RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and the University of Alabama at Birmingham.

Volume 5, Number 2 **Fall 1998**

SPECIAL ISSUE
STATISTICAL SIGNIFICANCE TESTING

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS (ISSN 1085-5300) publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of technology applications in the classroom, descriptions of innovative teaching strategies in research/measurement/statistics, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, RESEARCH IN THE SCHOOLS, School of Education, 233 Educ. Bldg., The University of Alabama at Birmingham, 901 13th Street, South, Birmingham, AL 35294-1250**. **Please direct questions to jmclean@uab.edu**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages, using 11-12 point type. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

RESEARCH IN THE SCHOOLS

Volume 5, Number 2 **Fall 1998**

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 1

Introduction to the Special Issue on Statistical Significance Testing

Alan S. Kaufman
Co-Editor, RESEARCH IN THE SCHOOLS
Clinical Professor of Psychology
Yale University, School of Medicine

The controversy about the use or misuse of statistical significance testing that has been evident in the literature for the past 10 years has become the major methodological issue of our generation. In addition to many articles and at least one book that have been written about the subject, several journals have devoted special issues to dealing with the issues surrounding its use. Because this issue has become so prevalent and it impacts on research in the schools in general and articles published in the *RESEARCH IN THE SCHOOLS* journal as well, James McLean and I--as co-editors of the journal--felt that a special issue that explored all sides of the controversy was in order. To me, personally, the topic is an exciting one. I have published a great many research articles during the past three decades, and often have felt that statistical significance was an imperfect tool. Why should a trivial difference in mean scores or a correlation that begins with a zero be significant simply because the sample is large? Yet, until I began reading articles that challenged the holiness of the birthright of statistical significance testing, I must confess that it never occurred to me to even ask questions such as, "Is there a better way to evaluate research hypotheses?" or "Is statistical significance testing essential to include in a research article?"

This special issue begins with three articles that explore the controversy from several perspectives (Nix and Barnette, McLean and Ernest, and Daniel). These three articles were submitted independently of each other, coincidentally at about the same time, and were peer-reviewed by our usual review process. I then asked the three sets of authors if they would be willing to have their articles serve as the stimuli for a special issue on the topic, and all readily agreed. I then solicited three respondents to the three articles (Thompson, Knapp, and Levin), researchers who seemed to represent the whole gamut of opinions on the topic of the use and possible misuse of statistical significance testing. I asked Bruce Thompson to respond to the articles, even though he had already served as a peer reviewer of these manuscripts, because of his eminence in the field. The three responses to the manuscript follow the three main articles. The special issue concludes with rejoinders from the three initial sets of authors. I believe that you will find the disagreements, none of which are vitriolic or personal, to be provocative and fascinating. Because co-editor James McLean was an author of one of the significance testing articles, he did not participate in editorial decisions with respect to this issue of the journal.

Both Jim McLean and I are very interested in your--the reader's--response to this special issue. We would like to know where our readership stands on the controversial topics debated in the pages of this special issue. We would like to invite you to send us your opinions on the use and misuse of statistical significance testing--what points you agree with and which ones you find not to be very persuasive. We intend to develop a unified policy on this topic for *RESEARCH IN THE SCHOOLS*, which we will base not only on the content of this special issue of the journal, but also on your opinions. We will print every letter that we receive on the topic in the same future issue of our journal that includes our policy statement.

Finally, this issue represents the completion of five years of publication of *RESEARCH IN THE SCHOOLS*. Both author and title indexes are included in this issue to commemorate that accomplishment and make past articles more accessible. In addition, the ERIC Clearinghouse on Assessment and Evaluation catalogs each issue, making *RESEARCH IN THE SCHOOLS* searchable through the ERIC database.

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.

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 15-22

The Role of Statistical Significance Testing In Educational Research

James E. McLean
University of Alabama at Birmingham

James M. Ernest
State University of New York at Buffalo

The research methodology literature in recent years has included a full frontal assault on statistical significance testing. The purpose of this paper is to promote the position that, while significance testing as the sole basis for result interpretation is a fundamentally flawed practice, significance tests can be useful as one of several elements in a comprehensive interpretation of data. Specifically, statistical significance is but one of three criteria that must be demonstrated to establish a position empirically. Statistical significance merely provides evidence that an event did not happen by chance. However, it provides no information about the meaningfulness (practical significance) of an event or if the result is replicable. Thus, we support other researchers who recommend that statistical significance testing must be accompanied by judgments of the event's practical significance and replicability.

The research methodology literature in recent years has included a full frontal assault on statistical significance testing. An entire edition of a recent issue of *Experimental Education* (Thompson, 1993b) explored this controversy. There are some who recommend the total abandonment of statistical significance testing as a research methodology option, while others choose to ignore the controversy and use significance testing following traditional practice. The purpose of this paper is to promote the position that while significance testing by itself may be flawed, it has not outlived its usefulness. However, it must be considered in the total context of the situation. Specifically, we support the position that statistical significance is but one of several criteria that must be demonstrated to establish a position empirically. Statistical significance merely pro-vides evidence that an event did not happen by chance. However, it provides no information about the meaning-fulness (practical significance) of an event or if the result is replicable.

This paper addresses the controversy by first provid-ing a critical review of the literature. Following the review are our summary and recommendations. While none of the recommendations by themselves are entirely new, they provide a broad perspective on the controversy a n d provide practical guidance for researchers employing statistical significance testing in their work.

James E. McLean is a university research professor and the director of the Center for Educational Accountability in the School of Education at the University of Alabama at Birmingham. James M. Ernest is a lecturer in the Department of Learning and Instruction, Graduate School of Education, State University of New York at Buffalo. Correspondence relevant to this article should be addressed to James E. McLean, Center for Educational Accountability, University of Alabama at Birmingham, 901 13th Street South, Birmingham, AL 35294-1250 or by e-mail to jmclean@uab.edu.

Review of the Literature

Scholars have used statistical testing for research purposes since the early 1700s (Huberty, 1993). In the past 300 years, applications of statistical testing have advanced considerably, most noticeably with the advent of the computer and recent technological advances. However, much of today's statistical testing is based on the same logic used in the first statistical tests and advanced in the early twentieth century through the work of Fisher, Neyman, and the Pearson family (see the appendix to Mulaik, Raju, & Harshman, 1997, for further information). Specifically, significance testing and hypothesis testing have remained at the cornerstone of research papers and the teaching of introductory statistics courses. (It should be noted that while the authors recognize the importance of Bayesian testing for statistical significance, it will not be discussed, as it falls outside the context of this paper.) Both methods of testing hold at their core basic premises concerning probability. In what may be termed Fisher's *p value approach*, after stating a null hypothesis and then obtaining sample results (i.e., "statistics"), the probability of the sample results (or sample results more extreme in their deviation from the null) is computed, assuming that the null is true in the population from which the sample was derived (see Cohen, 1994 or Thompson, 1996 for further explanation). The Neyman-Pearson or *fixed-alpha approach* specifies

a level at which the test statistic should be rejected and is set a priori to conducting the test of data. A null hypothesis (H_o) and an alternative hypothesis (H_a) are stated, and if the value of the test statistic falls in the rejection region the null hypothesis is rejected in favor of the alternate hypothesis. Otherwise the null hypothesis is retained on the basis that there is insufficient evidence to reject it.

Distinguishing between the two methods of statistical testing is important in terms of how methods of statistical analysis have developed in the recent past. Fisher's legacy of statistical analysis approaches (including ANOVA methods) relies on subjective judgments concerning differences between and within groups, using probability levels to determine which results are statistically significant from each other. Karl Pearson's legacy involves the development of correlational analyses and providing indexes of association. It is because of different approaches to analyses and different philosophical beliefs that the issue of testing for statistical significance has risen. In Huberty's (1993) historical review of the importance of statistical significance testing literature, the research community has shifted from one perspective to another, often within the same article. Currently we are in an era where the value of statistical significance testing is being challenged by many researchers. Both positions (arguing for and against the use of statistical significance tests in research) are presented in this literature review, followed by a justification for our position on the use of statistical significance testing as part of a comprehensive approach.

As previously noted, the research methodology literature in recent years has included a full frontal assault on statistical significance testing. Of note, an entire edi-tion of *Experimental Education* explored this controversy (Thompson, 1993b). An article was written for *Measure-ment and Evaluation in Counseling and Development* (Thompson, 1989). The lead section of the January, 1997 issue of *Psychological Science* was devoted to a series of articles on this controversy (cf., Hunter, 1997). An article suggesting editorial policy reforms was written for the American Educational Research Association (Thompson, 1996), reflected on (Robinson & Levin, 1997), and a rejoinder written (Thompson, 1997). Additionally, the American Psychological Association created a Task Force on Statistical Inference (Shea, 1996), which drafted an initial Report to the Board of Scientific Affairs in December 1996, and has written policy statements in the *Monitor*.

The assault is based on whether or not statistical significance testing has value in answering a research question posed by the investigators. As Harris (1991) noted, "There is a long and honorable tradition of blistering attacks on the role of statistical significance testing in the behavioral sciences, a tradition reminiscent of knights in shining armor bravely marching off, one by one, to slay a rather large and stubborn dragon . . . . Given the cogency, vehemence and repetition of such attacks, it is surprising to see that the dragon will not stay dead" (p. 375). In fact, null hypothesis testing still dominates the social sciences (Loftus & Masson, 1994) and still draws derogatory statements concerning the researcher's methodological competence. As Falk and Greenbaum (1995) and Weitzman (1984) noted, the researchers' use of the null may be attributed to the experimenters' ignorance, misunderstanding, laziness, or adherence to tradition. Carver (1993) agreed with the tenets of the previous statement and concluded that "the best research articles are those that include *no* tests of statistical significance" (p. 289, italics in original). One may even concur with Cronbach's (1975) statement concerning periodic efforts to "exorcize the null hypothesis" (p. 124) because of its harmful nature. It has also been suggested by Thompson, in his paper on the etiology of researcher resistance to changing practices (1998, January) that researchers are slow to adopt approaches in which they were not trained originally.

In response to the often voracious attacks on significance testing, the American Psychological Association, as one of the leading research forces in the social sciences, has reacted with a cautionary tone: "*An APA task force won't recommend a ban on significance testing, but is urging psychologists to take a closer look at their data*" (Azar, 1997, italics in original). In reviewing the many publications that offer advice on the use or misuse of statistical significance testing or plea for abstinence from statistical significance testing, we found the following main arguments for and against its use: (a) what statistical significance testing does and does not tell us, (b) empha-sizing effect-size interpretations, (c) result replicability, (d) importance of the statistic as it relates to sample size, (e) the use of language in describing results, and (f) the recognition of the importance of other types of information such as Type II errors, power analysis, and confidence intervals.

What Statistical Significance Testing Does and Does Not Tell Us

Carver (1978) provided a critique against statistical significance testing and noted that, with all of the criticisms against tests of statistical significance, there appeared to be little change in research practices. Fifteen years later, the arguments delivered by Carver (1993) in the *Journal of Experimental Education* focused on the negative aspects of significance testing and offered a series of ways to minimize the importance of statistical significance testing. His article indicted the research community for reporting significant differences when the results may be trivial, and called for the use of effect size estimates and study replicability. Carver's argument focused on what statistical significance testing *does not do*, and proceeded to highlight ways to provide indices of

practical significance and result replicability. Carver (1993) recognized that 15 years of trying to extinguish the use of statistical significance testing has resulted in little change in the use and frequency of statistical significance testing. Therefore the tone of the 1993 article differed from the 1978 article in shifting from a dogmatic anti-statistically significant approach to more of a bipartisan approach where the limits of significance testing were noted and ways to decrease their influence provided. Specifically, Carver (1993) offered four ways to minimize the importance of statistical significance testing: (a) insist on the word *statistically* being placed in front of significance testing, (b) insist that the results always be interpreted with respect to the data first, and statistical significance second, (c) insist on considering effect sizes (whether significant or not), and (d) require journal editors to publicize their views on the issue of statistical significance testing prior to their selection as editors.

