RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 3-14

The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing

Thomas W. Nix
University of Alabama

J. Jackson Barnette
University of Iowa

Null Hypothesis Significance Testing (NHST) is reviewed in a historical context. The most vocal criticisms of NHST that have appeared in the literature over the past 50 years are outlined. The authors conclude, based on the criticism of NHST and the alternative methods that have been proposed, that viable alternatives to NHST are currently available. The use of effect magnitude measures with surrounding confidence intervals and indications of the reliability of the study are recommended for individual research studies. Advances in the use of meta-analytic techniques provide us with opportunities to advance cumulative knowledge, and all research should be aimed at this goal. The authors provide discussions and references to more information on effect magnitude measures, replication techniques and meta-analytic techniques. A brief situational assessment of the research landscape and strategies for change are offered.

It is generally accepted that the purpose of scientific inquiry is to advance the knowledge base of humankind by seeking evidence of a phenomena via valid experiments. In the educational arena, the confirmation of a phenomena should give teachers confidence in their methods and policy makers confidence that their policies will lead to better education for children and adults. We approach the analysis of experimentation with the tools of statistics, more specifically, descriptive and inferential statistics. Little controversy surrounds the use of descriptive statistics to mirror the various states of nature, however the use of inferential statistics has a long and storied history. Today, there are at least four different schools of thought on inferential significance testing. They are the Fisherian approach, the Neyman-Pearson school, Bayesian Inference, and Likelihood Inference. A full description of each is beyond the scope of this paper, but a complete evaluation of each has been detailed by Oakes (1986). It is fair to state that not one of these inferential statistical methods is without controversy.

We first review the two most popular inferential approaches, the Fisherian and Neyman-Pearson schools, or what has come to be called null hypothesis significance testing (NHST). We then outline some of the major points found in critiques of NHST. Thirdly, we review the changing face of social science research with short primers on effect magnitude measures, meta-analytic methods, and replication techniques. Next, we assess how the development of these methods is coming face-to-face with the shortcomings of NHST. We outline how the primary researcher working on a single study of a phenomena can report more informative information using the same data now used for NHST and at the same time provide his/her study as the raw material for secondary research to be used by a meta-analytic researcher. We conclude with an assessment of the current situation and how change could be facilitated. Through this interchange of ideas and analysis, we can bring some order to what appears to be a chaotic world where the advancement of cumulative knowledge is slowed by a lack of information provided by NHST, misunderstandings about the meaning of NHST results, frustration with conflicting results, and bias in publication policies. Signals in the environment seem to indicate that discussions regarding whether NHST should be banned or not no longer seem to be germane. Rather, the informed stakeholders in the social sciences seem to be abandoning NHST, and with some guidance, we believe the transition to more enlightened statistical methods could be accomplished with minimal disruption.

Thomas W. Nix is a doctoral candidate at the University of Alabama. J. Jackson Barnette is associate professor of Preventive Medicine, Divisions of Community Health and Biostatistics, College of Medicine, University of Iowa. Correspondence regarding this article should be addressed to Thomas W. Nix, 700 Whippoorwill Drive, Birmingham, AL 35244 or by e-mail to tnix@bamaed.ua.edu.

Development of Null Hypothesis Significance Testing

To better understand how NHST achieved its status in the social sciences, we review its development. Most who read recent textbooks devoted to statistical methods are inclined to believe statistical significance testing is a unified, non-controversial theory whereby we seek to reject the null hypothesis in order to provide evidence of the viability of the alternative hypothesis. A *p*-value and an alpha level (α) are provided to determine the probability of the evidence being due to chance or sampling error. We also accept the fact there are at least two types of errors that can be committed in this process. If we reject the null hypothesis, a type I error, or a false positive result, can occur, and if we do not reject the null hypothesis, a type II error, or a false negative result, can occur. Most texts imply NHST is a unified theory that is primarily the work of Sir Ronald Fisher and that it has been thoroughly tested and is above reproach (Huberty, 1993). Nothing could be further from the truth.

The theory of hypothesis testing is not a unified theory at all. Fisher proposed the testing of a single binary null hypothesis using the *p*-value as the strength of the statistic. He did not develop or support the alternative hypotheses, type I and type II errors in significance testing, or the concept of statistical power. Jerzy Neyman, a Polish statistician, and Egon Pearson, son of Karl Pearson, were the originators of these concepts. In con-trast to Fisher's notion of NHST, Pearson and Neyman viewed significance testing as a method of selecting a hypothesis from a slate of candidate hypotheses, rather than testing of a single hypothesis.

Far from being in agreement with the theories of Neyman and Pearson, Fisher was harshly critical of their work. Although Fisher had many concerns about the work of Neyman and Pearson, a major concern centered around the way Neyman and Pearson used manufacturing acceptance decisions to describe what they saw as an extension of Fisher's theory. Fisher was adamant that hypothesis testing did not involve final and irrevocable decisions, as implied by the examples of Neyman and Pearson. However, his criticism was not always sparked by constructive scientific debate. Earlier in Fisher's career, he bitterly feuded with Karl Pearson while Pearson was the editor of the prestigious journal, *Biometrika* (Cohen, 1990). In fact, the rift became so great, Pearson refused to publish Fisher's articles in *Biometrika*. Although Neyman and the younger Pearson attempted to collaborate with Fisher after the elder Pearson retired, the acrimony continued from the 1930's until Fisher's death in July, 1962 (Mulaik, Raju, & Harshman, 1997).

