Practical Significance: A Concept Whose Time Has Come

Roger E. Kirk

Educational and Psychological Measurement 1996 56: 746
DOI: 10.1177/0013164496056005002

PRACTICAL SIGNIFICANCE: A CONCEPT WHOSE TIME HAS COME

ROGER E. KIRK
Baylor University

Statistical significance is concerned with whether a research result is due to chance or sampling variability; practical significance is concerned with whether the result is useful in the real world. A growing awareness of the limitations of null hypothesis significance tests has led to a search for ways to supplement these procedures. A variety of supplementary measures of effect magnitude have been proposed. The use of these procedures in four APA journals is examined, and an approach to assessing the practical significance of data is described.

For almost 70 years, null hypothesis significance testing has been an integral part of the research enterprise in which behavioral and educational researchers engage. And for almost 70 years, null hypothesis significance testing has been surrounded by controversy. The acrimonious exchanges between Ronald Fisher and his adversaries, Jerzy Neyman and Egon Pearson, set the pattern for the debate that has continued to this day. By 1925, Fisher had worked out most of the ideas underlying hypothesis testing, including the theory of point estimation, consistency, efficiency, sufficiency, randomization, and maximum likelihood estimation. Three years later, Neyman and Pearson (1928) contributed the ideas of Type I and Type II errors and a predetermined level of significance, the final ingredients of present-day hypothesis testing.

This article is based on my presidential address delivered at the Southwestern Psychological Association meeting in Houston, TX, April 5, 1996. Appreciation is expressed to Robert J. Boik, Harvey Keselman, Joel R. Levin, Robert Rosenthal, and Charles Wilkins for their thoughtful comments on an earlier draft of this article. Correspondence concerning this article should be addressed to Roger E. Kirk, Department of Psychology, Baylor University, Waco, TX 76798-7334.

Educational and Psychological Measurement, Vol. 56 No. 5, October 1996 746-759

One of the earliest serious challenges to the logic and usefulness of null hypothesis significance testing appeared in a 1938 article by Joseph Berkson in the *Journal of the American Statistical Association*. Since then, there has been a crescendo of challenges (Carver, 1978; Cohen, 1990, 1994; Falk & Greenbaum, 1995; Guttman, 1985; Lykken, 1968; Meehl, 1967; Oakes, 1986; Rozeboom, 1960; Schmidt, 1996b). Some of the articles are included in the edited book by Morrison and Henkel (1970) titled *The Significance Test Controversy* and a book that I edited titled *Statistical Issues* (Kirk, 1972). An excellent, more recent examination of hypothesis testing is contained in the 1993 summer volume of the *Journal of Experimental Education* edited by Bruce Thompson. The one individual most responsible for bringing the shortcomings of hypothesis testing to the attention of behavioral and educational researchers is Jacob Cohen. His *American Psychologist* articles, "Things I have learned (so far)" and "The earth is round ($p < .05$)," are classics (Cohen, 1990, 1994).

What are the major criticisms of classical null hypothesis significance testing? Three criticisms are mentioned frequently. The first criticism is that the procedure doesn't tell researchers what they want to know. To put it another way, null hypothesis significance testing and scientific inference address different questions. In scientific inference, what we want to know is the probability that the null hypothesis (H_0) is true given that we have obtained a set of data (D); that is, $p(H_0|D)$. What null hypothesis significance testing tells us is the probability of obtaining these data or more extreme data if the null hypothesis is true, $p(D|H_0)$. Unfortunately for researchers, obtaining data for which $p(D|H_0)$ is low does not imply that $p(H_0|D)$ also is low. Researchers reason incorrectly that if the p value associated with a test statistic is suitably small, say, less than .05, the null hypothesis is probably false. This form of deductive reasoning has been referred to by Falk and Greenbaum (1995) as the "illusion of probabilistic proof by contradiction." Associated with this form of reasoning are the incorrect, widespread beliefs that (a) the p value is the probability that the null hypothesis is correct, and (b) the complement of the p value is the probability that a significant result will be found in a replication.