Shaver (1993), in the same issue of *The Journal of Experimental Education*, provided a description of what significance testing is and a list of the assumptions involved in statistical significance testing. In the course of the paper, Shaver methodically stressed the importance of the assumptions of random selection of subjects and their random assignment to groups. Levin (1993) agreed with the importance of meeting basic statistical assumptions, but pointed out a fundamental distinction between statis-tical significance testing and statistics that provide estimates of practical significance. Levin observed that a statistically significant difference gives information about *whether* a difference exists. As Levin noted, if the null hypothesis is rejected, the *p* level provides an "a posteri-ori indication of the probability of obtaining the outcomes as extreme or more extreme than the one obtained, given the null hypothesis is true" (p. 378). The effect size gives an estimate of the noteworthiness of the results. Levin made the distinction that the effect size may be necessary to obtain the size of the effect; however, it is statistical significance that provides information which alludes to whether the results may have occurred by chance. In essence, Levin's argument was for the two types of significance being complementary and not competing concepts. Frick (in press) agreed with Levin: "When the goal is to make a claim about how scores were produced, statistical testing is still needed, to address the possibility of an observed pattern in the data being caused just by chance fluctuation" (in press). Frick's thesis concerning the utility of the statistical significance test was provided with a hypothetical situation in mind: the researcher is provided with two samples who together are the population under study. The researcher wants to know whether a particular method of learning to read is better than another method. As Frick (in press) noted,

> statistical testing is needed, despite complete knowledge of the population. The . . . experimenter wants to know if Method A is better than Method B, not whether the population of people learning with Method A is better than the population of people learning with Method B. The first issue is whether this difference could have been caused by chance, which is addressed with statistical testing. The example is imaginary, but a possible real-life analog would be a study of all the remaining speakers of a dying language, or a study of all of the split-brain patients in the world.

One of the most important emphases in criticisms of contemporary practices is that researchers must evaluate the practical importance of results, and not only statistical significance. Thus, Kirk (1996) agreed that statistical significance testing was a necessary part of a statistical analysis. However, he asserted that the time had come to include practical significance in the results. In arguing for the use of statistical significance as necessary, but insufficient for interpreting research, Suen (1992) used an 'overbearing guest' analogy to describe the current state of statistical significance testing. In Suen's analogy, statistical significance is the overbearing guest at a dinner party who

> inappropriately dominates the activities and conversation to the point that we forget who the host was. We cannot disinvite this guest. Instead, we need to put this guest in the proper place; namely as one of the many guests and by no means the host. (p. 78)

Suen's reference to a "proper place" is a call for researchers to observe statistical significance testing as a means to "filter out the sampling fluctuations hypothesis so that the observed information (difference, correlation) becomes slightly more clear and defined" (p. 79). The other "guests" that researchers should elevate to a higher level include ensuring the quality of the research design, measurement reliability, treatment fidelity, and using sound clinical judgment of effect size.

For Frick (in press), Kirk (1996), Levin (1993), and Suen (1992), the rationale for statistical significance testing is independent of and complementary to tests of practical significance. Each of the tests provides distinct pieces of information, and all three authors recommend the use of statistical significance testing; however, it must be considered in combination with other criteria. Specifically, statistical significance is but one of three criteria

that must be demonstrated to establish a position empirically (the other two being practical significance and replicability).

Emphasizing Effect-Size Interpretations

The recent American Psychological Association (1994) style manual noted that

> Neither of the two types of probability values [statistical significance tests] reflects the importance or magnitude of an effect because both depend on sample size . . . You are [therefore] *encouraged* to provide effect-size information. (p. 18, italics added)

Most regrettably, however, empirical studies of articles published since 1994 in psychology, counseling, special education, and general education suggest that merely "*encouraging*" effect size reporting (American Psychological Association, 1994) has *not* appreciably affected actual reporting practices (e.g., Kirk, 1996; Snyder & Thompson, in press; Thompson & Snyder, 1997, in press; Vacha-Haase & Nilsson, in press). Due to this lack of change, authors have voiced stronger opinions concerning the "emphasized" recommendation. For example, Thompson (1996) stated "AERA should venture beyond APA, and *require* such [effect size] reports in all quantitative studies" (p. 29, italics in original).

In reviewing the literature, the authors were unable to find an article that argued against the value of including some form of effect size or practical significance estimate in a research report. Huberty (1993) noted that "of course, empirical researchers should not rely exclusively on statistical significance to assess results of statistical tests. Some type of measurement of magnitude or importance of the effects should also be made" (p. 329). Carver's third recommendation (mentioned previously) was the inclu-sion of terms that denote an effect size measure; Shaver (1993) believed that "studies should be published without tests of statistical significance, but not without effect sizes" (p. 311); and Snyder and Lawson (1993) contri-buted a paper to *The Journal of Experimental Education* special edition on statistical significance testing titled "Evaluating Results Using Corrected and Uncorrected Effect Size Estimates." Thompson (1987, 1989, 1993a, 1996, 1997) argued for effect sizes as one of his three recommendations (the language use of statistical significance and the inclusion of result replicability results were the other two); Levin (1993) reminded us that "statistical significance (alpha and *p* values) and practical significance (effect sizes) are not *competing* concepts– they are *complementary* ones" (p.379, italics in original), and the articles by Cortina and Dunlap (1997), Frick (1995, in press), and Robinson and Levin (1997) agreed that a measure of the size of an effect is indeed important in providing results to a reader.

We agree that it is important to provide an index of not only the statistical significance, but a measure of its magnitude. Robinson and Levin (1997) took the issue one step further and advocated for the use of adjectives such as *strong/large*, *moderate/medium*, etc. to refer to the effect size and to supply information concerning *p* values. However, some authors lead us to believe that they feel it may be necessary only to provide an index of practical significance and that it is unnecessary to provide statistical significance information. For example, it could be con-cluded from the writings of Carver (1978, 1993) and Shaver (1993) that they would like to abandon the use of statistical significance testing results. Although Cohen (1990, 1994) did not call for the outright abandonment of statistical significance testing, he did assert that you can attach a *p*-value to an effect size, but "it is far more in-formative to provide a confidence interval" (Cohen, 1990, p. 1310). Levin, in his 1993 article and in an article co-authored with Robinson (1997), argued against the idea of a single indicator of significance. Using hypothetical examples where the number of subjects in an experiment equals two, the authors provide evidence that practical significance, while noteworthy, does not provide evi-dence that the results gained were not gained by chance.

It is therefore the authors' opinion that it would be prudent to include both statistical significance and estimates of practical significance (not forgetting other important information such as evidence of replicability) within a research study. As Thompson (in press) discussed, any work undertaken in the social sciences will be based on subjective as well as objective criteria. The importance of subjective decision-making, as well as the idea that social science is imprecise and based on human judgment as well as objective criteria, helps to provide common benchmarks of quality. Subjectively choosing alpha levels (and in agreement with many researchers this does not necessarily denote a .05 or .01 level), power levels, and adjectives such as *large effects* for practical significance (cf. Cohen's [1988] treatise on power analysis, or Robinson and Levin's [1997] criteria for effect size estimates) are part of establishing common benchmarks or creating objective criteria. Robinson and Levin (1997) expressed the relationship between two types of significance quite succinctly: "First convince us that a finding is *not due to chance*, and only then, assess how *impressive* it is" (p. 23, italics in original).

Result Replicability

Carver (1978) was quick to identify that neither significance testing nor effect sizes typically inform the researcher regarding the likelihood that results will be replicated in future research. Schafer (1993), in response to the articles in *The Journal of Experimental Education*,

felt that much of the criticism of significance testing was misfocused. Schafer concluded that readers of research should not mistakenly assume that statistical significance is an indication that the results may be replicated in the future; the issue of replication provides the impetus for the third recommendation provided by Thompson in both his 1989 *Measurement and Evaluation in Counseling and Development* article and 1996 AERA article.

According to Thompson (1996), "If science is the business of discovering replicable effects, because statistical significance tests do not evaluate result replicability, then researchers should use and report some strategies that *do* evaluate the replicability of their results" (p. 29, italics in original). Robinson and Levin (1997) were in total agreement with Thompson's recommendations of external result replicability. However, Robinson and Levin (1997) disagreed with Thompson when they concluded that internal replication analysis constitutes "an acceptable substitute for the genuine 'article'" (p. 26). Thompson (1997), in his rejoinder, recognized that external replication studies would be ideal in all situations, but concludes that many researchers do not have the stamina for external replication, and internal replicability analysis helps to determine where noteworthy results originate.

In terms of statistical significance testing, all of the arguments offered in the literature concerning replicability report that misconceptions about what statistical significance tells us are harmful to research. The authors of this paper agree, but once again note that misconceptions are a function of the researcher and not the test statistic. Replicability information offers important but somewhat different information concerning noteworthy results.

Importance of the Statistic as it Relates to Sample Size

According to Shaver (1993), a test of statistical significance "addresses only the simple question of whether a result is a likely occurrence under the null hypothesis with randomization and a sample of size *n*" (p. 301). Shaver's inclusion of "a sample of size *n*" indicates the importance of sample size in the H_o decision-making process. As reported by Meehl (1967) and many authors since, with a large enough sample and reliable assess-ment, practically every association will be statistically significant. As noted previously, within Thompson's (1989) article a table was provided that showed the relationship between *n* and statistical significance when the effect size was kept constant. Two salient points applicable to this discussion were highlighted in Thompson's article: the first noted the relationship of *n* to statistical significance, providing a simulation that shows how, by varying *n* to create a large enough sample, a difference between two values can change a non-significant result into a statistically significant result. The second property of significance testing Thompson alluded to was an indication that "superficial understanding of significance testing has led to serious distortions, such as researchers interpreting significant results involving large effect sizes" (p. 2). Following this line of reasoning, Thompson (1993a) humorously noted that "tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and they are tired" (p. 363). Thus, as the sample size increases, the importance of significance testing is reduced. However, in small sample studies, significance testing can be useful, as it provides a level of protection from reporting random results by providing information about the chance of obtaining the sample statistics, given the sample size *n*, when the null hypothesis is exactly true in the population.

The Use of Language in Describing Results

Carver (1978, 1993), Cronbach (1975), Morrison and Henkel (1970), Robinson and Levin (1997), and Thompson (1987, 1989, 1993a, 1996, 1997) all stressed the need for the use of better language to describe significant results. As Schneider and Darcy (1984) and Thompson (1989) noted, significance is a function of at least seven interrelated features of a study where the size of the sample is the most influential characteristic. Thompson (1989) used an example of varying sample sizes with a fixed effect size to indicate how a small change in sample size affects the decision to reject, or fail to reject, H_o. The example helped to emphasize the cautionary nature that should be practiced in making judgements about the null hypothesis and raised the important issue of clarity in writing. These issues were the basis of Thompson's (1996) AERA article, where he called for the use of the term "statistically significant" when referring to the process of rejecting H_o based on an alpha level. It was argued that through the use of specific terminology, the phrase "statistically significant" would not be confused with the common semantic meaning of *significant*.