Huberty's (1993) review of textbooks outlines the evolution of these two schools of thought and how they came to be perceived as a unified theory. He found that in the 1930s, writers of statistics textbooks began to refer to Fisher's methods, while a 1940 textbook was the first book in which the two types of error are identified and discussed. It was not until 1949 that specific references to Neyman and Pearson contributions were listed in textbooks, in spite of the fact that Neyman and Pearson's work was contemporary to that of Fisher. By 1950, the two separate theories began to be unified in textbooks but without the consent or agreement of any of the originators. By the 1960's the unified theory was accepted in a number of disciplines including economics, education, marketing, medicine, occupational therapy, psychology, social research, and sociology. At the end of the 1980s, NHST, in its unified form, had become so ubiquitous that over 90% of articles in major psychology journals justified conclusions from data analysis with NHST (Loftus, 1991).

Objections to Null Hypothesis Statistical Testing (NHST)

Criticism of NHST provides much evidence that it is flawed and misunderstood by the many who routinely use it. It has even been suggested that dependence on NHST has retarded the advancement of scientific knowledge (Schmidt, 1996b). Objections to NHST began in earnest in the early 1950s as NHST was gaining acceptance. While reviewing the accomplishments in statistics in 1953, Jones (1955) said, "Current statistical literature attests to increasing awareness that the usefulness of conventional hypothesis testing methods is severely limited" (p. 406). By 1970, an entire book was devoted to criticism of NHST in wide ranging fields such as medicine, soci-ology, psychology, and philosophy (Morrison & Henkel, 1970). Others, including Rozeboom (1960), Cohen (1962), Bakan (1966), Meehl (1978), Carver (1978), Oakes (1986), Cohen (1994),Thompson (1995, Novem-ber) and Schmidt (1996a), have provided compelling evidence that NHST has serious limiting flaws that many educators and researchers are either unaware of or have chosen to ignore. Below, we examine some of the often quoted arguments. They relate to: a) the meaning of the null hypothesis, b) the concept of statistical power, c) sample size dependence, and d) misuse of NHST information.

The Concept of a Null Hypothesis

In traditional NHST, we seek to reject the null hypothesis (H_o) in order to gain evidence of an

alternative or research hypothesis (H_a). The null hypothesis has been referred to as the hypothesis of no relationship or no difference (Hinkle, Wiersma, & Jurs, 1994). It has been argued that, only in the most rare of instances, can we fail to reject the hypothesis of no difference (Cohen, 1988; Meehl, 1967, 1978). This statement has merit when we consider that errors can be due to treatment differences, measurement error and sampling error. Intuitively, we know that in nature it is extremely rare to find two identical cases of anything. The test of differences in NHST posits an almost impossible situation where the null hypothesis differences will be exactly zero. Cohen points out the absurdity of this notion when he states, ". . . things get downright ridiculous when . . . (the null hypothesis). . . (states) that the effect size is 0, that the proportion of males is .5, that the rater's reliability is 0" (Cohen, 1994). Others have pointed out, "A glance at any set of statistics on total populations will quickly confirm the rarity of the null hypothesis in nature" (Bakan, 1966). Yet we know that there are tests where the null hypothesis is not rejected. How can this happen given the situation described above? To understand this we turn to the problems associated with statistical power, type I errors, and type II errors in NHST.

Type I Errors, Type II Errors, and Statistical Power

Neyman and Pearson provided us with the two types of errors that occur in NHST. They are type I errors or errors that occur when we indicate the treatment was effective when it was not (a false positive) and type II errors or errors that occur when we indicate there was no treatment effect when in fact there was (a false negative). The probability of a type I error is the level of significance or alpha (α). That is, if we choose a .05 level of significance, the probability of a type I error is .05. The lower the value we place on alpha, for example .01, the more exact the standard for acceptance of the null hypothesis and the lower the probability of a type I error. However, all things being equal, the lower the probability of a type I error, the lower the power of the test.

Power is the probability that a statistical test will find statistical significance (Rossi, 1997, p. 177). As such, moderate power of .5 indicates one would have only a 50% chance of obtaining a significant result. The comple-ment of power (1 - power), or beta (β), is the type II error rate in NHST. Cohen (1988, p. 5) pointed out the weight-ing procedure the researcher must consider prior to a null hypothesis test. For example, if alpha is set at .001, the risk of a type I error is minuscule, but the researcher may reduce the power of the test to .10, thereby setting the risk of a type II error at (1 - .10) or .90! A power level of .10, as in the previous example, would mean the researcher had only a 10% chance of obtaining significant results.

Many believe the emphasis on type I error control used in popular procedures such as the analysis of variance follow up tests and the emphasis on teaching the easier concept of type I errors may have contributed to the lack of power we now see in statistical studies. One only needs to turn to the popular Dunn-Bonferroni, Scheffé, Tukey, and Newman-Keuls follow up procedures in the analysis of variance to see examples of attempts to stringently control type I errors. However, when type I errors are stringently controlled, the price that is paid is a lack of control of the inversely related type II error, lowered test power, and less chance of obtaining a significant result.

How much power do typical published studies have? Cohen (1962) was one of the first to point out the problem of low power when he reviewed 78 articles appearing in the 1960 *Journal of Abnormal and Social Psychology.* He found the mean power value of studies, assuming a medium effect size, was only .48, where effect size is the degree to which a phenomenon exists in a study. This finding indicated the researchers had slightly less than a 50 - 50 chance of rejecting the null hypothesis. For stud-ies with small effects the odds were lower, and only when authors had large effects did they have a good chance, approximately 75%, of rejecting the null hypothesis.

With this information in hand, one would suspect researchers would be more cognizant of the power of their studies. However, when Sedlmeier and Gigerenzer (1989) replicated Cohen's study by reviewing 1984 articles, they found that the mean power of studies had actually declined from .48 to .37. It should be noted that Cohen's original methodology, used in these power studies, uses sample size and Cohen's definitions of large, medium, and small effects size to determine power rather than actual effect size (Thompson, 1998). As a result, the outcomes of these studies have been questioned. Nevertheless, they do point out the fact that decades of warnings about low power studies had done nothing to increase the power of studies.