A second criticism of null hypothesis significance testing is that it is a trivial exercise. As John Tukey (1991) wrote, "the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'Are the effects different?' is foolish" (p. 100). Because the null hypothesis is always false, a decision to reject it simply indicates that the research design had adequate power to detect a true state of affairs, which may or may not be a large effect or even a useful effect. It is ironic that a ritualistic adherence to null hypothesis significance testing has led researchers to focus on controlling the Type I error that cannot occur because all null hypotheses are false while allowing the Type II error that can occur to exceed acceptable levels, often as high as .50 to .80 (Cohen, 1962, 1969, 1990, 1994).

A third criticism of null hypothesis significance testing is that by adopting a fixed level of significance, a researcher turns a continuum of uncertainty into a dichotomous reject-do-not-reject decision. The use of this decision strategy can lead to the anomalous situation in which two researchers obtain identical treatment effects but draw different conclusions from their research. One researcher, for example, might obtain a p value of .06 and decide to not reject the null hypothesis. The other researcher uses slightly larger samples and obtains a p value of .05, which leads to a rejection. What is troubling here is that identical treatment effects can lead to different decisions. The comment by Rosnow and Rosenthal (1989) is pertinent: "Surely, God loves the .06 nearly as much as the .05" (p. 1277). Another problem associated with the dichotomous decision rule is that some researchers mistakenly interpret a failure to reject the null hypothesis as evidence for accepting it.

Supplementing the Null Hypothesis Significance Test

These criticisms and others (see Carver, 1978; Cronbach, 1975; Oakes, 1986; Shulman, 1970) led quantitative psychologists to look for ways to supplement a null hypothesis significance test. As we have seen, the rejection of a null hypothesis is not very informative. We know in advance that the hypothesis is false. In spite of this, we compute a test statistic that enables us to specify the probability of obtaining a difference as large as or larger than that observed if the null hypothesis is true. If the p value is equal to or less than, say, .05, we conclude that we have obtained a result for which chance or sampling variability is an unlikely explanation, and we reject the null hypothesis. Notice that the emphasis is on rejecting the null hypothesis and the size of the p value. The emphasis should be on the data and whether the data support the scientific hypothesis. This is not a new idea. It was originally touched on by Karl Pearson in 1901 and more explicitly in 1925 by Ronald Fisher. Fisher (1925) proposed that researchers supplement the significance test in analysis of variance with the correlation ratio or eta, which measures the strength of the association between the independent and dependent variables. Since then, quantitative psychologists have proposed a variety of supplementary measures. I will use the term *effect magnitude* to refer to all such measures. The measures fall into one of three categories as shown in Table 1. The categories are (a) measures of strength of association, (b) measures of effect size (typically, standardized mean differences), and (c) other measures. Forty measures are listed in Table 1. As we will see, a survey of four APA journals found that only two of the measures are used frequently.

The idea of supplementing the null hypothesis significance test reappears from time to time. The idea received a major boost in 1940 when Peters and VanVoorhis, in their classic text, advocated reporting Kelley's epsilon, another measure of strength of association, with the analysis of variance F statistic. The reason they gave is that

Table 1
Measures of Effect Magnitude

Measures of Strength of Association	Measures of Effect Size	Other measures
r, r_{pb}, r^2, R, R^2, η, η^2, η^2_{mult}, ϕ	Cohen's (1988) d, f, g, h, q, w	Cohen's (1988) U_1, U_2, U_3
Cohen's (1988) f^2	Glass's (1976) g'	Logit d'
Contingency coefficient	Hedges's (1981) g	McGraw and Wong's (1992) common language effect size (CL)
Cramér's (1946) V	Rosenthal and Rubin's (1989) Π	Odds Ratio ($\hat{\omega}^2$)
Fisher's (1921) Z	Tang's (1938) ϕ	Preece's (1983) ratio of success rates
Hays's (1963) ω^2 and ρ_I		Probit d'
Kelley's (1935) ε^2		Relative risk
Kendall's (1963) W		Risk difference
Tatsuoka's (1973) $\hat{\omega}^2_{mult.\ c}$		Rosenthal and Rubin's (1982) binomial effect size display (BESD)
		Rosenthal and Rubin's (1994) counternull value of an effect size ($ES_{counternull}$)

