. Oakes (1986) has shown empirically that the operational decision rule used by researchers is indeed "if it is not significant, it is zero." Use of this decision rule amounts to an implicit belief on the part of researchers that the power of significance tests is perfect or nearly perfect. Such a belief would account for the surprise typically expressed by researchers when informed of the low level of statistical power in most studies.

The confidence of researchers in a research finding is not a linear function of its significance level. Rosenthal and Gaito (1963) studied the confidence that researchers have that a difference is real as a function of the p value of the significance test. They found a precipitous decline in confidence as the p value increased from .05 to .06 or .07. There was no similar "cliff effect" as the p value increased from .01 to .05. This finding suggests that researchers believe that any finding significant at the .05 level or beyond is real and that any finding with a larger p value—even one only marginally larger—is zero.

Researchers must be disabused of the false belief that if a finding is not significant, it is zero. This belief has probably done more than any of the other false beliefs about significance testing to retard the growth of cumulative knowledge in psychology. Those of us concerned with the development of meta-analysis methods hope that demonstrations of the sort given earlier in this article will effectively eliminate this false belief.

I believe that these false beliefs are a major cause of the addiction of researchers to significance tests. Many researchers believe that statistical significance testing confers important benefits that are in fact completely imaginary. If we were clairvoyant and could enter the mind of a typical researcher, we might eavesdrop on the following thoughts:

> Significance tests have been repeatedly criticized by methodological specialists, but I find them very useful in interpreting my research data, and I have no intention of giving them up. If my findings are not significant, then I know that they probably just occurred by chance and that the true difference is probably zero. If the result is significant, then I know I have a reliable finding. The p values from the significance tests tell me whether the relationships in my data are large enough to be important or not. I can also determine from the p value what the chances are that these findings would replicate if I conducted a new study. These are very valuable things for a researcher to know. I wish the critics of significance testing would recognize this fact.

Every one of these thoughts about the benefits of significance testing is false. I ask the reader to ponder this question: Does this describe your thoughts about the significance test?

Analysis of Costs and Benefits

We saw earlier that meta-analysis reveals clearly the horrendous costs in failure to attain cumulative

knowledge that psychology pays as the price for its addiction to significance testing. I expressed the hope that the appreciation of these massive costs will do what 40 years of logical demonstrations of the deficiencies of significance testing have failed to do: convince researchers to abandon the significance test in favor of point estimates of effect sizes and confidence intervals. But it seems unlikely to me that even these graphic demonstrations of costs will alone lead researchers to give up statistical significance testing. We must also consider the perceived benefits of significance testing. Researchers believe that significance testing confers important imaginary benefits. Many researchers may believe that these "benefits" are important enough to outweigh even the terrible costs that significance testing extracts from the research enterprise. It is unlikely that researchers will abandon significance testing unless and until they are educated to see that they are not getting the benefits they believe they are getting from significance testing. This means that quantitative psychologists and teachers of statistics and other methodological courses have the responsibility to teach researchers not only the high costs of significance testing but also the fact that the benefits typically ascribed to them are illusory. The failure to do the latter has been a major oversight for almost 50 years.

Current Situation in Data Analysis in Psychology

There is a fundamental contradiction in the current situation with respect to quantitative methods. The research literatures and conclusions in our journals are now being shaped by the results and findings of meta-analyses, and this development is solving many of the problems created by reliance on significance testing (Cooper & Hedges, 1994; Hunter & Schmidt, 1990b). Yet the content of our basic graduate statistics courses has not changed (Aiken et al., 1990); we are training our young researchers in the discredited practices and methods of the past. Let us examine this anomaly in more detail.

Meta-analysis has explicated the critical role of sampling error, measurement error, and other artifacts in determining the observed findings and the statistical power of individual studies. In doing so, it has revealed how little information there typically is in any single study. It has shown that, contrary to widespread belief, a single primary study can rarely resolve an issue or answer a question. Any individual study must be considered a data point to be contributed to a future meta-analysis. Thus the scientific status and value of the individual study is necessarily lower than has typically been imagined in the past.