One can only speculate on the damage to cumulative knowledge that has been cast upon the social sciences when study authors have only approximately a 50% chance of rejecting the null hypothesis and getting significant results. If the author does not obtain significant results in his/her study, the likelihood of being published is severely diminished due to the publication bias that exists for statistically significant results (Begg, 1994). As

a result there may be literally thousands of studies with meaningful effect sizes that have been rejected for publication or never submitted for publication. These studies are lost because they do not pass muster with NHST. This is particularly problematic in educational research where effect sizes may be subtle but at the same time may indicate meritorious improvements in instruction and other classroom methods (Cohen, 1988).

Sample Size Dependence

The power of a statistical test, or how likely the test is to detect significant results, depends not only on the alpha and beta levels but also on the reliability of the data. Reliability is related to the dispersion or variability in the data, and as a result it can be controlled by reducing measurement and sampling error. However, the most common way of increasing reliability and increasing the power of a test is to increase the sample size.

With increased sample size, we incur yet another problem, that is the sample size dependency of tests used in NHST. Bakan (1966) reported on the results of a battery of tests he had collected on 60,000 subjects in all parts of the United States. When he conducted significance tests on these data, he found that every test yielded significant results. He noted that even arbitrary and nonsensical divisions, such as east of the Mississippi versus west of the Mississippi and Maine versus the rest of the country, gave significant results. "In some instances the differences in the sample means were quite small, but nonetheless, the *p* values were all very low" (p. 425). Nunnally (1960) reported similar results using correlation coefficients on 700 subjects and Berkson (1938) found similar problems using a chi-square test. Berkson stated, ". . . we have something here that is apt to trouble the conscience of a reflective statistician . . . a large sample is always better than a small sample . . . (and) . . . if we know in advance the *p* will result from . . . a test of a large sample . . . (then) . . . there would seem to be no use in doing it on a smaller one . . . since the result . . . is known, it is no test at all" (p. 526). Therefore, a small difference in estimates of population parameters from large samples, no matter how insignificant, yields significant results.

Ironically, if we have low test power, we cannot detect statistical significance, but if we have high test power, via a large sample size, all differences, no matter how small, are significant. Schmidt (1996a) has pointed out a troubling problem associated with solving power problems with large sample sizes. He suggested that scientific inquiry can be retarded because many worthwhile research projects cannot be conducted, since the sample sizes required to achieve adequate power may be difficult, if not impossible, to attain. It is not unusual for the educational researcher to have to settle for smaller samples than desired. Therefore, it is not likely that educational studies can escape the bane of low power as long as NHST is the statistical tool used. But before we worry too much about power problems in NHST, perhaps we should consider the thoughts of Oakes (1986) and later Schmidt (1996a). Schmidt noted that the power of studies "is a legitimate concept only within the context of statistical significance testing . . . (and) . . . if significance testing is no longer used, then the concept of statistical power has no place and is not meaningful" (p. 124).

Misunderstanding of p Values

With the advent of easy to use computer programs for statistical analysis, the researcher no longer has to depend on tables and the manual procedures for NHST, instead computerized statistical packages provide the researcher with a *p* value that is used to determine whether we reject, or fail to reject, the null hypothesis. As such, *p* values lower than the alpha value are viewed as a rejection of the null hypothesis, and *p* values equal to or greater than the alpha value are viewed as a failure to reject. The *p* value tells us nothing about the magnitude of significance nor does it tell us anything about the probability of replication of a study. The *p* value's use is limited to either rejecting or failing to reject the null hypothesis. It says nothing about the research or alternative hypothesis (Carver, 1978). The *p* value is primarily a function of effect size and sampling error (Carver, 1993). Therefore, differences of even trivial size can be judged to be statistically significant when sampling error is small (due to a large sample size and/or a large effect size) or when sampling error is large (due to a small sample size and/or a small effect size). However, NHST does not tell us what part of the significant differences is due to effect size and what part is due to sampling error.

The easy access to *p* values via statistical software has led in some instances to misunderstanding and misuse of this information. Since many researchers focus their research on *p* values, confusion about the meaning of a *p* value is often revealed in the literature. Carver (1978) and Thompson (1993), among others, have indicated that users of NHST often misinterpret the meaning of a *p* value as being a magnitude measure. This is evidenced by such common phrases, as "almost achieving significance" and "highly significant" (Carver, 1978, p. 386). They right-fully point out that many textbooks make the same mistake and that some textbooks have gone one step further by implying that a

statistically significant *p* value indicates the probability that the results can be replicated. This is evidenced in statements such as "reliable dif-ference" or the "results were reliable" (Carver, 1978, p. 385). No part of the logic of NHST implies this.

Thompson (1995, November) has noted that many researchers use the *p* value as a vehicle to "avoid judgment" (p. 10). He implies that when a significant result is obtained, the analyst is generally provided with the confidence to conclude his/her analysis. The devotion to *p* values to determine if a result is statistically significant suspends further analysis. Analysis should continue to determine if the statistically significant result is due to sampling error or due to effect size. For this information, the researcher will need to determine the effect size, using one of many available effect magnitude measures. He/she will then construct confidence intervals to assess the effect of sample size and error. As a last step, he/she will look to other methods to provide an indication of the replicability of the results. With this information in hand, the researcher can then not only better assess his/her results but can also provide more guidance to other researchers.

As this brief summary has shown, the simplicity and appeal of the dichotomous decision rule, posited by *p* values, is alluring. But, it can lead to misinterpretation of statistical significance, and more importantly it can dis-tract us from a higher goal of scientific inquiry. That is, to determine if the results of a test have any practical value or not.