> the F and z tests employed with analysis of variance do not directly indicate the strength of the relation that is present, but only its reliability. . . . Epsilon on the other hand, shows in language with a uniform meaning what is the strength of the relation that is present. (p. 353)

Twenty-six years later, Glass and Hakstian (1969) observed that

> periodically, researchers have been reminded that test statistics (e.g., t-ratios, F-ratios) serve only to indicate the inferential stability (statistical significance) of observed results. Various measures of association have been developed over the years to address the question of the strength of relationship between the independent and dependent variables in comparative experiments. (p. 403)

Three measures of strength of association between a categorical variable and a continuous variable are shown in Table 2. Kelley (1935) developed the epsilon squared measure to correct for the positive bias in eta squared. Neither epsilon squared nor Hays's (1963) omega squared that appeared in 1963 are unbiased estimators, and neither captured the fancy of researchers like d, introduced by Cohen (1969). Cohen's d was the first effect size measure that was explicitly labeled as such. As the formula in Table 2 shows, d expresses the size of the population treatment effect in units of the common population standard deviation. What made Cohen's contribution unique is that he provided guidelines for interpreting the magnitude of d. According to Cohen (1992), a medium effect of .5 is visible to the naked eye of a careful observer.

Table 2
Early Measures of Effect Magnitude

Measures of Strength of Association		
Fisher (1925)	$\hat{\eta}^2 = \frac{SSBG}{SSTO}$	
Kelley (1935)	$\hat{\varepsilon}^2 = \frac{SSBG - (p-1)\,MSWG}{SSTO}$	
Hays (1963)	$\hat{\omega}^2 = \frac{SSBG - (p-1)\,MSWG}{SSTO + MSWG}$	
Measures of Effect Size		
Cohen (1969)	$d = \frac{\mu_1 - \mu_2}{\sigma}$	$\begin{cases} d = .2 \text{ is a small effect} \\ d = .5 \text{ is a medium effect} \\ d = .8 \text{ is a large effect} \end{cases}$
Glass (1976)	$g' = \frac{\bar{Y}_E - \bar{Y}_C}{S_C}$	
Hedges (1981)	$g = \frac{\bar{Y}_E - \bar{Y}_C}{S_{\text{Pooled}}}$	

Several surveys have found that .5 approximates the average size of observed effects in various fields (Cooper & Findley, 1982; Haase, Waechter, & Solomon, 1982; Sedlmeier & Gigerenzer, 1989). A small effect of .2 is noticeably smaller than medium but not so small as to be trivial. A large effect of .8 is the same distance above medium as small is below it. These operational definitions turned his measure of effect size into a much more useful statistic. For the first time, researchers had guidelines for interpreting the size of treatment effects. But the usefulness of *d* did not stop there. The *d* parameter could be used to estimate the sample size necessary to detect small, medium, and large effects and to assess the power of a research design to detect various size effects. Cohen extended his work by developing guidelines for interpreting correlation coefficients, regression coefficients, differences between correlation coefficients, proportions, differences between proportions, contingency table data, and differences among means in analyses of variance. The meaning of small, medium, and large effects remained approximately the same across the various measures of effect size.

The effect size concept was used by Gene Glass (1976) in his pioneering work on meta-analysis. However, as shown in Table 2, he used a sample analogue of *d* in which the population standard deviation was replaced by the sample standard deviation of the control group. He reasoned that if there were several experimental groups, pairwise pooling of the standard deviations would result in a different standard deviation for each experimental-control contrast. Hence, the same size difference between experimental and control