In response, Robinson and Levin (1997) referred to Thompson's comments in the same light as Levin (1993) had done previously. While applauding Thompson for his "insightful analysis of the problem and the general spirit of each of his three article policy recommendations" (p. 21), Robinson and Levin were quick to counter with quips about "language police" and letting editors focus on content and substance and not on dotting the i's and crossing the t's. However, and interestingly, Robinson and Levin (1997) proceeded to concur with Thompson on the importance of language and continued their article

with a call for researchers to use words that are more specific in nature. It is Robinson and Levin's (1997) recommenda-tion that, instead of using the word statistically *significant*, researchers use statistically *nonchance* or statistically *real*, reflecting the test's intended meaning. The authors' rationale for changing the terminology reflects their wish to provide clear and precise information.

Thompson's (1997) rejoinder to the charges brought forth by Robinson and Levin (1997) was, fundamentally, to agree with their comments. In reference to the question of creating a "language police," Thompson admitted that "I, too, find this aspect of my own recommendation troublesome" (p. 29). However, Thompson firmly believes the recommendations made in the AERA article should stand, citing the belief that "over the years I have reluctantly come to the conclusion that confusion over what statistical significance evaluates is sufficiently serious that an exception must be made in this case" (p. 29).

In respect to the concerns raised concerning the use of language, it is not the practice of significance testing that has created the statistical significance debate. Rather, the underlying problem lies with careless use of language and the incorrect assumptions made by less knowledgeable readers and practitioners of research. Cohen (1990) was quick to point out the rather sloppy use of language and statistical testing in the past, noting how one of the most grievous errors is the belief that the *p* value is the exact probability of the null hypothesis being true. Also, Cohen (1994) in his article; "The Earth is Round (*p* less than .05)" once again dealt with the ritual of null hypothesis significance testing and an almost mechanical dichotomous decision around a sacred $\alpha = .05$ criterion level. As before, Cohen (1994) referred to the misinterpretations that result from this type of testing (e.g., the belief that *p*-values are the probability that the null hypothesis is false). Cohen again suggested exploratory data analysis, graphical methods, and placing an emphasis on estimating effect sizes using confidence intervals. Once more, the basis for the argument against statistical significance testing falls on basic misconceptions of what the *p*-value statistic represents.

One of the strongest rationales for not using statistical significance values relies on misconceptions about the meaning of the *p*-value and the language used to describe its purpose. As Cortina and Dunlap (1997) noted, there are many cases where drawing conclusions based on *p* values are perfectly reasonable. In fact, as Cortina and Dunlap (1997), Frick (1995), Levin (1993), and Robinson and Levin (1997) pointed out, many of the criticisms of the *p* value are built on faulty premises, misleading examples, and incorrect assumptions concerning popula-tion parameters, null hypotheses, and their relationship to samples. For example, Cortina and Dunlap emphasized the incorrect use of logic (in particular the use of syllo-gisms and the Modus Tollens rule) in finding fault with significance testing, and Frick provides an interesting theoretical paper where he shows that in some circum-stances, and based on certain assumptions, it *is* possible for the null hypothesis to be true.

It should be noted that several journals have adopted specific policies regarding the reporting of statistical results. The "Guidelines for Contributors" of the *Journal of Experimental Education* include the statement, "authors are *required* to report and interpret magnitude-of-effect measures in conjunction with every *p* value that is reported" (Heldref Foundation, 1997, pp. 95-96, italics added). The *Educational and Psychological Measurement* "Guidelines for Authors" are even more emphatic. They state:

> We will go further [than mere encouragement]. Authors reporting statistical significance will be *required* to both report and interpret effect sizes. However, their effect sizes may be of various forms, including standardized differences, or uncorrected (e.g., r^2, R^2, eta^2) or corrected (e.g., adjusted R^2, $omega^2$) variance-accounted-for statistics. (Thompson, 1994, p. 845, italics in original)

At least one APA journal is also clear about this requirement. The following is from an editorial in the *Journal of Applied Psychology*.

> If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

For these journals, the reporting of effect size is required and the editors will consider statistical significance tests in their proper contexts. However, for most journals, the use of statistical and practical significance is determined by the views of the reviewers, and the editors and authors are subject to the decisions made by the reviewers they draw for their submissions.

The Recognition of the Importance of Other Types of Information

Other types of information are important when one considers statistical significance testing. The researcher should not ignore other information such as Type II errors, power analysis, and confidence intervals. While

all of these statistical concepts are related, they provide different types of information that assist researchers in making decisions. There is an intricate relationship between power, sample size, effect size, and alpha (Cohen, 1988). Cohen recommended a power level of .80 for no other reason than that for which Fisher set an alpha level of .05 — it seemed a reasonable number to use. Cohen believed that the effect size should be set using theory, and the alpha level should be set using what degree of Type I error the researcher is willing to accept based on the type of experiment being conducted. In this scenario, *n* is the only value that may vary, and through the use of mathematical tables, is set at a particular value to be able to reach acceptable power, effect size, and alpha levels. Of course, in issues related to real-world examples, money is an issue and therefore sample sizes may be limited.

It is possible that researchers have to use small *n*'s because of the population they are studying (such as special education students). Cohen (1990) addresses the problems mentioned above by asking researchers to plan their research using the level of alpha risk they want to take, the size of the effect they wish to find, a calculated sample size, and the power they want. If one is unable to use a sample size of sufficient magnitude, one must compromise power, effect size, or as Cohen puts it, "even (heaven help us) increasing your alpha level" (p. 1310). This sentiment was shared by Schafer (1993) who—in reviewing the articles in the special issue of *The Journal of Experimental Education*—believed that researchers should set alpha levels, conduct power analysis, decide on the size of the sample, and design research studies that would increase effect sizes (e.g., through the careful addition of covariates in regression analysis or extending treatment interventions). It is necessary to balance sample size against power, and this automatically means that we do not fix one of them. It is also necessary to balance size and power against cost, which means that we do not arbitrarily fix sample size. All of the recommendations may be conducted prior to the data collection and therefore before the data analysis. The recommendations, in effect, provide evidence that methodological prowess may overcome some of the a posteriori problems researchers find.

Summary and Recommendations

We support other researchers who state that statistical significance testing must be accompanied by judgments of the event's practical significance and replicability. However, the likelihood of a chance occurrence of an event must not be ignored. We acknowledge the fact that the importance of significance testing is reduced as sample size increases. In large-sample experiments, particularly those involving multiple variables, the role of significance testing diminishes because even small, non-meaningful differences are often statistically significant. In small-sample studies where assumptions such as random samp-ling are practical, significance testing provides meaningful protection from random results. It is important to remember that statistical significance is only one criterion useful to inferential researchers. In addition to statistical significance, practical significance, and replicability, researchers must also consider Type II Errors and sample size. Furthermore, researchers should not ignore other techniques such as confidence intervals. While all of these statistical concepts are related, they provide different types of information that assist researchers in making decisions.

Our recommendations reflect a moderate mainstream approach. That is, we recommend that in situations where the assumptions are tenable, statistical significance testing still be applied. However, we recommend that the analy-ses always be accompanied by at least one measure of practical significance, such as effect size. The use of con-fidence intervals can be quite helpful in the interpretation of statistically significant or statistically nonsignificant results. Further, do not consider a hypothesis or theory "proven" even when both the statistical and practical significance have been established; the results have to be shown to be replicable. Even if it is not possible to establish external replicability for a specific study, intern-al approaches such as jackknife or bootstrap procedures are often feasible. Finally, please note that as sample sizes increase, the role of statistical significance becomes less important and the role of practical significance increases. This is because statistical significance can provide false comfort with results when sample sizes are large. This is especially true when the problem is multivariate and the large sample is representative of the target population. In these situations, effect size should weigh heavily in the interpretations.

References

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

Azar, B. (1997). *APA task force urges a harder look at data* [On-line]. Available: http://www.apa.org/monitor/mar97/stats.html

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

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

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

Cohen, J. (1990). Things I have learned so far. *American Psychologist, 45*(12), 1304-1312.

Cohen, J. (1994). The Earth is Round (*p* less than .05). *American Psychologist*, *49*(12), 997-1003.

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

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

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

Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition*, *23*, 132-138.

Frick, R. W. (In press). Interpreting statistical testing: processes, not populations and random sampling. *Behavior Research Methods, Instruments, & Computers.*

Harris, M. J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory & Psychology*, *1*, 375-382.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education*, *65*, 95-96.

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

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*(1), 3-7.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-59.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *The Journal of Experimental Education, 61*(4), 378-382.

Loftus, G. R., & Masson, M. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*, 476-490.

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

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

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and 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: Erlbaum.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology*, *82*, 3-5.

Robinson D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Schneider, A. L. & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. *Evaluation Review, 8,* 573-582.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education, 61*(4), 293-316.

Shea, C. (1996). Psychologists debate accuracy of "significance test." *Chronicle of Higher Education, 42*(49), A12, A16.

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

Snyder, P. A., & Thompson, B. (in press). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly*.

Suen, H. K. (1992). Significance testing: Necessary but insufficient. *Topics in Early Childhood Special Education, 12(1*), 66-81.

Task Force on Statistical Inference Initial Draft Report (1996). *Report to the Board of Scientific Affairs*. American Psychological Association [On-line]. Available: http://www.apa.org/science/tfsi.html.

Thompson, B. (1987, April). *The use (and misuse) of statistical significance testing. Some recommendations for improved editorial policy and practice*. Paper pre-sented at the annual meeting of the American Educa-tional Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868).

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22,* 2-5.

Thompson, B. (1993a). The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Education, 61*(4), 361-377.

Thompson, B. (Guest Ed.). (1993b). Statistical significance testing in contemporary practice [Special issue]. *The Journal of Experimental Education, 61*(4).

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29-32.

Thompson, B. (1998, January). *Why "encouraging" effect size reporting isn't working: The etiology of researcher resistance to changing practices*. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document ED Number forthcoming)

Thompson, B. (in press). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), *Reading and under-standing multivariate statistics* (*Vol. 2*). Washington, DC: American Psychological Association.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. *Journal Experimental Education, 66*, 75-83.

Thompson, B., & Snyder, P. A. (in press). Statistical significance testing and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development*.

Vacha-Haase, T., & Nilsson, J. E. (in press). Statistical significance reporting: Current trends and usages within *MECD*. *Measurement and Evaluation in Counseling and Development*.

Weitzman, R. A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports, 54*, 355-363.

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 23-32

Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals

Larry G. Daniel
University of North Texas

Statistical significance tests (SSTs) have been the object of much controversy among social scientists. Proponents have hailed SSTs as an objective means for minimizing the likelihood that chance factors have contributed to research results; critics have both questioned the logic underlying SSTs and bemoaned the widespread misapplication and misinterpretation of the results of these tests. The present paper offers a framework for remedying some of the common problems associated with SSTs via modification of journal editorial policies. The controversy surrounding SSTs is overviewed, with attention given to both historical and more contemporary criticisms of bad practices associated with misuse of SSTs. Examples from the editorial policies of Educational and Psychological Measurement and several other journals that have established guidelines for reporting results of SSTs are overviewed, and suggestions are provided regarding additional ways that educational journals may address the problem.

Statistical significance testing has existed in some form for approximately 300 years (Huberty, 1993) and has served an important purpose in the advancement of inquiry in the social sciences. However, there has been much controversy over the misuse and misinterpretation of statistical significance testing (Daniel, 1992b). Pedhazur and Schmelkin (1991, p. 198) noted, "Probably few methodological issues have generated as much controversy among sociobehavioral scientists as the use of [statistical significance] tests." This controversy has been evident in social science literature for some time, and many of the articles and books exposing the problems with statistical significance have aroused remarkable interest within the field. In fact, at least two articles on the topic appeared in a list of works rated by the editorial board members of *Educational and Psychological Measurement* as most influential to the field of social science measurement (Thompson & Daniel, 1996b). Interestingly, the criticisms of statistical significance testing have been pronounced to the point that, when one reviews the literature, "it is more difficult to find specific arguments for significance tests than it is to find arguments decrying their use" (Henkel, 1976, p. 87); nevertheless, Harlow, Mulaik, and Steiger (1997), in a new book on the controversy, present chapters on both sides of the issue. This volume, titled *What if There Were No Significance Tests?*, is highly recommended to those interested in the topic, as is a thoughtful critique of the volume by Thompson (1998).