Defenders of NHST

With the plethora of shortcomings of NHST that have been documented for over 60 years, one would suspect there are few defenders of a procedure that suffers from so many weaknesses. In fact, Oakes (1986) has expressed, "It is extraordinarily difficult to find a statistician who argues explicitly in favor of retention of significance tests" (p. 71). Schmidt (1996a) reported that a few psychologists have argued in favor of retention of NHST, but "all such arguments have been found to be logically flawed and hence false" (p.116). As in all areas of endeavor, change is often difficult to accept, especially movement away from a phenomenon that has become an integral part of the work of so many people for so many years.

Winch and Campbell (1969), Frick (1996), and Cortina and Dunlap (1997) are among those who have spoken for the retention of significance testing. However, all of these defenders acknowledge the problematic nature and limited use of NHST. Winch and Campbell (1969), while defending NHST, stated, " . . . we advocate its use in a perspective that demotes it to a relatively minor role in the valid interpretation of . . . comparisons" (p. 140). The timidity of the typical defense was echoed by Levin (1993), when he stated, " . . . until something better comes along significance testing just might be science's best alternative" (p. 378).

With few strident defenders and almost universal detractors, the salient question is where do we go from here? Since our hallmark statistical test is flawed, do we have a replacement? We not only believe there is a replacement available now, but the replacement methods have the potential, if properly used, to move us out of the current morass described by Meehl (1978) more than 20 years ago. He described a situation in social sciences where theories are like fads. They come to the forefront with a flurry of enthusiasm, then they slowly fade away as both positive and negative results are gleaned from empirical data, and the results get more and more confusing and frustrating. This typical mixture of negative and positive findings is most likely the result of low power studies that sometimes reach statistical significance and sometimes do not.

Instead of all research effort contributing to the body of research knowledge, only the studies that are lucky enough to reach statistical significance via large sample size, or via chance, ever reach the research community. We would like to see a situation where all studies that were adequately designed, controlled, and measured would be reported, regardless of statistical significance. Below, we provide brief primers, along with appropriate references, to the tools that we believe will eventually replace the much flawed NHST.

Effect Magnitude Measures

In search of an alternative to NHST, methodologists have developed both measures of strength of association between the independent and dependent variables and measures of effect size. Combined, these two categories of measures are called "effect magnitude measures" (Maxwell & Delaney, 1990). Table 1 provides information on the known effect magnitude measures.

Table 1
Effect Magnitude Measures

Measures of Strength of Association	Measures of Effect Size
r, r_{pb}, R, R^2, η, η^2, η_{mult}	Cohen (1988) d, f, g, h,q, w
Cohen (1988) f^2	Glass (1976) g'
Contingency coefficient	Hedges (1981) g
Cramer (1946) v	Tang (1938) ϕ

Fisher (1921) z
Hays (1963) ω^2 and ρ_I
Kelly (1935) ε^2
Kendall (1963) W
Tatsuoka (1973) $\hat{\omega}^2_{mult.\ c}$

Note. Eta squared (η^2) in ANOVA , called the correlation ratio, is the sum of squares (SS) for an effect divided by the SS_{total}. R^2 is the proportional reduction in error, or PRE, measure in regression. R^2 is the $SS_{regression}$ divided by $SS_{total.}$ Both η^2 and R^2 are analogous to the coefficient of determination (r^2). Adapted from Kirk, "Practical significance: A concept whose time has come." *Educational and Psychological Measurement, 56*(5), p.749. Copyright 1996 by Sage Publication, Inc. Adapted with permission.

Measures of association are used for examining proportion of variance (Maxwell & Delaney, 1990, p. 98), or how much of the variability in the dependent variable(s) is associated with the variation in the independent variable(s). Common measures of association are the family of correlation coefficients (r), eta squared (η^2) in ANOVA, and R^2 (proportional reduction in error) in regression analysis.

Measures of effect size involve analyzing differences between means. Any mean difference index, estimated effect parameter indices, or standardized difference between means qualify as measures of effect size. It should be noted that effect size indices can be used with data from both correlational and experimental designs (Snyder & Lawson,1993). Both measures of association and effect size can provide us with measures of practical significance when properly used.

Measures of Association

Kirk (1996) has reviewed the history of the development of these measures. Oddly, it was noted that Ronald Fisher, the father of NHST, was one of the first to suggest that researchers augment their tests of significance with measures of association (p. 748). Kirk found that effect magnitude measures other than the traditional measures of variance-accounted-for, such as r^2, are rarely found in the literature (p. 753). He believes this is due not to an awareness of the limitations of NHST but rather to the widespread use of regression and correlation procedures that are based on the correlation coefficient. However, the low instance of use of these measures could be due to their lack of availability in popular statistical software.

Snyder and Lawson (1993) have warned us of the perils of indiscriminate use of measures of association. They indicate that experimental studies and more homogeneous samples result in smaller measures of association and that studies that involve subject-to-variable ratios of 5:1 or less will usually contain noteworthy positive bias (p. 339). Issues such as the study design (fixed or random effects designs) and whether we are using univariate or multivariate measures also impact the choice of measure of association. In general, formulas designed to estimate measures of association in other samples are less biased than formulas designed for estimating measures of asso-ciation in the population. Also, a study that has a large effect size and a large sample size will typically need no correction for bias, however smaller effect sizes and smaller sample sizes should use measures corrected for positive bias. For a detailed explanation of appropriate measures of association as well as computational form-ulas, the reader is referred to either Snyder and Lawson (1993) or Maxwell and Delaney (1990). Various measures of association are shown in Table 1.