Table 3
Conversion Formulas for Measures of Effect Magnitude

	Two-Sample Case		Multisample Case	
	1	2	3	4
1.	$g = \sqrt{\frac{df(n_1+n_2)r_{pb}^2}{n_1n_2+(1-r_{pb}^2)}}$	$= \frac{2t}{\sqrt{n_1+n_2}}$	$\hat{f} = \sqrt{\frac{\hat{\omega}^2}{1-\hat{\omega}^2}}$	$= \sqrt{\frac{df_{BG}(F-1)}{N}}$
2.	$r_{pb} = \sqrt{\frac{g^2n_1n_2}{g^2n_1n_2+(n_1+n_2)\,df}}$	$= \frac{t}{\sqrt{t^2+df}}$	$\hat{\omega}^2 = \frac{\hat{f}^2}{1+\hat{f}^2}$	$= \frac{df_{BG}(F-1)}{df_{BG}(F-1)+N}$
3.	$d = .2$ $\rho_{pb} = .10$ is a small effect[a] $d = .5$ $\rho_{pb} = .24$ is a medium effect $d = .8$ $\rho_{pb} = .37$ is a large effect		$f = .10$ $\omega^2 = .010$ is a small effect[a] $f = .25$ $\omega^2 = .059$ is a medium effect $f = .40$ $\omega^2 = .138$ is a large effect	

a. ρ_{pb} denotes the population point biserial correlation coefficient.

means would result in different effect size values when the standard deviations of the contrasts differed. Larry Hedges (1981) had a different solution to this problem. He pooled the standard deviations of the experimental groups with that for the control group to obtain one standard deviation for all contrasts. His pooled population estimator, shown in Table 2, is identical to the usual within-groups mean square in analysis of variance.

Two categories of effect magnitude have been described thus far. The leading researchers in this and related areas differ in their preferences for the various measures. Fortunately, it is a simple matter to convert from one category to the other. Several examples are shown in Table 3. In row 1, column 1, for example, Hedges's g is expressed as a function of the point biserial correlation coefficient, r_{pb}. Row 2, column 1 shows that the point biserial correlation coefficient can be expressed as a function of Hedges's g. Guidelines are available for interpreting each of the measures of effect magnitude as shown in row 3. Columns 2 and 4 of Table 3 show that each of the measures of effect magnitude can be computed if a researcher knows the value of the test statistic, say t or F, its degrees of freedom, and the sample sizes: information that should be contained in any research report.

Quantitative psychologists continue to search for ways to supplement the null hypothesis significance test. Most of the attention has focused on measures of strength of association and effect size. However, a variety of other measures have been proposed. A number of these measures are listed in the Other Measures column of Table 1. From an examination of the literature, I have concluded that none of these measures has much appeal to researchers in psychology and education.

Table 4
Percentage of Journal Articles With One or More Measures of Effect Magnitude

Journal	Number of Articles That Used an Inferential Statistic	Percentage With 0 Measures	Percentage With 1 Measure	Percentage With 2 Measures	Percentage With ≥ 3 Measures
Journal of Applied Psychology	57	23	47	21	9
Journal of Educational Psychology	49	45	27	20	8
Journal of Experimental Psychology, Learning & Memory	111	88	11	1	0
Journal of Personality and Social Psychology	174	53	28	9	9

Reporting Measures of Effect Magnitude in the Literature

As we have seen, for more than 70 years, researchers have been encouraged to supplement reports of null hypothesis significance tests with measures of effect magnitude (Brewer, 1978; Cohen, 1988; Fisher, 1925; Fleiss, 1969; Rosenthal, 1978). Are researchers following this advice? To answer this question, the 1995 volumes of four APA journals were examined. The results of the examination are shown in Table 4. The table gives the number of articles that used inferential statistics and the percentage of these articles that contained 0, 1, 2, and 3 or more measures of effect magnitude. There is considerable variability among the journals: 77% of the articles in the *Journal of Applied Psychology* contained one or more measures of effect magnitude; the comparable figure for the *Journal of Experimental Psychology* was only 12%. Before anyone concludes that authors of articles in the *Journal of Applied Psychology* are more aware of the limitations of null hypothesis significance testing, remember that these authors are more likely to use regression and correlation procedures. Computer packages routinely provide R^2 for these procedures. Authors in the *Journal of Experimental Psychology* are more likely to use analysis of variance procedures. Computer packages do not routinely provide measures of effect magnitude for these procedures.