Larry G. Daniel is a professor of education at the University of North Texas. The author is indebted to five anonymous reviewers whose comments were instrumental in improving the quality of this paper. Address correspondence to Larry G. Daniel, University of North Texas, Denton, TX 76203 or by e-mail to daniel@tac.coe.unt.edu.

Thompson (1989b) noted that researchers are increas-ingly becoming aware of the problem of over-reliance on statistical significance tests (referred to herein as "SSTs"). However, despite the influence of the many works critical of practices associated with SSTs, many of the problems raised by the critics are still prevalent. Researchers have inappropriately utilized statistical significance as a means for illustrating the importance of their findings and have attributed to statistical significance testing qualities it does not possess. Reflecting on this problem, one psycho-logical researcher observed, "the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; . . . a great deal of mischief has been associated with its use" (Bakan, 1966, p. 423).

Because SSTs have been so frequently misapplied, some reflective researchers (e.g., Carver, 1978; Meehl, 1978; Schmidt, 1996; Shulman, 1970) have recommended that SSTs be completely abandoned as a method for evaluating statistical results. In fact, Carver (1993) not only recommended abandoning statistical significance testing, but referred to it as a "corrupt form of the scientific method" (p. 288). In 1996, the American Psychological Association (APA) appointed its Task Force on Statistical Inference, which considered among other actions recommending less or even no use of statistical significance testing within APA journals

(Azar, 1997; Shea, 1996). Interestingly, in its draft report, the Task Force (Board of Scientific Affairs, 1996) noted that it "does not support any action that could be interpreted as banning the use of null hypothesis significance testing" (p. 1). Furthermore, SSTs still have support from a number of reflective researchers who acknowledge their limita-tions, but also see the value of the tests when appropri-ately applied. For example, Mohr (1990) reasoned, "one cannot be a slave to significance tests. But as a first approximation to what is going on in a mass of data, it is difficult to beat this particular metric for communication and versatility" (p. 74). In similar fashion, Huberty (1987) maintained, "there is nothing wrong with statistical tests themselves! When used as guides and indicators, as opposed to a means of arriving at definitive answers, they are okay" (p. 7).

"Statistical Significance" Versus "Importance"

A major controversy in the interpretation of SSTs has been "the ingenuous assumption that a statistically significant result is necessarily a noteworthy result" (Daniel, 1997, p. 106). Thoughtful social scientists (e.g., Berkson, 1942; Chow, 1988; Gold, 1969; Shaver, 1993; Winch & Campbell, 1969) have long recognized this problem. For example, even as early as 1931, Tyler had already begun to recognize a trend toward the misinter-pretation of statistical significance:

> The interpretations which have commonly been drawn from recent studies indicate clearly that we are prone to conceive of statistical significance as equivalent to social significance. These two terms are essentially different and ought not to be confused. . . . Differences which are statistically significant are not always socially important. The corollary is also true: differences which are not shown to be statistically significant may nevertheless be socially significant. (pp. 115-117)

A decade later, Berkson (1942) remarked, "statistics, as it is taught at present in the dominant school, consists almost entirely of tests of significance" (p. 325). Like-wise, by 1951, Yates observed, "scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective. Results are signifi-cant or not significant and this is the end of it" (p. 33). Similarly, Kish (1959) bemoaned the fact that too much of the research he had seen was presented "at the primitive level" (p. 338). Twenty years later, Kerlinger (1979) recognized that the problem still existed:

> statistical significance says little or nothing about the magnitude of a difference or of a relation. With a large number of subjects . . . tests of significance show statistical significance even when a difference between means is quite

> small, perhaps trivial, or a correlation coefficient is very small and trivial. . . . To use statistics adequately, one must understand the principles involved and be able to judge whether obtained results are statistically significant *and* whether they are meaningful in the particular research context. (pp. 318-319, emphasis in original)

Contemporary scholars continue to recognize the existence of this problem. For instance, Thompson (1996) and Pedhazur and Schmelkin (1991) credit the continuance of the misperception, in part, to the tendency of researchers to utilize and journals to publish manuscripts containing the term "significant" rather than "statistically significant"; thus, it becomes "common practice to drop the word 'statistical,' and speak instead of 'significant differences,' 'significant correlations,' and the like" (Pedhazur & Schmelkin, 1991, p. 202). Similarly, Schafer (1993) noted, "I hope most researchers understand that *significant* (statistically) and *important* are two different things. Surely the term *significant* was ill chosen" (p. 387, emphasis in original). Moreover, Meehl (1997) recently characterized the use of the term "significant" as being "cancerous" and "misleading" (p. 421) and advocated that researchers interpret their results in terms of confidence intervals rather than *p* values.

SSTs and Sample Size

Most tests of statistical significance utilize some test statistic (e.g., *F*, *t*, chi-square) with a known distribution. An SST is simply a comparison of the value for a particular test statistic based on results of a given analysis with the values that are "typical" for the given test statistic. The computational methods utilized in gene-rating these test statistics yield larger values as sample size is increased, given a fixed effect size. In other words, for a given statistical effect, a large sample is more likely to guarantee the researcher a statistically significant result than a small sample is. For example, suppose a researcher was investigating the correlation between scores for a given sample on two tests. Hypothesizing that the tests would be correlated, the researcher posited the null hypothesis that *r* would be equal to zero. As illustrated in Table 1, with an extremely small sample, even a rather appreciable *r*-value would not be statistically significant ($p < .05$). With a sample of only 10 persons, for example, an *r* as large as .6, indicating a moderate to large statistical effect, would not be statistically significant; by contrast, a negligible statistical effect of less than 1% ($r^2 = .008$) would be statistically significant with a sample size of 500!

Table 1
Critical Values of *r* for Rejecting the Null Hypothesis ($r = 0$) at the .05 Level Given Sample Size *n*

n	*r*
3	.997
5	.878
10	.632
20	.444
50	.276
100	.196
500	.088
1,000	.062
5,000	.0278
10,000	.0196

Note: Values are taken from Table 13 in Pearson and Hartley (1962).

As a second example, suppose a researcher is conducting an educational experiment in which students are randomly assigned to two different instructional settings and are then evaluated on an outcome achievement measure. This researcher might utilize an analysis of vari-ance test to evaluate the result of the experiment. Prior to conducting the test (and the experiment), the researcher would propose a null hypothesis of no difference between persons in varied experimental conditions and then compute an *F* statistic by which the null hypothesis may be evaluated. *F* is an intuitively-simple ratio statistic based on the quotient of the mean square for the effect(s) divided by the mean square for the error term. Since mean squares are the result of dividing the sum of squares for each effect by its degrees of freedom, the mean square for the error term will get smaller as the sample size is increased and will, in turn, serve as a smaller divisor for the mean square for the effect, yielding a larger value for the *F* statistic. In the present example (a two-group, one-way ANOVA), a sample of 302 would be five times as likely to yield a statistically significant result as a sample of 62 simply due to a larger number of error degrees of freedom (300 versus 60). In fact, with a sample as large as 302, even inordinately trivial differences between the two groups could be statistically significant considering that the *p* value associated with a large *F* will be small.

As these examples illustrate, an SST is largely a test of whether or not the sample is large, a fact that the researcher knows even before the experiment takes place. Put simply, "Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a

statistical test to evaluate whether there were a lot of subjects" (Thompson, 1992, p. 436). Some 60 years ago, Berkson (1938, pp. 526-527) exposed this circuitous logic based on his own observation of statistical significance values associated with chi-square tests with approximately 200,000 subjects:

> an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the *P*'s tend to come out small . . . and no matter how small the discrepancy between the normal curve and the true curve of observations, the chi-square *P* will be small if the sample has a sufficiently large number of observations it. . . . If, then, we know in advance the *P* that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all!

Misinterpretation of the Meaning of "Statistically Significant"

An analysis of past and current social science literature will yield evidence of at least six common misperceptions about the meaning of "statistically significant." The first of these, that "statistically significant" means "important," has already been addressed herein. Five additional misperceptions will also be discussed briefly: (a) the misperception that statistical significance informs the researcher as to the likelihood that a given result will be replicable ("the replicability fantasy" – Carver, 1978); (b) the misperception that statistical significance informs the researcher as to the likelihood that results were due to chance (or, as Carver [1978, p. 383] termed it, "the odds-against-chance fantasy"); (c) the misperception that a statistically significant result indicates the likelihood that the sample employed is representative of the population; (d) the misperception that statistical significance is the best way to evaluate statistical results; and (e) the misperception that statistically significant reliability and validity coefficients based on scores on a test administered to a given sample imply that the same test will yield valid or reliable scores with a different sample.

SSTs and replicability. Despite misperceptions to the contrary, the logic of statistical significance testing is NOT an appropriate means for assessing result replicability (Carver, 1978; Thompson, 1993a). Statistical significance simply indicates the probability that the null hypothesis is true in the population. However, Thompson (1993b) provides discussion of procedures that may provide an estimate of replicability. These procedures (cross validation, jackknife methods, and bootstrap methods) all involve sample-splitting logics and allow for the computation of statistical estimators across multiple configurations of the same sample in a single study. Even though these methods are biased to some degree (a single sample is utilized in each of the procedures), they represent the next best alternative to conducting a replication of the given study (Daniel, 1992a). Ferrell (1992) demonstrated how results from a single multiple regression analysis can be cross validated by randomly splitting the original sample and predicting dependent variable scores for each half of the sample using the opposite group's weights. Daniel (1989) and Tucker and Daniel (1992) used a similar logic in their analyses of the generalizability of results with the sophis-ticated "jackknife" procedure. Similar heuristic presenta-tions of the computer-intensive "bootstrap" logic are also available in the extant literature (e.g., Daniel, 1992a).

SSTs and odds against chance. This common misperception is based on the naive perception that statistical significance measures the degree to which results of a given SST occur by chance. By definition, an SST tests the probability that a null hypothesis (i.e., a hypothesis positing no relationship between variables or no difference between groups) is true in a given population based on the results of a sample of size *n* from that population. Consequently, "a test of significance provides the *probability of a result occurring by chance in the long run under the null hypothesis* with random sampling and sample size *n*; it provides *no basis for a conclusion about the probability that a given result is attributable to chance*" (Shaver, 1993, p. 300, emphasis added). For example, if a correlation coefficient *r* of .40 obtained between scores on Test X and Test Y for a sample of 100 fifth graders is statistically significant at the 5% ($\alpha = .05$) level, one would appropriately conclude that there is a 95% likelihood that the correlation between the tests in the population is not zero assuming that the sample employed is representative of the population. However, it would be *inappropriate* to conclude (a) that there is a 95% likelihood that the correlation is .40 in the population or (b) that there is only a 5% likelihood that the result of that particular

statistical significance test is due to chance. This fallacy was exposed by Carver (1978):

> the *p* value is the probability of getting the research results when it is first assumed that it is actually true that chance caused the results. It is therefore impossible for the *p* value to be the probability that chance caused the mean difference between two research groups since (a) the *p* value was calculated by assuming that the probability was 1.00 that chance did cause the mean difference, and (b) the *p* value is used to decide whether to accept or reject the idea that probability is 1.00 that chance caused the mean difference. (p. 383)

SSTs and sampling. This misperception states that the purpose of statistical significance testing is to determine the degree to which the sample represents the population. Representativeness of the sample cannot be evaluated with an SST; the only way to estimate if a sample is representative is to carefully select the sample. In fact, the statistical significance test is better conceptualized as answering the question, "*If* the sample represents the population, how likely is the obtained result?"