Measures of Effect Size

Perhaps no one has done more than Jacob Cohen to make researchers aware of the use of effect size measures, as well as the problem of low test power in NHST. Cohen (1988) also provides us with definitions of effect size as well as conventions that can be used in the absence of specific information regarding a phenomenon. The various effect size measures are outlined in Table 1. Effect size is defined "without any necessary implication of causality . . . (as) . . . the degree to which the phenomenon is present in the population . . . or . . . the degree to which the null hypothesis is false" (p. 9). Cohen further states, "the null hypothesis always means the effect size is zero" (p. 10). A generalized form of effect size *d* is used for independent samples in a one-tailed, directional case:

$$d = \mu_1 - \mu_2 \ / \sigma$$

where *d* is the effect size index for the *t* test for means, μ_1 and μ_2 are population means, and σ is the population standard deviation. As such, the value of the difference in the population means is divided by the population standard deviation to yield a standardized, scale invariant, or metric-free, estimate of the size of the effect.

Substituting sample statistics in the formula as estimates of the population parameters can also be applied. The standard deviation can either be the standard deviation of a control group, assuming equality of variance, or alternatively the pooled (within) population standard devi-ation can be used (Wolf, 1986). Cohen has developed methods of converting most of the popular significance tests to effect size measures. For example,

there are effect size measures for differences between correlation coefficients (q), differences between proportions (h), the chi-square test for goodness of fit and contingency tables (w), ANOVA and ANCOVA (f), multiple regression and other multivariate methods (f^2). The reader is referred to Cohen (1988) for a full treatment of this subject.

Interpreting Effect Size

Various interpretation methods have been developed for effect size measures. Cohen (1988) developed three measures of overlap or U measures. With the assumptions of normality and equality of variance satisfied, and with two populations, A and B, U_1 is defined as the percentage of combined area not shared by the two populations distributions. U_2 is the percentage in the B population that exceeds the same percentage in the A population. U_3 is the percentage of the A population which the upper half of the cases of the B population exceeds. Cohen provides tables to determine the U measures for effect sizes 0 - 4 (p. 22). The U_3 measure of overlap can be interpreted using the tabled values of the standard normal distri-bution. For example, if effect size, *d*, is .5 (a medium effect), the area under the normal curve would be .6915 (.5 +.1915). This means that the treatment effect would be expected to move a typical person from the 50th percentile to the 69th percentile of the control group. Generally, the result of this outcome is graphically displayed for easier interpretation. The reader is referred to Glass (1976) for one of the earliest uses of this interpretive device.

Rosenthal and Rubin (1982) have described a method for evaluating the practical significance of the effect size measures that has shown promise. This procedure trans-forms *r*, or other effect measures, to chi-square (χ^2) to form a binomial effect size display (BESD) for 2 x 2 tables. The relatively easy calculations provide us with the estimated difference in success probabilities between the treatment and control groups. This method holds promise, but criticism has surfaced that attacks the method as distorting the data (McGraw, 1991), especially in cases where differences are highly divergent from 50-50 (Strahan, 1991), and as misinterpreting the data (Crow, 1991). Rosenthal (1991) has responded by noting that this method is context specific and was not intended to assess all situations. As a result, caution should be exer-cised when using BESD tables, especially in cases where differences in treatment and control groups are large.

Interpretation of the effect size is best accomplished by comparing the study effect size to the effect size of similar studies in the field of study. Methods for determining a general effect size in a particular field of study have been limited to studies of the *median* effect size of studies in a particular journal (Haase, Waechter, & Solomon, 1982). This type of study converts traditional test statistics into a distribution of effect sizes and provides a convenient method of comparing results of a single test to that of results in the field as a whole**.** We believe more studies of this type, along with periodic updates, would provide the primary researcher with the most valid assessment of a particular effect size. In lieu of this type of information, Cohen (1988) has provided general conventions for the use of effect size. A small effect is defined as .2, a medium effect as .5, and a large effect as .8. Cohen warns that these conventions are analogous to the conventions for significance levels (α = .05) and should be used with great caution, and only in the case where previous research is unavailable (p. 12). How-ever, Kirk (1996) has noted that the average effect size of observed effects in many fields approximates .5 and the meaning of effect size remains the same without regard to the effect size measure. In general, the ultimate judgment regarding the significance of the effect size measure "rests with the researcher's personal value system, the research questions posed, societal concerns and the design of a particular study" (Snyder & Lawson, 1993, p. 347). Both Snyder and Lawson (1993) and Thompson (1993a, pp. 365-368) provide very readable information on the calculation, as well as the use and limitations of univariate and multivariate effect magnitude measures.

Confidence Intervals

The traditional NHST provides us only with information about whether chance is or is not an explanation for the observed differences. Typically, the use of confidence intervals is treated as an alternative to NHST since both methods provide the same outcome. Point estimates of differences, surrounded by confidence intervals, provide all the information that NHST does, but additionally they provide the degree of precision observed, while requiring no more data than NHST. Surprisingly, based on a review of recent literature, the superiority of this method is not recognized or has been ignored by the research community (Kirk, 1996, p. 755). Why should we routinely report confidence intervals? Not only do they serve to remind the researcher of the error in his/her results and the need to improve measurement and sampling techniques, they also provide a basis for assessing the impact of sample size. Note that confidence intervals are an analogue for test power. A larger sample size, higher power test will have a smaller

confidence interval, while a smaller sample size, lower power test will have a larger confidence interval.

Work on asymmetric confidence intervals and expanding the use of confidence intervals to apply to multivariate techniques and causal models has been underway for some time. Many of the methods have been available but were so complex that they were seldom used. However, the use of high speed computers makes calculations of these confidence intervals more realistic. A detailed look at more recent and appropriate appli-cations of confidence intervals have been described by Reichardt and Gollob (1997) and Serlin (1993).