As a result, there has been a shift of the focus of scientific discovery in our research literatures from the individual primary study to the meta-analysis, creating a major change in the relative status of reviews. Journals that formerly published only primary studies and refused to publish reviews are now publishing meta-analytic reviews in large numbers. Today, many discoveries and advances in cumulative knowledge are being made not by those who do primary research studies but by those who use meta-analysis to discover the latent meaning of existing research literatures. This is apparent not only in the number of meta-analyses being published but also—and perhaps more important—in the shifting pattern of citations in the literature and in textbooks from primary studies to meta-analyses. The same is true in education, social psychology, medicine, finance, accounting, marketing, and other areas (Hunter & Schmidt, 1990a, chap. 1).

In my own substantive area of industrial/organizational psychology there is even some evidence of reduced reliance on significance testing in analyses of data within individual studies. Studies are much more likely today than in the past to report effect sizes and more likely to report confidence intervals. Results of significance tests are usually still reported, but they are now often sandwiched into parentheses almost as an afterthought and are often given appropriately minimal attention. It is rare today in industrial/organizational psychology for a finding to be touted as important solely on the basis of its p value.

Thus when we look at the research enterprise being conducted by the established researchers of our field, we see major improvements over the situation that prevailed even 10 years ago. However—and this is the worrisome part—there have been no similar improvements in the teaching of quantitative methods in graduate and undergraduate programs. Our younger generations of upcoming researchers are still being inculcated with the old, discredited methods of reliance on statistical

significance testing. When we teach students how to analyze and interpret data in individual studies, we are still teaching them to apply *t* tests, *F* tests, chi-square tests, and ANOVAS. We are teaching them the same methods that for over 40 years made it impossible to discern the real meaning of data and research literatures and have therefore retarded the development of cumulative knowledge in psychology and the social sciences. We must introduce the reforms needed to solve this serious problem.

It will not be easy. At Michigan State University, John Hunter and Ralph Levine reformed the graduate statistics course sequence in psychology over the last 2 years along the general lines indicated in this article. The result was protests from significance testing traditionalists among the faculty. These faculty did not contend that the new methods were erroneous; rather, they were concerned that their graduate students might not be able to get their research published unless they used traditional significance testing–based methods of data analysis. They did not succeed in derailing the reform, but it has not been easy for these two pioneers. But this must be done and done everywhere. We can no longer tolerate a situation in which our upcoming generation of researchers are being trained to use discredited data analysis methods while the broader research enterprise of which they are to become a part has moved toward improved methods.

References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45,* 721–734.

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

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

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

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

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

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

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

Cooper, H. M., & Hedges, L. V. (1994). *Handbook of research synthesis.* New York: Russell Sage Foundation.

Cronbach, L. J. (1975). The two disciplines of scientific psychology revisited. *American Psychologist, 30,* 116–127.

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

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

Gergen, K. J. (1982). *Toward transformation in social knowledge.* New York: Springer-Verlag.

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

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

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

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

Huberty, C. J. (1993). Historical origins of statistical testing practices. *Journal of Experimental Education, 61,* 317–333.

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, NY.

Hunter, J. E., & Schmidt, F. L. (1990a). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75,* 334–349.

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

Hunter, J. E., & Schmidt, F. L. (in press). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law.*

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). Statistics and research design. *Annual Review of Psychology, 6,* 405–430.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review, 24,* 328–338.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review, 41,* 429–471.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48,* 1181–1209.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36,* 102–105.

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, CA.

Lykken, D. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

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

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

Neyman, J., & Pearson, E. S. (1932). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society, 29,* 492–516.

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, A231,* 289–337.

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

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

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

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55,* 33–38.

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., & Hunter, J. E. (1978). Moderator research and the law of small numbers. *Personnel Psychology, 31,* 215–232.

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

Schmidt, F. L., Hunter, J. E., & Urry, V. E. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473–485.

Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 38,* 509–524.

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

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

Received April 27, 1995
Revision received September 7, 1995
Accepted September 25, 1995 ■