SSTs and evaluation of results. This misperception, which states that the best (or correct) way to evaluate the statistical results is to consult the statistical significance test, often accompanies the "importance" misperception but actually may go a step beyond the importance misperception in its corruptness. The importance misperception, as previously noted, simply places emphasis on the wrong thing. For example, the researcher might present a table of correlations, but in interpreting and discussing the results, only discuss whether or not each test yielded a statistically significant result, making momentous claims for statistically significant correlations no matter how small and ignoring statistically nonsignificant values no matter how large. In this case, the knowledgeable reader could still look at the correlations and draw more appropriate conclusions based on the magnitude of the *r* values. However, if the researcher were motivated by the "result evaluation" misperception, he or she might go so far as to fail to report the actual correlation values, stating only that certain relationships were statistically signifi-cant. Likewise, in the case of an analysis of variance, this researcher might simply report the *F* statistic and its *p* value without providing a breakdown of the dependent variable sum of squares from which an estimate of effect size could be determined. Thompson (1989a, 1994) discussed several suggestions for improvement of these practices, including the reporting of (a) effect sizes for all parametric analyses and (b) "what if" analyses "indicating at what different sample size a given fixed effect would become statistically significant or would have no longer been statistically significant" (1994, p. 845). In regard to (b), Morse (1998) has designed a PC-compatible comput-er program for assessing the sensitivity of results to sample size. Moreover, in the cases in which statistically nonsignificant results are obtained, researchers should consider conducting statistical power analyses (Cohen, 1988).

SSTs and test score characteristics. Validity and reliability are characteristics of test *scores* or test *data*. However, contemporary scholarly language (e.g., "the test is reliable," "the test is valid") often erroneously implies that validity and reliability are characteristics of tests themselves. This fallacious use of language is sometimes accompanied by another fallacy related to statistical significance testing, namely, the use of null hypothesis SSTs of reliability or validity coefficients. Statistical tests of these coefficients are nonsensical. As Witta and Daniel (1998) noted:

> In the case of a reliability coefficient, these statistical significance tests evaluate the null hypothesis that a set of scores is totally unreliable, a hypothesis that is meaningless considering that large reliability or validity coefficients may often be statistically significant even when based on extremely small samples (Thompson, 1994) whereas minute reliability or validity coefficients will eventually become statistically significant if the sample size is increased to a given level (Huck & Cormier, 1996). Further, considering that reliability and validity coefficients are sample specific, statistical significance tests do not offer any promise of the generalizability of these coefficients to other samples. (pp. 4-5)

Journal Policies and Statistical Significance

As most educational researchers are aware, social sci-ence journals have for years had a bias towards accepting manuscripts documenting statistically significant findings and rejecting those with statistically nonsignificant findings. One editor even went so far as to boast that he had made it a practice to avoid accepting for publication results that were statistically significant at the .05 level, desiring instead that results reached at least the .01 level (Melton, 1962). Because of this

editorial bias, many researchers (e.g., Mahoney, 1976) have paid homage to SSTs in public while realizing their limitations in private. As one observer noted a generation ago, "Too, often . . . even wise and ingenious investigators, for varieties of reasons, not the least of which are the editorial policies of our major psychological journals, . . . tend to credit the test of significance with properties it does not have" (Bakan, 1966, p. 423).

According to many researchers (e.g., Neuliep, 1991; Shaver, 1993), this bias against studies that do not report statistical significance or that present results that did not meet the critical alpha level still exists. Shaver (1993) eloquently summarized this problem:

> Publication is crucial to success in the academic world. Researchers shape their studies, as well as the manuscripts reporting the research, according to accepted ways of thinking about analysis and interpretation and to fit their perceptions of what is publishable. To break from the mold might be courageous, but, at least for the untenured faculty member with some commitment to self-interest, foolish. (p. 310)

Because this bias is so prevalent, it is not uncommon to find examples in the literature of studies that report results that are statistically nonsignificant with the disclaimer that the results "approached significance." Thompson (1993a) reported a somewhat humorous, though poignant, response by one journal editor to this type of statement: "How do you know your results were not working very hard to *avoid* being statistically significant?" (p. 285, emphasis in original).

Likewise, results that are statistically significant at a conservative alpha level (e.g, .001), are with some frequency referred to as "highly significant," perhaps with the authors' intent being to make a more favorable impression on some journal editors and readers than they could make by simply saying that the result was statistically significant, period. This practice, along with the even more widespread affinity for placing more and more zeroes to the right of the decimal in an attempt to make a calculated *p* appear more noteworthy, has absolutely nothing to do with the practical significance of the result. The latter practice has often been the focus of tongue-in-cheek comments. For example, Popham (1993) noted, "Some evaluators report their probabilities so that they look like the scoreboard for a no-hit baseball game (e.g., $p < .000000001$)" (p. 266); Campbell (1982) quipped, "It is almost impossible to drag authors away from their *p* values, and the more zeroes after the decimal point, the harder people cling to them" (p. 698); and McDonald (1985), referring to the tendency of authors to place varying numbers of stars after statistical results re-ported in tabular form as a means for displaying differing levels of statistical significance, bantered that the practice resembled "grading of hotels in guidebooks" (p. 20).

If improvements are to be made in the interpretation and use of SSTs, professional journals (Rozeboom, 1960), and, more particularly, their editors will no doubt have to assume a leadership role in the effort. As Shaver (1993) articulated it, "As gatekeepers to the publishing realm, journal editors have tremendous power. . .[and perhaps should] become crusaders for an agnostic, if not atheistic, approach to tests of statistical significance" (pp. 310-311). Hence, Carver (1978, 1993) and Kupfersmid (1988) suggested that journal editors are the most likely candidates to promote an end to the misuse and misinterpretation of SSTs.

Considering this, it is encouraging to note that at least some journals have begun to adopt policies relative to statistical significance testing that address some of the problems discussed here. For several years, *Measurement and Evaluation in Counseling and Development* (1992, p. 143) has included three specific (and appropriate) author guidelines related to statistical significance testing, including the encouragement for authors to (a) index results of SSTs to sample size, (b) provide readers with effect size estimates as well as SSTs, and (c) provide power estimates of protection against Type II error when statistically nonsignificant results are obtained.

Educational and Psychological Measurement (*EPM*) has developed a similar set of editorial policies (Thompson, 1994) which are presently in their fourth year of implementation. These guidelines do not for the most part ban the use of SSTs from being included in authors' manuscripts, but rather request that authors report other information along with the SST results. Specifically, these editorial guidelines include the following:

1. Requirement that authors use "statistically signi-ficant" and not merely "significant" in discussing results.
2. Requirement that tests of statistical significance generally NOT accompany validity and reliability coefficients (Daniel & Witta, 1997; Huck & Cormier, 1996; Witta & Daniel, 1998). This is the one scenario in which SSTs are expressly forbidden according to *EPM* editorial policy.

3. Requirement that all statistical significance tests be accompanied by effect size estimates.
4. Suggestion that authors may wish to report the "what if" analyses alluded to earlier. These analyses should indicate "at what different sample size a given fixed effect would become statistically significant or would have no longer been statistically significant" (Thompson, 1994, p. 845).
5. Suggestion that authors report external replicability analyses via use of data from multiple samples or else internal replicability analyses via use of cross-validation, jackknife, or bootstrap procedures.

A number of efforts have been utilized by the *EPM* editors to help both authors and reviewers become familiar with these guidelines. For the first two years that these guidelines were in force, copies of the guidelines editorial (Thompson, 1994) were sent to every author along with the manuscript acceptance letter. Although copies are no longer sent to authors, the current manuscript acknowledgment letter includes a reference to this and two other author guidelines editorials the journal has published (Thompson, 1995; Thompson & Daniel, 1996a), and it directs authors to refer to the several editorials to determine if their manuscripts meet editorial policy. More recently, the several editorials have been made available via the Internet at Web address: "http://acs.tamu.edu/~bbt6147/".

In addition to this widescale distribution policy, the guidelines are referenced on each review form (see Appendix) sent to the masked reviewers. As a part of the review process, reviewers must determine if manuscripts contain material that is in violation of the editorial policies relative to statistical significance testing and several other methodological issues. To assure that reviewers will take this responsibility seriously, several questions relative to the guidelines editorials are included on the review form and must be answered by the reviewers. No manuscripts are accepted for publication by either of the two current editors if they violate these policies, although these violations do not necessarily call for outright rejection of the first draft of a manuscript. It is the hope of the editors that this comprehensive policy will over time make a serious impact on *EPM* authors' and readers' ideas about correct practice in reporting the results of SSTs.

More recently, two additional journals have adopted editorial policies that are likely to prompt additional scrutiny of the reporting and interpretation of SSTs. The current author guidelines of the *Journal of Experimental Education* (Heldref Foundation, 1997) indicate that "authors are *required* to report and interpret magnitude-of-effect measures in conjunction with every *p* value that is reported" (pp. 95-96, emphasis added). Further, the editor of one of the APA journals, *Journal of Applied Psychology*, recently stated:

> If an author decides not to report an effect size estimate along with the outcome of a [statistical] significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

Recommendations for Journal Editors

As the previous discussion has illustrated, there is a trend among social science journal editors to either reject or demand revision of manuscripts in which authors employ loose language relative to their interpretations of SSTs or else overinterpret the results of these tests; however, more movement of the field toward this trend is needed. Pursuant to the continued movement toward this trend, the following ten recommendations are offered to journal editors and scholars at large as a means for encouraging better practices in educational journals and other social science journals.

1. *Implement editor and reviewer selection policies*. First, following the suggestions of Carver (1978, 1993) and Shaver (1993), it would be wise for professional associations and publishers who hire/appoint editors for their publications to require potential editors to submit statements rel-ative to their positions on statistical significance testing. Journal editors might also require a sim-ilar statement from persons who are being con-sidered as members of editorial review boards.
2. *Develop guidelines governing SSTs*. Each editor should adopt a set of editorial guidelines that will promote correct practice relative to the use of SSTs. The *Measurement and Evaluation in Counseling and Development* and *Educational and Psychological Measurement* guidelines referenced in this paper could serve as a model for policies developed for other journals.