In summary, there is a multitude of effect magnitude measures available to provide the practical significance of effects revealed in a study. When used in combination with confidence intervals that describe sampling error, magnitude measures present the researcher with more information than is provided by NHST. However, the use of these measures has not yet received widespread acceptance by the research community. We believe the lack of acceptance is due not to active resistance but to a lack of familiarity with effect magnitude measures and confidence intervals when compared with NHST. Some may argue that the interpretation of these measures is more subjective than the dichotomous interpretation of significance tests. However, those arguments fail to consider the subjectivity of the significance level in NHST and the general subjective nature of all empirical science (Thompson, 1993).

Simulated Replications

Fisher (1971), among others, has acknowledged the need for replication of studies in order to verify results and, in the current vernacular, to advance cumulative knowledge. However, there are many factors working against replication studies. Among them are a general disdain for non-original research by journal editors and dissertation committees, lack of information on another's study to replicate it, and the bias that is implied when the researcher replicates his/her own study. Additionally, replication of one's own study immediately following its completion is likely to invoke a strong fatigue factor. Nevertheless, some indication of the likelihood of replicability of results is in the interest of good science.

Fortunately, there are alternatives to full-scale replication. Schmidt (1996a) has noted that the power of a test provides us with an estimate of the probability of replication (p.125), and Thompson (1993a) describes three methods that can be used to indicate the likelihood of replication. Two of the methods, crossvalidation and the jackknife techniques, use split samples to empirically compare results across the sample splits. The third method, bootstrapping, involves sampling equal size samples with replacement from the original data set. After several thousand iterations, one is provided with an analogue to the sampling distribution of means. The resulting data have a variety of uses including estimating the standard error of the means, developing confidence intervals around the estimate of the population mean, and providing a vehicle for viewing the skewness and kurtosis in a simulated population distribution. Thompson pointed out two practical uses of the bootstrap method: 1) to descriptively evaluate the stability of the results of the study, and 2) to make inferences using confidence intervals (p. 372). Statistical software designed by researchers for the specific purpose of conducting bootstrap studies are available (p. 369). The one thing the researcher should always consider when conducting a bootstrap study is the inherent limitations of the original data that are carried over to the bootstrap method. As a result, caution and thoughtfulness in the interpretation of data are called for in this, as in all statistical analyses. In summary, the reporting of studies should include some indication of the replicability of the data. No matter what method the author chooses, it will provide more informa-tion than is available from NHST.

Meta-analysis

Meta-analysis is defined as, ". . . the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3). In the past, subjective literature reviews or simplistic vote counting of significant and non-significant results were used. Light and Pillemer (1984) described these methods as subjective, scien-tifically unsound, and an inefficient way to extract useful information. Cooper and Hedges (1994) describing the early meta-analyses stated, "research synthesis in the 1960s was at best an art, at worst a form of yellow journalism" (p. 7). However, the field of meta-analysis has seen a burst of activity since Glass (1976) first coined the term and used Cohen's effect size and overlap measures to analyze psychotherapy outcome research. Glass paved the way for a plethora of meta-analytic studies in the 1980s and 1990s that used effect size as the dependent variable. Cooper and Hedges (1994) observed that "much of the power and flexibility of quantitative research synthesis is owed to the existence of effect size estimators such as *r* and *d*" (p. 24). The power of these statistics comes from their ability to measure the effects in terms of their own standard deviations.

With the advances in the development of effect size measures and meta-analytic techniques, the field of meta-analysis now has a body of statistics specifically for combining the results of studies (Hedges & Olkin, 1985). Additionally, many of the early methods of meta-analysis have been "standardized" and many of the early criticisms of meta-analysis (Wittrock, 1986) have been addressed (Cooper & Hedges, 1994). Today, we see the role of meta-analysis taking on more and more importance in scientific inquiry. This is evidenced by a growing number of meta-analytic studies published in journals that formerly refused to publish literature reviews, as well as shifting patterns of citations in the literature (Schmidt, 1996a). In a recent development, meta-analytic methods have now been broadened to the empirical study of variability of test score reliability coefficients across samples. This reliability generalization method along with extant validity generalization methods makes meta-analysis an even more powerful method of data synthesis (Vacha-Haase, 1998). The interested reader should consult Cooper and Hedges' (1994) text on methods, statistics and limitations of current meta-analytic practices. The development of meta-analysis as an "independent specialty within the statistical sciences" (p. 6) allows the secondary researcher to use sound statistical methods to combine the results of years of research to interpret a phenomena.

Research Registries

Despite the fact that many of the methods of meta-analysis come from the social sciences, the more dramatic use of these methods has been in the field of health care. This development was most likely due to the availability of registries of studies in the health care field. By tracking all known research studies in specialty areas, the field had a wealth of data to draw upon. Meta-analysis has been so successful in medical research that federal legislation has authorized creation of an agency for health care policy research that is required to develop guidelines based on a systematic synthesis of research evidence (Cooper & Hedges, 1994, p. 7).

One of the problems facing the registries in health care is lack of knowledge in the field about their availability. There are so many registries for so many clinical trials that registries of registries have had to be formed. In the social sciences we can learn a lesson from the ad hoc nature of establishing registries that has developed in medical science. Dickersin (1994) notes that the institutional review system for research registration already exists for all research involving human subjects. She has identified a national system that exists in Spain that mandates cooperation between local institutional review boards and a centralized national board (p. 71). With the availability of high speed electronic transfer of data, what would have seemed like a pipe dream some years ago now has the possibility of becoming a reality. A national system for the social sciences, working through local review boards, could be stimulated through con-certed action by a coalition of professional organizations and the federal government. However, if government intervention is unthinkable, perhaps professional organi-zations could muster the manpower and resources to de-velop research registries in education and/or psychology.