The three most frequently used inferential procedures in the four journals were analysis of variance, the *t* test for means, and regression analysis. The average number of inferential tests per article was similar for the four journals. The *Journal of Experimental Psychology, Learning and Memory* had the fewest number per article, 2.3; the *Journal of Educational Psychology* had the most inferential tests per article, 3.6.

Table 5
Reporting Frequency for Measures of Effect Magnitude

Measure	*Journal of Applied Psychology*	*Journal of Educational Psychology*	*Journal of Experimental Psychology, Learning & Memory*	*Journal of Personality and Social Psychology*	Σ
Variance-accounted-for	19	21	4	43	87
R^2	25	14	3	30	72
r	0	2	3	24	29
R	4	4	0	14	22
$\hat{\eta}^2$	6	3	0	3	12
d, g, g'	3	2	2	4	11
r_{pb}	1	0	0	6	7
$\hat{\rho}_I$	1	0	0	5	6
$\hat{\omega}_2$	4	1	0	0	5
$\hat{\eta}$	0	2	0	2	4
r^2	1	0	1	1	3
Odds ratio	1	0	1	1	3
Kendall's W	1	0	0	1	2
$\hat{\rho}_I^2$	0	0	0	1	1
Relative risk	0	0	0	1	1
Tang's $\hat{\Phi}$	1	0	0	0	1

What measures of effect magnitude do researchers report? A summary of the measures in the four journals is shown in Table 5. The general category "variance-accounted-for" was used whenever an article used the phrase "variance accounted for" and did not specify a particular statistic such as R^2, eta squared, and so on. The failure to identify the statistic used to determine the variance-accounted-for is only one of many examples of sloppy reporting in the literature. The second most frequently cited measure is R^2. Together, variance-accounted-for and R^2 represented 60% of the 16 measures that were used. This is not surprising considering that regression programs in stat packages always report R^2. Measures that are not routinely provided in statistics packages such as Hedges's g and Hays's $\hat{\omega}^2$ rarely appeared in the four journals.

Practical Significance, an Alternative

As we have seen, the null hypothesis significance test is often misinterpreted. One response to this unfortunate state of affairs is to admonish researchers to clean up their act, start interpreting significance tests correctly, and get on with the business of science. I believe that even when a significance test is interpreted correctly, the business of science does not progress

as it should. This is not a new observation. Critics have been saying it for years. For example, Frank Yates (1951), a contemporary of Fisher, observed that the use of the null hypothesis significance test

> has caused scientific research workers to pay undue attention to the results of the tests of significance that they perform on their data and too little attention to the estimates of the magnitude of the effects they are investigating. . . . The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers often have regarded the execution of a test of significance on an experiment as the ultimate objective. (pp. 32-33)

A more strongly worded criticism of null hypothesis significance testing was written by Paul Meehl (1978):

> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (p. 817)

The criticism that an emphasis on null hypothesis significance tests detracts researchers from the main business of science—interpreting the outcome of research, theory development, and so on—is shared by many researchers (Cohen, 1994; Dar, 1987; Schmidt, 1996b; Thompson, 1996; Tukey, 1991).

What does a researcher learn from a failure to reject the null hypothesis? Because all null hypotheses are false, John Tukey (1991) observed that a nonrejection simply means that the researcher is unable to specify the direction of the difference between the conditions. On the other hand, a rejection means that the researcher is pretty sure of the direction of the difference. Is this any way to develop psychological theory? I think not. How far would physics have progressed if their researchers had focused on discovering ordinal relationships? What we want to know is the size of the difference between A and B and the error associated with our estimate; knowing that A is greater than B is not enough.