3. *Develop a means for making the policies known to all involved*. Editors should implement a mechanism whereby authors and reviewers will be likely to remember and reflect upon the policies. The procedures mentioned previously that are currently utilized by the editors of *Educational and Psychological Measurement* might serve as a model that could be adapted to the needs of a given journal.
4. *Enforce current APA guidelines for reporting SSTs*. Considering that most journals in education and psychology utilize APA publication guidelines, editors could simply make it a requirement that the guidelines for reporting results of SSTs included in the fourth edition *Publication Manual of the American Psychological Association* (APA, 1994, pp. 17-18) be followed. Although the third edition *Publication Manual* was criticized for using statistical signi-ficance reporting examples that were flawed (Pedhazur & Schmelkin, 1991; Shaver, 1993), the fourth edition includes appropriate examples as well as suggestions encouraging authors to report effect size estimates.
5. *Require authors to use "statistically" before "significant."* Despite the fact that some journal editors will be resistant to the suggestion (see, for example, Levin's [1993; Robinson & Levin, 1997] criticism that such a practice smacks of policing of language), requiring authors to routinely use the term "statistically significant" rather than simply "significant" (cf. Carver, 1993; Cohen, 1994; Daniel, 1988; Shaver, 1993; Thompson, 1996) when referring to research findings will do much to minimize the "statistical significance as importance" problem and to make it clear where the author intends to make claims about the "practical significance" (Kirk, 1996) of the results.
6. *Require effect size reporting*. Editors should require that effect size estimates be reported for all quantitative analyses. These are strongly *suggested* by APA (1994); however, Thompson (1996, p. 29, emphasis in original) advocated that other professional associations that publish professional journals "venture beyond APA, and *require* such reports in all quantitative analyses."
7. *Encourage or require replicability and "what if" analyses*. As previously discussed, replicability analyses provide reasonable evidence to support (or disconfirm) the generalizability of the find-ings, something that SSTs do NOT do (Shaver, 1993; Thompson, 1994). "What if" analyses, if used regularly, will build in readers and authors a sense of always considering the sample size when conducting SSTs, and thereby considering the problems inherent in particular to cases involving rather larger or rather small samples.
8. *Require authors to avoid using SSTs where they are not appropriate*. For example, as previously noted, *EPM* does not allow manuscripts to be published if SSTs accompany certain validity or reliability coefficients.
9. *Encourage or require that power analyses or replicability analyses accompany statistically nonsignificant results*. These analyses allow for the researcher to address power considerations or to determine if a result with a small sample has evidence of stability in cases in which an SST indicates a statistically nonsignificant result.
10. *Utilize careful copyediting procedures.* Careful copyediting procedures will serve to assure that very little sloppy language relative to SSTs will end up in published manuscripts. In addition to the suggestions mentioned above, editors will want to make sure language such as "highly significant" and "approaching significance" is edited out of the final copies of accepted manuscripts.

References

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

Azar, B. (1997). APA task force urges a harder look at data. *APA Monitor*, *28*(3), 26.

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

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526-536.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325-335.

Board of Scientific Affairs. (1996). *Task Force on Statistical Inference initial report (DRAFT)* [On-line]. Available: http://www.apa.org/science/tsfi/html

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology*, *67*, 691-700.

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

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

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, *70*, 426-443.

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

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

Daniel, L. G. (1988). [Review of *Conducting educational research* (3rd ed.)]. *Educational and Psychological Measurement*, *48*, 848-851.

Daniel, L. G. (1989, January). *Use of the jackknife statis-tic to establish the external validity of discriminant analysis results*. Paper presented at the annual meet-ing of the Southwest Educational Research Asso-ciation, Houston. (ERIC Document Reproduction Service No. ED 305 382)

Daniel, L. G. (1992a, April). *Bootstrap methods in the principal components case*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 346 135)

Daniel, L. G. (1992b, November). *Perceptions of the quality of educational research throughout the twentieth century: A comprehensive review of the literature*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Daniel, L. G. (1997). Kerlinger's research myths: An overview with implications for educational researchers. *Journal of Experimental Education*, *65*, 101-112.

Daniel, L. G., & Witta, E. L. (1997, March). *Implications for teaching graduate students correct terminology for discussing validity and reliability based on a content analysis of three social science measurement journals*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 408 853)

Ferrell, C. M. (1992, February). *Statistical significance, sample splitting and generalizability of results*. Paper presented at the annual meeting of the Southwest Educational Research Association. (ERIC Document Reproduction Service No. ED 343 935)

Gold, D. (1969). Statistical tests and substantive significance. *American Sociologist*, *4*, 42-46.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education*, *65*, 95-96.

Henkel, C. G. (1976). *Tests of significance*. Newbury Park, CA: Sage.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher*, *16*(8), 4-9.

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*, 317-333.

Huck, S. W., & Cormier, W. G. (1996). *Reading statistics and research* (2nd ed.). New York: HarperCollins.

Kerlinger, F. N. (1979). *Behavioral research: A conceptual approach*. New York: Holt, Rinehart and Winston.

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

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

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist*, *43*, 635-642.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.

Mahoney, M. J. (1976). *Scientist as subject: The psycho-logical imperative*. Cambridge, MA: Ballinger.

McDonald, R. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

Measurement and Evaluation in Counseling and Development. (1992). Guidelines for authors. *Measurement and Evaluation in Counseling and Development*, *25*, 143.

Meehl, P. (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.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence

intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393-426). Mahwah, NJ: Erlbaum.

Melton, A. (1962). Editorial. *Journal of Experimental Psychology*, *64*, 553-557.

Mohr, L. B. (1990). *Understanding significance testing*. Newbury Park, CA: Sage.

Morse, D. T. (1998). MINSIZE: A computer program for obtaining minimum sample size as an indicator of effect size. *Educational and Psychological Measurement*, *58*, 142-153.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology*, *82*, 3-5.

Neuliep, J. W. (Ed.). (1991). *Replication in the social sciences*. Newbury Park, CA: Sage.

Pearson, E. S., & Hartley, H. O. (Eds.). (1962). *Biometrika tables for statisticians* (2nd ed.). Cambridge, MA: Cambridge University Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Popham, W. J. (1993). *Educational evaluation* (3rd ed.). Boston, MA: Allyn and Bacon.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21-26.

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

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

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

Shaver, J. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Educa-tion*, *61*, 293-316

Shea, C. (1996). Psychologists debate accuracy of "significance" test. *Chronicle of Higher Education*, *42*(9), A12,A19.

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

Thompson, B. (1989a). Asking "what if" questions about significance tests. *Measurement and Evaluation in Counseling and Development*, *22*, 66-67.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development*, *22*, 2-6.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, *70*, 434-438.

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

Thompson, B. (1993b). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361-377.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, *55*, 525-534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26-30.

Thompson, B. (1998). [Review of *What if there were no significance tests?*]. *Educational and Psychological Measurement*, *58*, 334-346.

Thompson, B., & Daniel, L. G. (1996a). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, *56*, 197-208.

Thompson, B., & Daniel, L. G. (1996b). Seminal readings on reliability and validity: A "hit parade" bibliography. *Educational and Psychological Measurement*, *56*, 741-745.

Tucker, M. L., & Daniel, L. G. (1992, January). *Investigating result stability of canonical function equations with the jackknife technique*. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 343 914)

Tyler, R. W. (1931). What is statistical significance? *Educational Research Bulletin*, *10*, 115-118,142.

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

Witta, E. L., & Daniel, L. G. (1998, April). *The reliability and validity of test scores: Are editorial policy changes reflected in journal articles*? Paper

presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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.

APPENDIX
EPM MANUSCRIPT REVIEW FORM

epmrevew.new

Educational and Psychological Measurement

Manuscript Review Form

Reviewer Code #_____ MS #_____

Due Date: ____/____/____

Omit criteria that are not relevant in evaluating a given ms. Return the rating sheet and comments to the appropriate Editor in the attached return envelope.

Manuscripts under review should be treated as confidential, proprietary information (not to be cited, quoted, etc.). After review, the ms should be discarded.

Part I ("N.A." = Not Applicable) *Criteria associated with the editorials in the Winter, 1994 (vol. 54, no. 4), August, 1995 (vol. 55, no. 4), and April, 1996 (vol. 56, no. 2) issues. Guidelines editorials are also available on the Internet at Web address "http://acs.tamu.edu/~bbt6147/":*

YES	NO	N.A.	For each reported statistical significance test, is an effect size also reported?
YES	*NO*	N.A.	Is a null hypothesis test of no difference used to evaluate measurement statistics (e.g., concurrent validity or score reliability)?
YES	NO	N.A.	If statistical significance tests are reported, were "what if" analyses of sample sizes presented?
YES	*NO*	N.A.	In discussing score validity or reliability, do the au(s) ever use inappropriate language (e.g., "the test was reliable" or "the test was valid")?
YES	NO	N.A.	If statistically non-significant results were reported, was either a power analysis or a replicability analysis reported?
YES	*NO*	N.A.	Was a stepwise analysis conducted?

Part II *General Criteria*

Worst	1	2	3	4	5	Best	Noteworthiness of Problem
Worst	1	2	3	4	5	Best	Theoretical Framework
Worst	1	2	3	4	5	Best	Adequacy of Sample
Worst	1	2	3	4	5	Best	Appropriateness of Method
Worst	1	2	3	4	5	Best	Insightfullness of Discussion
Worst	1	2	3	4	5	Best	Writing Quality

Part III *Overall recommendation.* Check one of the following seven recommendations.

Reject Now.

_____ Even with substantial revision, the ms. is unlikely to meet **EPM** standards.

_____ The ms. is not appropriate for **EPM**. A more appropriate journal would be:

_____

Accept Now.

_____ An important contribution. Accept "as is" or with very minor revisions.

_____ An important contribution, but needs specific revisions. *Tentatively* accept pending revisions reviewed by the editor.

Marginal: A decision can be made now.

_____ A sound contribution. Publish if EPM has space.

Request revision from author: Decision cannot be made now. (Note: "Full review" involves review of the revision by all initial referees).

_____ Likely to be an important contribution if suitably revised. Encourage major revision with full review of the revision.

_____ May possibly be an important contribution if suitably revised. Allow revision, require full review of the revision.

Based on the quality of the present draft of the manuscript, what is the likelihood that the author will produce an acceptable revision?

_____ 10% _____ 30% _____ 50%

_____ 70% _____ 90%

Part IV *Please provide the au(s) with constructive suggestions, helpful references, and related comments,* attaching additional sheets as needed.

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 33-38

Statistical Significance and Effect Size Reporting: Portrait of a Possible Future

Bruce Thompson
Texas A&M University and Baylor College of Medicine

The present paper comments on the matters raised regarding statistical significance tests by three sets of authors in this issue. These articles are placed within the context of contemporary literature. Next, additional empirical evidence is cited showing that the APA publication manual's "encouraging" effect size reporting has had no appreciable effect. Editorial policy will be required to affect change, and some model policies are quoted. Science will move forward to the extent that both effect size and replicability evidence of one or more sorts are finally seriously considered within our inquiry.

I appreciate the opportunity to comment on matters raised by Daniel (1998), McLean and Ernest (1998), and Nix and Barnette (1998) as regards statistical significance tests. Theme issues of journals such as the present one (see also Thompson (1993)) allow various perspectives to be articulated and help slowly but inexorably move the field toward improved practices. Of course, an important recent book (Harlow, Mulaik, & Steiger, 1997) also presents diverse perspectives regarding these continuing controversies (for reviews see Levin (1998) and Thompson (1998c)).

At the outset perhaps I should acknowledge possible conflicts of interest. First, co-editor Kaufman asked me to serve as one of the five or so referees who read each of these manuscripts in their initial form. Second, in a somewhat distant past, prior to his ascending to tenure, full professorship, and directorship of a research center, I chaired Larry Daniel's dissertation committee at the Uni-versity of New Orleans (boy, does reciting these facts make me feel old!).

These Articles and My Views in Context

It might be helpful to readers to frame these three articles, and my own views, within the context of views presented within the literature. Certainly at one extreme some authors (cf. Carver, 1978; Schmidt, 1996) have argued that statistical significance tests should be banned from publications. For example, Rozeboom (1997) recently argued that:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students . . . [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism . . . (p. 335)

Schmidt and Hunter (1997), virulent critics of statistical significance testing, similarly argued that, "Statistical significance testing retards the growth of scientific knowledge; it *never* makes a positive contribution" (p. 37, emphasis added).