Where We Go from Here

Based on our review of the arguments and logic of NHST and the vast literature on augmentation and replacement methods, we have come to the conclusion (albeit not a unique or new conclusion) that individual studies can best be analyzed by using point estimates of effect size as a measure of the magnitude of effect and confidence limits as a measure of the sampling error. Reporting these findings will provide more detailed information and certainly more raw information than is contained in significance tests (Schafer, 1993). Additionally, individual studies should indicate the likelihood of replication through the use of simulation methods. The researchers who believe the *p* value provides this information are thinking appropriately, but incorrectly, in that replication is the only way to reach consensus on the evidence provided by individual studies. However, statistical tools that simulate replications are the best methods of providing evidence of replicability, short of full-scale replication. We also believe the academic community should rethink the importance and the role of full-scale replication studies in scientific investigation and promote them to a status equal to that of original research. These recommendations should bring some order to the chaotic situation that currently exists in the analysis of individual studies. Using the described methods and with the availa-bility of research registries, the meta-analytic researcher will have access to more studies (including those formerly unsubmitted or rejected as non-significant), and the studies will be reported in a manner that is more conducive to meta-analytic studies.

We believe a major advancement of knowledge will come from a synthesis of many individual studies regarding a particular phenomenon using meta-analytic methods. With the primary researcher providing raw materials, the meta-analytic secondary researcher can analyze trends in various areas of research endeavor and

provide the raw materials for more rational educational policy.

Changing Times

There are signs that the mountain of criticism that has befallen NHST has finally reached fruition. There is evidence in the research environment that change is taking place and the abandonment of NHST for the use of point estimates of effect size with confidence intervals is underway. In 1996, the American Psychological Association's Board of Scientific Affairs formed a task force to study and make recommendations about the conduct of data analysis (APA Monitor, 1997). The initial report of the committee fell short of recommending a ban on NHST, however it did report that " . . . (data analysis) . . . include both direction and size of effect and their confidence intervals be provided routinely . . ." (APA Science Agenda, 1997, p. 9). Two years earlier, and almost unnoticed, the fourth edition of the APA Publication Manual (1994) stated, "You are encouraged to provide effect-size information. . . whenever test statistics and samples sizes are reported" (p. 18). Kirk (1996) reported the APA is also seeking involvement from the AERA, APS, Division 5, the Society for Mathematical Psychology and the American Statistical Association in its study of the NHST issue (p. 756). Schmidt (1996a) reported that studies today are more likely to report effect sizes, and "it is rare today in industrial/organizational psychology for a finding to be touted as important solely on the basis of its *p* value" (p. 127). Additionally, government entities are now seeing the importance of meta-analytic studies and the effect size measures they use and are calling for more studies to guide policy decisions (Sroufe, 1997). Popular statistical software is also being reprogrammed to provide measures of power and effect size (J. McLean, personal communication, November 12, 1997).

Despite the fact that Michigan State has reformed its graduate statistics course sequence in psychology to include teaching of effect size measures and a de-emphasis of NHST (Schmidt, 1996a), it is acknowledged that "there have been no similar improvements in the teaching of quantitative methods in graduate and undergraduate programs" (p. 127). This mirrors a report (Aiken, West, Secrest, & Reno, 1990) that reviewed Ph.D. programs in psychology and concluded that "the statistics . . . curriculum has advanced little in 20 years" (p. 721). Thompson (1995) has also noted that his review of AERA publications and of papers presented at (the) annual meetings suggest that the calls for new methods haven't affected contemporary practice. Based on our own knowledge of teaching methods and statistics textbooks, we do not believe the academic community or textbook publishers have changed appreciably since the 1990 report issued by Aiken, et al. (1990).

Strategies for Change

We respect democratic principles so we cannot in good faith call for a ban on significance testing since this would represent censorship and infringement on individual freedoms. However, we believe that most statisticians would welcome orderly change that would lead to abandonment of NHST. In no way would it prohibit the diehard researcher from using NHST, but all emphasis would be on improved methods of legitimate research. These methods would be directed at ways and means of facilitating meta-analytic studies. This would include editorial policies that require: a) validity and reliability measures on all instruments used; b) use of appropriate effect magnitude measures with confidence intervals to describe studies; c) use of information such as effect size studies of the phenomena of interest, BESD methods, odds ratio's, Cohen's effect size interpretations and other measures to interpret the results; and d) an indication of the replicability of the results obtained using bootstrap or other legitimate methods. Educational research registries would be put in place to attempt to replicate the registries that have demonstrated success in the health care field. Statistical software would be modified to emphasize the procedures and caveats for the newer statistical methods (including meta-analysis), and textbooks would be revised to reflect the changes in emphasis.

We see the various stakeholders, or interest groups, in the discussion we have presented as: a) professional associations, b) journal editors, c) researchers, d) educators, e) statistics textbook writers, and f) statistical software developers. The first steps in replacing NHST have taken place with professional organizations addressing the issue of NHST. We believe this step will eventually influence editorial policies used by journal editors. This, we believe, will be the critical path for change since it will, in turn, influence the researchers' data analyses and writings, as well as their educational practices.

For the above scenario to occur with minimal disruption, a joint project of the leading professional organizations needs to take the first step with a well developed master plan for change. Prominent practitioners, not dissimilar from the extant APA task force on significance testing, would outline a general framework for change following suggestions outlined in this and other works that have taken a critical look at the issues surrounding current research practice.

Following the development of the general plan, several other task forces of prominent practitioners would

be formed to flesh out the details for the master plan. We envision these task forces addressing the issues of editorial policies for scholarly journals, revisions required to be made by textbook and statistical software publishers, and development of research registries. Once the individual task forces had reported, their work would be put out for review and comment by the interested professionals.

The original master plan task force would coordinate the final development of the master plan, based on the input of the various task forces and the public comment. The professional organization would then announce the date for the change-over that would give all stakeholders time to prepare. An analogy would be the rollout of a new computer operating system, where software developers, vendors and users are aware of and prepared for the change that is going to take place long before it actually occurs. Users are kept aware of the progress of change through periodic, well publicized and distributed infor-mation. This process would allow an orderly and expe-dited process. We would envision the above described process entailing approximately 24 to 36 months of concerted effort.