The computation of a point estimate of the difference between A and B and a confidence interval for that difference requires no more information than a null hypothesis significance test. A confidence interval contains all of the information provided by a significance test and, in addition, provides a range of values within which the true difference is likely to lie. It is important to understand that a confidence interval is just as useful as a null hypothesis significance test for deciding whether chance or sampling variability is an unlikely explanation for an observed difference. Furthermore, a point estimate and confidence interval use the same unit of measurement as the data. This facilitates the interpretation of results and makes trivial effects harder to ignore. However, in spite of the superiority of confidence intervals, they rarely appear in psychology and education journals. What we see is a reject-

nonreject decision strategy that does not tell us what we want to know and a preoccupation with p values that are several steps removed from examining the data.

Consider a researcher who believes that a medication will improve the intelligence test performance of Alzheimer patients. She randomly assigns 12 patients to experimental and control groups and administers the medication to the experimental group and a placebo to the control group. In due time, she administers an intelligence test to the patients and computes a t test, $t(10) = 1.61$, $p = .14$. To her dismay, the p value is larger than .05, which means that the null hypothesis cannot be rejected. What's wrong with this typical scenario? The researcher focused on the null hypothesis and p value without asking whether the data supported her scientific hypothesis. Unfortunately, a result that is not statistically significant is interpreted as providing no support for the scientific hypothesis, even though the data are consistent with the hypothesis. Suppose that the mean for the experimental group is 13 IQ points above that for the control group. This information should make any rational researcher think that the data provides some support for the scientific hypothesis. In fact, the best guess that can be made is that the population mean difference is 13 IQ points. A 95% confidence interval for the population mean difference indicates that it is likely to be between –6.3 and 32.3 IQ points. The nonsignificant t test does not mean that there is no difference between the IQs; all it means is that the researcher cannot rule out chance or sampling variability as an explanation for the observed difference.

The appeal of null hypothesis significance testing is that it is considered to be an objective, scientific procedure for advancing knowledge. In fact, focusing on p values and rejecting null hypotheses actually distracts us from our real goals: deciding whether data support our scientific hypothesis and are practically significant or useful. For measuring scales that are familiar, such as the IQ scale, a point estimate of a difference and confidence interval can be used to decide whether results are trivial, useful, or important. It is true that an element of subjectivity is introduced into the decision process when researchers make this kind of judgment. And the judgment inevitably involves a variety of considerations, including the researcher's value system, societal concerns, costs and benefits, and so on. However, I believe that researchers have an obligation to make this kind of judgment. No one is in a better position than the researcher who collected and analyzed the data to decide whether or not the results are trivial. It is a curious anomaly that researchers are trusted to make a variety of complex decisions in the design and execution of an experiment, but in the name of objectivity, they are not expected or even encouraged to decide whether data are practically significant.

Decisions regarding scientific hypotheses and practical usefulness are less straightforward when a measuring scale involves unfamiliar units. In such cases, it is necessary to (a) compute an effect magnitude and a confidence

interval for that effect magnitude and (b) develop guidelines for deciding whether the effect magnitude is useful. A variety of measures of effect magnitude are available to researchers. And considerable progress has been made in developing unbiased estimators of effect magnitudes and associated confidence intervals. For example, Hedges and Olkin (1985) have derived an unbiased estimator of *d* and an exact confidence interval for the estimator. Similar results have been obtained for other effect magnitude parameters (Fleishman, 1980; Fowler, 1985). With respect to determining the practical significance of results, Cohen's definitions of small, medium, and large effects represent a good beginning. However, much more systematic research is needed to extend his work. Certainly, the task of scaling practical significance is no more difficult than scaling other variables in psychology and education. It is important to not sanctify effect size numbers such as .2, .5, and .8 as has been done with the .05 and .01 levels of significance. If practical significance is to be a useful concept, its determination must not be ritualized.