At the other extreme (cf. Cortina & Dunlap, 1997; Frick, 1996), Abelson (1997) argued that, "Significance tests fill an important need in answering some key research questions, and if they did not exist they would have to be invented" (p. 118). Similarly, Harris (1997) argued that

> Null hypothesis significance testing (NHST) as applied by most researchers and journal editors can provide a very useful form of social control over researchers' understandable tendency to "read too much" into their data . . . [E]ven NHST alone would be an improvement over the current lack of attention to sampling error. (pp. 145, 164)

Bruce Thompson is a professor and distinguished research scholar in the Department of Educational Psychology at Texas A & M University. He is also an adjunct professor of community medicine at the Baylor College of Medicine. Correspondence regarding this article should be addressed to Bruce Thompson, Department of Educational Psychology, Texas A & M University, College Station, TX 77843-4225 or by e-mail to e100bt@tamvm1.tamu.edu. Related reprints and the author can be accessed on the Internet via URL: "**http://acs.tamu.edu/~bbt6147/**".

Some of these defenses of statistical tests have been thoughtful, but others have been flawed (Thompson, 1998b).

I see Nix and Barnette (1998) as somewhat approaching the Carver (1978)/Rozeboom (1997) end of the continuum. They "believe that most statisticians would [and seemingly should] welcome orderly change that would lead to abandonment of NHST." The authors feel constrained from supporting a ban, not on the merits, but only because of concerns regarding "democratic principles" and "censorship and infringement on individual freedoms."

McLean and Ernest (1998) believe that "our recommendations reflect a moderate mainstream approach." Certainly, their views are intellectually "moderate." A call that their views are "mainstream" requires a factual judgment as regards a moving target – our moving discipline. McLean and Ernest (1998) suggest that tests of statistical significance "must be accompanied by judgments of the event's practical significance and replicability."

I also see Daniel's (1998) views as being moderate, though they may tend a bit more toward the Carver (1978)/Rozeboom (1997) end of the continuum. Thus, the three articles do not include advocacy that the status quo is peachy-keen, and that no changes are warranted (a deficiency that will doubtless be corrected via additional commentaries).

My own views are fairly similar to those of McLean and Ernest (1998) and Daniel (1998). That is, on numerous occasions I certainly have pointed out the myriad problems with rampant misuse and misinterpretation of statistical tests.

However, I have never argued that statistical significance tests should be banned. If I felt these tests were intrinsically evil, as an editor of three journals, I necessarily would have written author guidelines proscribing these tests. And as an author I would also never report *p* values.

Instead, I generally find statistical tests to be largely irrelevant. Like Cohen (1994), I do not believe that *p* values evaluate the probability of what we want to know (i.e., the population). Rather, we assume the null hypothesis describes the population, and then evaluate the probability of the sample results (Thompson, 1996).

I am especially disinterested in statistical tests when what Cohen (1994) termed "nil" null hypotheses are used, particularly when testing reliability or validity coefficients. Daniel (1998) makes some excellent points here. We expect reliability and validity coefficients to be .7 or .8. As his table shows, with a *n* of 10 or 15, we will always attain statistical significance even for minimally acceptable reliability and validity coefficients, so what is the value of such tests with these or larger sample sizes? Abelson (1997) put the point fairly clearly:

> And when a reliability coefficient is declared to be nonzero, that is the ultimate in stupefyingly vacuous information. What we really want to know is whether an estimated reliability is .50'ish or .80'ish. (p. 121)

Thus, editorial policies of *Educational and Psychological Measurement* proscribe use of statistical testing of reliability and validity coefficients, if (and only if) "nil" nulls are used for this purpose.

I believe that evidence of result replicability is very important and is ignored by those many people who do not understand what statistical tests do (e.g., believe that their tests evaluate the probability of the population). Daniel (1998) at one point says, "Statistical significance simply indicates the probability that the null hypothesis is true in the population" (a view I do not accept), but says later that these tests answer the question, "*If* the sample represents the population, how likely is the obtained [sample] result?" (a view I do endorse).

Empirical studies consistently show that many researchers do not fully understand the logic of statistical tests (cf. Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Similarly, many textbooks teach misconceptions regarding these tests (Carver, 1978; Cohen, 1994).

More than anything else, I especially want to see authors always report effect sizes. I concur with the views of McLean and Ernest (1998), who noted that, "In reviewing the literature, the authors were unable to find an article that argued against the value of including some form of effect size or practical significance estimate in a research report." Kirk (1996) and Snyder and Lawson (1993) present helpful reviews of the many types of effect sizes that can be computed.

Regarding effect sizes, some (cf. Robinson & Levin, 1997) have argued that we must always first test statistical significance, and if results are statistically significant, "only if so: (2) effect size information should be provided" (Levin & Robinson, in press).

In Thompson (in press-b) I used a hypothetical to portray the consequences of this view. Two new proteins that suppress cancer metastasis and primary tumor growth in mice are discovered. Two hundred teams of researchers begin clinical trials with humans. Unfortunately, the 200 studies are underpowered, because the researchers slightly overestimate expected effects, or

perhaps because the researchers err too far in their fears of "over-powering" (Levin, 1997) their studies. Low and behold, all 200 studies yield noteworthy "moderate" effects for which $p_{CALCULATED}$ values are all .06.

> [A]m I to understand that these moderate effect sizes involving a pretty important criterion variable may not permissibly be discussed or even reported? . . . In the Thompson world, . . . [i]n this happy example, considerable direct replication evidence is available, so the noteworthy effect is interpreted even though none (zero, nada) of the 200 results is statistically significant. Thus, this is a world in which, in at least some cases, 'surely, God loves the .06 nearly as much as the .05' level of statistical significance (Rosnow & Rosenthal, 1989, p. 1277). (Thompson, in press-b)

Effect Size Reporting

Nix and Barnette (1998) cite others in suggesting that "studies today are more likely to report effect sizes," perhaps because the APA (1994) publication manual "encourages" (p. 18) such reports. However, McLean and Ernest (1998, emphasis in original) diametrically disagree, arguing that "encouraging" effect size reporting "has *not* appreciably affected actual reporting practices," and then cite five *empirical* studies corroborating their views.

Most regrettably, I believe that the pessimistic views of McLean and Ernest (1998) are correct. Indeed, let me cite five additional *empirical* studies of journal reporting practices that present similar findings (Keselman et al., in press; Lance & Vacha-Haase, 1998; Ness & Vacha-Haase, 1998; Nilsson & Vacha-Haase, 1998; Reetz & Vacha-Haase, 1998). In fact, Keselman et al. (in press) concluded that, "as anticipated, effect sizes were almost never reported along with *p*-values."

I have offered various reasons why the APA "encouragement" has been such a failure. First, an "encouragement" is too vague to enforce. Second, the APA policy

> presents a self-canceling mixed-message. To present an "encouragement" in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "these myriad requirements count, this encouragement doesn't." (Thompson, in press-b)

Of course, mindless adherence to old habits may also partly explain the glacial movement of the field, because "changing the beliefs and practices of a lifetime . . . naturally . . . provokes resistance" (Schmidt & Hunter, 1997, p. 49). As Rozeboom (1960) observed nearly 40 years ago, "the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

It is my view (Thompson, 1998a; Vacha-Haase & Thompson, 1998) that most authors will simply not change their practices until editorial policies *require* them to do so. These three sets of authors cite three editorial policies (Heldref Foundation, 1997; Murphy, 1997; Thompson, 1994) requiring effect size reporting. Here are some additional editorial policies on this point. [Should *RESEARCH IN THE SCHOOLS* adopt such a policy? Hint. Hint.]

The editor of the *Journal of Consulting and Clinical Psychology* noted in passing that effect sizes are required in that journal, and furthermore that

> Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the *required* effect size but also a consideration of clinical significance. That is, . . . it is also important for the evaluator to consider the degree to which the outcomes are clinically significant (e.g., normative comparisons). . . . A treatment that produces a significant reduction in depressed mood must also be examined to determine whether the reduction moved participants from within to outside the defining boundary of scores for depression. (Kendall, 1997, p. 3, emphasis added)

The editor of the *Journal of Educational Psychology* called for "the provision of both strength-of-relationship measures and 'sufficient statistics' (the latter to permit independent confirmation of a study's statistical findings, statistical power calculations, and access to relevant information for meta-analyses, among others)" (Levin, 1995, p. 3).

The editor of the *Journal of Family Psychology* argued that, "In addition, reporting clinical significance . . . as opposed to mere statistical significance would also make treatment research more relevant to practitioners" (Levant, 1992, p. 6). Finally, the editor of the *Journal of*

Experimental Psychology: Learning, Memory, and Cognition argued that

> In reporting results, authors should still provide measures of variability and address the issue of the generalizability and reliability of their empirical findings across people and materials. There are a number of acceptable ways to do this, including reporting MSEs and confidence intervals and, in case of within-subject or within-items designs, the number of people or items that show the effect in the reported direction. (Neeley, 1995, p. 261)

Highlights of the Three Articles

The three articles each had highlights that particularly appealed to me. For example, Nix and Barnette (1998) present a nice albeit short review of the controversies between Fisher as against Neyman and Pearson, which were never effectively resolved (the consequence of this failed resolution being the hodge-podge of practices we see today). I very much liked their statement, "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." As I have noted elsewhere,

> The calculated *p* values in a given study are a function of several study features, but are particularly influenced by the confounded, joint influence of study sample size and study effect sizes. Because *p* values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $p_{CALCULATED}$, and 100 studies with the same single effect size could each have 100 different values for $p_{CALCULATED}$. (Thompson, in press-a)

Daniel (1998) does a nice job of presenting older quotations to illustrate that we have been haunted by these controversies virtually since the inception of statistical tests. I particularly liked his citation of Berkson, arguing in 1938 that testing significance when the *n* is 200,000 is not very enlightening!

Daniel's (1998) review of editorial policies and how they are applied was also informative. He emphasizes a point that some authors do not appreciate: editors will not accept articles that violate their published editorial policies, so prudent authors must take these policies seriously. I find myself in general agreement with Daniel's (1998) very specific recommendations for improving our scholarship.

As regards McLean and Ernest (1998), I very much appreciated their recognition that science is subjective and that statistical tests cannot make it otherwise (Thompson, in press-c). I also very much liked their treatment of the "language controversy."

McLean and Ernest (1998) prefer to keep statistical tests within the researcher's arsenal but are more than willing to provide both effect size and replicability evidence of one or more sorts. I am somewhat less interested than they in the results of statistical tests, but science will move forward to the extent that the latter two issues are finally seriously considered within our inquiry.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

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

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

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

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

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2) 23-32.

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

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145-174). Mahwah, NJ: Erlbaum.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.

Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, *15*, 3-5.

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (in press). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research.*

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

Lance, T., & Vacha-Haase, T. (1998, August). *The Counseling Psychologist: Trends and usages of statistical significance testing*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Levant, R. F. (1992). Editorial. *Journal of Family Psychology*, *6*, 3-9.

Levin, J. R. (1995). Editorial: Journal alert! *Journal of Educational Psychology*, *87*, 3-4.

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging*, *12*, 84-106.

Levin, J. R. (1998). To test or not to test H_0? *Educational and Psychological Measurement, 58*, 311-331.

Levin, J. R., & Robinson, D. H. (in press). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools*, *5*(2)15-22.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology*, *82*, 3-5.

Neeley, J. H. (1995). Editorial. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 261.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, *41*, 1299-1301.

Ness, C., & Vacha-Haase, T. (1998, August). *Statistical significance reporting: Current trends and usages within Professional Psychology: Research and Practice*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Nilsson, J., & Vacha-Haase, T. (1998, August). *A review of statistical significance reporting in the Journal of Counseling Psychology*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, *5*(2)3-14.

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

Reetz, D., & Vacha-Haase, T. (1998, August). *Trends and usages of statistical significance testing in adult development and aging research: A review of Psychology and Aging*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21-26.

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

Rosnow, R. L., & 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.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-392). Mahwah, NJ: Erlbaum.

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

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

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

Thompson, B. (Guest Ed.). (1993). Special issue on statistical significance testing, with comments from various journal editors. *Journal of Experimental Education, 61*(4).