Summary

With the evidence that has been provided, it is reasonable to state that NHST, with its many shortcomings, has failed in its quest to move the social sciences toward verisimilitude and may have actually stymied the advance-ment of knowledge. NHST promised an improved method of determining the significance of a study, and no doubt was enlightening in the 1930s when researchers were saddled with fewer methods of inquiry. Some sixty years later, we can now state that methods with the track record of NHST have no place in scientific inquiry. In the past, we may have had to tolerate the shortcomings of NHST because there were no viable alternatives. Today viable and continually evolving alternatives are available. The use of effect magnitude measures, replication meas-ures, and the statistics that drive meta-analytic studies are no longer embryonic, and we believe they merit a central role in scientific inquiry.

The loss of NHST techniques will not mean that older studies are meaningless. In fact, many studies that have failed to pass the NHST test and were not published or presented can be resurrected and updated with effect size measures. As a result, the loss of NHST will not retard the growth of scientific knowledge but will, ironically, advance scientific knowledge. We strongly believe a major step in advancing cumulative knowledge will be the establishment of research registries to compile all studies of a particular phenomenon for meta-analysis. Contro-versy will always surround statistical studies, and this paper in no way proposes that current effect magnitude measures and meta-analytic techniques are without limitations. We will see misuses of the measures that we propose, just as we have seen misuses of NHST, but we should remain vigilant and not allow these misuses to be institutionalized as they apparently have been with NHST. With change, the new century promises more advanced and enlightened methods will be available to help forge more rational public policies and advance the cumulative knowledge of educational research, in particular, and the social sciences, in general.

References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. L. (1990). Graduate training in statistics, methodology, and measurement in psychology, a survey of Ph.D. programs in North America. *American Psychologist, 45*(6), 721-734.

APA Monitor. (1997, March). *APA task force urges a harder look at data, 28*(3), 26. Washington, D.C.: Author.

APA Science Agenda (1997, March-April). *Task force on statistical inference identifies charge and produces report, 10*(2), 9-10. Washington, D.C.: Author.

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, D.C.: Author.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423-437.

Begg, C. B., (1994). Publication bias. In H. Cooper & L. V. Hedges, (Eds.) *The Handbook of Research Synthesis*. (pp. 399-409). New York: Russell Sage Foundation.

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.

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

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*(4), 287-292.

Cohen, J. (1962). The statistical power of abnormal social psychology research. *Journal of Abnormal and Social Psychology, 65*(3), 145-153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, (2nd ed.). Hillsdale, N.J.; Academic Press.

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*(12), 997-1003.

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

Cortina, J. M. & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*(2), 161-172.

Crow, E. L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology?" *American Psychologist, 46*, 1083.

Dickersin, K. (1994). Research registries. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research syn-thesis* (p. 71). New York: Russell Sage Foundation.

Fisher, R. A. (1971). *The design of experiments*. (8th ed.) New York: Hafner Publishing.

Frick, R. W. (1996) The appropriate use of null hypothesis testing. *Psychological Methods, 1*(4), 379-390.

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

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

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego; Academic Press.

Hinkle, D. E., Wiersma, W., & Jurs, S.G. (1994). *Applied statistics for the behavioral sciences* (3rd ed.). Boston; Houghton Mifflin Company.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*(4), 317-333.

Jones, L. V. (1955). Statistical theory and research design. *Annual Review of Psychology, 6*, 405-430.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*(4), 378-381.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36*(2), 102-105.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth Publishing.

McGraw, K. O. (1991). Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?" *American Psychologist, 46*, 1084-1086.

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*(4), 806-834.

Morrison, D. E., & Henkel, R. E. (1970). *The significance testing controversy - A reader*. Chicago: Aldine Publishing.

Mulaik, S. A, Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Lawrence Erlbaum Associates.

Nunnally, J. (1960). The place of statistics in psychology. *Education and Psychological Measurement, 20*, 641-650.

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

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical significance tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259-284). Mahwah, N.J.: Lawrence Erlbaum Associates.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD and alternative indices. *American Psychologist, 46*, 1086-1087.

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

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp.176-197). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Schafer, J. P. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education*, *61*(4), 383-387.

Schmidt, F. L. (1996a). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115-129.

Schmidt, F. L. (1996b). What do data really mean? Research findings, meta analysis and cumulative knowledge in psychology. *American Psychologist*, *47*(10), 1173-1181.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105* (2), 309-316.

Serlin, R. C. (1993). Confidence intervals and the scientific method: Case for Holm on the range. *Journal of Experimental Education*, *61*(4), 350-360.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, *61*(4), 334-349.

Sroufe, G. E. (1997). Improving the "awful reputation" of educational research. *Educational Researcher*, *26*(7), 26-28.

Strahan, R. F. (1991). Remarks on the binomial effect size display. *American Psychologist*, *46*, 1083-1084.

Thompson, B. (1993a). Foreword. *Journal of Experimental Education*, *61*(4), 285-286.

Thompson, B. (1993b). The use of significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, 6(4), 361-377.

Thompson, B. (1995). *Inappropriate statistical practices in counseling research: Three pointers for readers of research literature*. Washington, D. C. Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. 391 990).

Thompson, B. (1995, November). *Editorial policies regarding statistical significance testing: Three suggested reforms*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS.

Thompson, B. (1998). [Review of the book *What if there were no significance tests?*] *Educational and Psychological Measurement*, (in press).

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6-20.

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

Wittrock, M. C. (1986). *Handbook of Research on Teach-ing*, (3rd ed.). New York: MacMillan Publishing.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*, (Series no. 07-059). Newbury Park, CA: Sage Publications.