Let's return to the Alzheimer experiment. Recall that the IQ of the experimental group was 13 points higher than that of the control group. Following Hedges and Olkin (1985, pp. 81-91), an unbiased estimate of Cohen's *d* is .86, which suggests that the difference represents a large effect. Anyone who has worked with intelligence tests probably would agree that 13 IQ points is a large effect. An exact 95% confidence interval for our .86 estimate is from –.3 to 2.0. As we have seen, the data provide considerable support for the researcher's scientific hypothesis, although she cannot rule out chance sampling variability as a possible explanation for the difference. Will the results replicate? Are they real? There is only one way to find out: Do a replication. Does the medication appear to have promise with Alzheimer patients? I think so. Notice the difference in our reasoning process when we shift attention from the *t* test and *p* value to deciding whether the data support our scientific hypothesis and are useful. Our science has paid a high price for its ritualistic adherence to null hypothesis significance testing.

APA Review of Null Hypothesis Significance Testing

In spite of repeated criticisms of null hypothesis significance testing, the procedure continues to dominate psychological and educational research. According to Schmidt (1996a), there is reason to believe that the situation will change. The APA Board of Scientific Affairs recently appointed a task force to study the desirability of phasing out the use of null hypothesis significance testing in course texts, journal articles, and so on. The board, which is seeking the involvement of AERA, APS, Division 5, the Society for Mathematical Psychology, and the American Statistical Association, appears to be very receptive to the idea of doing away with null hypothesis significance testing. If such a recommendation ultimately comes from the task force, the change could be phased in over several years by changing the instructions

to authors in psychology and education journals. This change would cause a chain reaction: Statistics teachers would change their courses, textbook authors would revise their statistics books, and journal authors would modify their inference strategies. The winds of change are about us. Many researchers share the belief that if our science is to progress as it should, we must get over our obsession with null hypothesis significance tests and focus on the practical significance of our data. The appointment of the task force may mark the beginning of a more enlightened approach to the interpretation of data.

References

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-542.

Brewer, J. K. (1978). Effect size: The most troublesome of the hypothesis testing considerations. *Phi Delta Kappa, 11*, 7-10.

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. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

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

Cooper, H., & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 8*, 168-173.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Cronbach, L. J. (1975). Beyond two disciplines of scientific psychology. *American Psychologist, 30*, 116-127.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologists, 42*, 145-151.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*, 75-98.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1*, 1-32.

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

Fleishman, A. I. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement, 40*, 659-670.

Fleiss, J. L. (1969). Estimating the magnitude of experimental effects. *Psychological Bulletin, 72*, 273-276.

Fowler, R. L. (1985). Point estimates and confidence intervals in measures of association. *Psychological Bulletin, 98*, 160-165.

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

Glass, G. V., & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal, 6*, 403-414.

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

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29*, 58-65.

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

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

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

Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences, 21*, 554-559.

Kendall, M. G. (1963). *Rank correlation methods* (3rd ed.). London: Griffin.

Kirk, R. E. (Ed.). (1972). *Statistical issues*. Monterey, CA: Brooks/Cole.

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

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361-365.

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 progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 29A*, Part I: 175-240; Part II: 263-294.

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

Pearson, K. (1901). On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, 195*, 1-47.

Peters, C. C., & VanVoorhis, W. R. (1940). *Statistical procedures and their mathematical bases*. New York: McGraw-Hill.

Preece, P.F.W. (1983). A measure of experimental effect size based on success rates. *Educational and Psychological Measurement, 43*, 763-766.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*, 185-193.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106*, 332-337.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5*, 329-334.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

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

Schmidt, F. (1996a). APA Board of Scientific Affairs to study issue of significance testing, make recommendations. *Score, 19*, 1, 6.

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

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

Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40*, 371-393.

Tang, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. *Statistics Research Memorandum, 2*, 126-149.

Tatsuoka, M. M. (1973). *An examination of the statistical properties of a multivariate measure of strength of association.* (Final Report to U.S. Office of Education on Contract No. OEG-5-72-0027.)

Thompson, B. (Ed.). (1993). Statistical significance testing in contemporary practice: Some proposed alternatives with comments from journal editors. *Journal of Experimental Education, 61*(4).

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*, 26-30.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Yates, F. (1951). The influence of "statistical methods for research workers" on the development of the science of statistics. *Journal of the American Statistical Association, 46*, 19-34.