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26-30.

Thompson, B. (1998a, April). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas*. Invited address presented at the annual meeting of the American Educational Research Association, San Diego.

Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Thompson, B. (1998c). Review of *What if there were no significance tests?* by L. Harlow, S. Mulaik & J. Steiger (Eds.). *Educational and Psychological Measurement, 58*, 332-344.

Thompson, B. (in press-a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*. [Invited address presented at the 1997 annual meeting of the American Psychological Association, Chicago.]

Thompson, B. (in press-b). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*.

Thompson, B. (in press-c). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology*.

Vacha-Haase, T., & Thompson, B. (1998, August). *APA editorial policies regarding statistical significance and effect size: Glacial fields move inexorably (but glacially)*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49-53.

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 39-41

Comments on the Statistical Significance Testing Articles

Thomas R. Knapp
The Ohio State University

This review assumes a middle-of-the-road position regarding the controversy. The author expresses that significance tests have their place, but generally prefers confidence intervals. His remarks concentrate on ten errors of commission or omission that, in his opinion, weaken the arguments. These possible errors include using the jackknife and bootstrap procedures for replicability purposes, omitting key references, misrepresenting the null hypothesis, omitting the weaknesses of confidence intervals, ignoring the difference between a hypothesized effect size and an obtained effect size, erroneously assuming a linear relationship between p and F, claiming Cohen chose power level arbitrarily, referring to the "reliability of a study," inferring that inferential statistics are primarily for experiments, and recommending "what if" analyses.

Since I take a middle-of-the-road position regarding the significance testing controversy (I think that significance tests have their place, I generally prefer confidence intervals, and I don't like meta-analysis!), I would like to concentrate my remarks on ten errors of commission or omission that in my opinion weaken the arguments in these otherwise thoughtful papers. In this article, the three articles under review are referred to as Daniel (1998), McLean and Ernest (1998), and Nix and Barnette (1998).

1. Each of the authors discusses something they call "internal replicability analysis." The term is apparently due to Thompson (1994), and it represents a misinterpretation of the work on the jackknife and the bootstrap in the statistical literature. I challenge all of the authors to find in that literature (e.g., Diaconis & Efron, 1983; Efron & Gong, 1983; Mooney & Duval, 1993; Mosteller & Tukey, 1977) any reference to either approach providing evidence for the replicability of a finding. They are simply procedures for estimating sampling error without making the traditional parametric assumptions. The confusion may arise from the fact that both require the creation of several replications of the statistic of principal interest (the jackknife by "re-sampling" the sample data without replacement; the bootstrap by "re-sampling" the data with replacement).

2. None of the authors cite the article by Abelson (1997), and two of the authors (McLean and Ernest (1998) and Nix and Barnette (1998)) do not even cite the book on the significance testing controversy (Harlow, Mulaik, & Steiger, 1997) in which that article appears. It is the best defense of the use of significance tests I have ever read. Since the controversy has been going on for many years it is impossible to cite every relevant source, but McLean and Ernest (1998) don't even cite Schmidt (1996), the most vocal critic of significance tests and strongest advocate of meta-analysis. Daniel (1998) cites Thompson's (1998) review of the Harlow et al. compendium, but does not cite Levin's (1998) review that appeared in the same source.

3. Two of the authors make mistakes when discussing what a null hypothesis is. Daniel (1998) gives an example where the null hypothesis is said to be: *r* (the sample *r*) is equal to zero, and claims that "by definition" a test of significance tests the probability that a null hypothesis is true (the latter is OK in Bayesian analysis but not in classical inference). Both Daniel (1998) and Nix and Barnette (1998) refer to the null hypothesis as the hypothesis of no relationship or no difference; no, it is the hypothesis that is tested, and it need not have zero in it anyplace.

4. None of the authors point out the weaknesses of confidence intervals or how they can be misinterpreted just as seriously as significance tests. For example, it is not uncommon to see statements such as "the probability is .95 that the population correlation is between a and b." A population correlation doesn't have a probability and it is not "between" anything; it is a fixed, usually unknown, parameter that may be bracketed or covered by a particular confidence interval, but it doesn't vary.

Thomas R. Knapp is a professor of nursing and education at The Ohio State University. Correspondence regarding this article should be addressed to Thomas R. Knapp, College of Nursing, The Ohio State University, 1585 Neil Avenue, Columbus, OH 43210-1289 or send e-mail to knapp.5@osu.edu.

5. None of the authors make sufficiently explicit the necessary distinction between a hypothesized effect size and an obtained effect size. It is the former that is relevant in determining an appropriate sample size; it is the latter that provides an indication of the "practical significance" of a sample result and around which a confidence interval can be constructed. Cohen (1988) at least tried to differentiate the two when he put the subscript s on the d for the obtained effect size. Some of the confusion in the significance testing controversy could be avoided if we had different terms for those two kinds of "effect sizes." (A similar confusion has arisen recently regarding prospective and retrospective power – see Zumbo & Hubley, 1998.)

6. Daniel (1998) claims that a df of 300 for an ANOVA error term is five times more likely to produce a statistically significant difference than a df of 60. That's not true; the relationship between p and F is not linear.

7. McLean and Ernest (1998) claim that Cohen (1988) recommended a power of .80 as arbitrarily as Fisher recommended an alpha of .05. That's not fair. He (Cohen) argued there, and elsewhere, that Type I errors are generally more serious than Type II errors and therefore beta (= 1 - power) can be chosen to be considerably larger than alpha.

8. Nix and Barnette (1998) refer to "the reliability of the study." There is no such thing as the reliability of a study. Measuring instruments have varying degrees of reliability (I think the claim by Daniel (1998), and others, that reliability pertains to scores, not instruments, is much ado about nothing); statistics have varying degrees of reliability, in the sense of sampling error; studies do not.

9. Nix and Barnette (1998) also seem to suggest that inferential statistics in general and significance testing in particular are primarily relevant for experiments (given their several references to "treatments"). Statistical inference actually gets very complicated for experiments, since it is not clear what the population(s) of interest is (are). Experiments are almost never carried out on random samples, but all true experiments have random assignment. What inference is being made (from what to what) is a matter of no small confusion. (See the reaction by Levin, 1993 to Shaver, 1993 regarding this issue.)

10. Daniel (1998) advocates, as does Thompson, "what if" analyses (not to be confused with the "What if . . . ?" title of the Harlow book). Although such analyses are not wrong, they are unlikely to be very useful. Researchers have actual sample sizes and actual values for their statistics; speculating as to what might have happened if they had bigger or smaller sample sizes, or the population correlations had been bigger or smaller, or whatever, is the sort of thinking that should be gone through before a study is carried out, not after. (See Darlington, 1990, pp. 379-380 regarding this matter.)

But to end on a positive note, I commend Daniel (1998) for his point that a significance test tells you nothing about the representativeness of a sample; McLean and Ernest (1998) for their contention that significance tests (*and* confidence intervals?) aren't very important for huge sample sizes; and Nix and Barnette (1998) for bringing to the attention of the readers of this journal that there are both significance tests and confidence intervals available for multivariate analyses. Curiously, most of the controversy about significance testing has been confined to univariate and bivariate contexts.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would have to be invented). In L.L Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

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

Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *RESEARCH IN THE SCHOOLS*, *5*(2), 23-32.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, *248*, 116-130.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36-48.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.

Levin, J. R. (1998). To test or not to test H_O. *Educational and Psychological Measurement*, *58*, 311-331.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *RESEARCH IN THE SCHOOLS*, *5*(2), 15-22.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Mosteller, F., &: Tukey, J.W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A Review of null hypothesis significance testing. *RESEARCH IN THE SCHOOLS*, *5*(2), 3-14.

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

Shaver, J. P. (1993) What statistical significance testing is, and what it is not. *Journal of Experimental Education*, *61*, 293-316.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality*, *62*, 157-176.

Thompson, B. (1998). Review of *What if there were no significance tests?* by L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.). *Educational and Psychological Measurement*, *58*, 332-344.

Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, *47*, 385-388.

RESEARCH IN THE

1998, Vol. 5, No. 2, 43-53

What If There Were No More Bickering About Statistical Significance Tests?

Joel R. Levin
University of Wisconsin – Madison

Questions and concerns are directed to those who advocate replacing statistical hypothesis testing with alternative data-analysis strategies. It is further suggested that: (1) commonly recommended hypothesis-testing alternatives are anything but perfect, especially when allowed to stand alone without an accompanying inferential filtering device; (2) various hypothesis-testing modifications can be implemented to make the hypothesis-testing process and its associated conclusions more credible; and (3) hypothesis testing, when implemented intelligently, adds importantly to the story-telling function of a published empirical research investigation.

From the local pubs to our professional "pubs," everyone in social-science academic circles seems to be talking about it these days. Not that there's anything wrong with talking about it, mind you, even to a more practically oriented crowd such as the readership of this journal. But as with the "gates" of Washington politics on the one coast and the Gates of Washington state on the other, when do we stand up and say "Enough already!"? When do we decide that ample arguments have been uttered and sufficient ink spilled for us to stop talking about it and instead start doing something about it?

The "it," of course, is the "significance test contro-versy" (Morrison & Henkel, 1970), which, in its most extreme form is whether or not conductors/reporters of scholarly research should continue (or even be *allowed* to continue) the time-honored tradition of testing statistical hypotheses. As has been carefully documented in our current forum on the issue in this issue of *RESEARCH IN THE SCHOOLS*, the topic isn't one that just recently arrived on the science scene. Not at all. Eminent statisticians, applied researchers, and just plain folks have been debating the virtues and vices of statistical significance testing for decades, with the debate crescendoing every couple of decades or so – consistent with principles of GC ("generational correctness").

The decade of the 1990s has been a critical one in hypothesis testing's protracted struggle for survival. During this decade especially vitriolic attacks, by

Joel R. Levin is a professor of educational psychology at the University of Wisconsin. Correspondence concerning this article should be addressed to Joel R. Levin, Department of Educational Psychology, 1025 W. Johnson St., University of Wisconsin, Madison, WI 53706 (E-mail address: LEVIN@MACC.WISC.EDU).

especially viable attackers, in especially visible outlets (e.g., Cohen, 1990, 1994; Kirk, 1996; Schmidt, 1996), have been mounted for the greater good of God, country, and no more significance testing! Even more critically for the life-and-death struggle, in the 1990's we also witnessed the first formal establishment of task forces and committees representing professional organizations [e.g., the American Psychological Association (APA), the American Educational Research Association (AERA), the American Psychological Society (APS)] to study the "problem" and make recommendations. As the deliber-ations of such task forces have proceeded apace, so have the spoken and written words: for example, in semi-civilized debates at professional meetings [e.g., "Significance tests: Should they be banned from APA journals?" (APA, 1996); "Should significance tests be banned?" (APS, 1996); "A no-holds-barred, tag-team debate over the statistical significance testing contro-versy" (AERA, 1998)] and in the most comprehensive, most indispensable, source on the topic, the edited volume *What if there were no significance tests*? (Harlow, Mulaik, & Steiger, 1997; reviewed by Levin, 1998, and Thompson, 1998).[1]

In the typical argument scenario, hypothesis testing is cast as the "bad guy," the impeder of all scientific progress. The prosecution prosecutes the accused, and then the defense defends. That is the basic approach taken in Harlow et al.'s (1997) four focal chapters ("The Debate: Against and For Significance Testing"), as well as in the recent professional meeting set-to's. As each piece of hypothesis-testing evidence is trotted out for public display, the typical juror-consumer goes through a "good point, that sounds reasonable, I hadn't thought of that" self-dialogue before deciding whether to convict or acquit, or just to quit and retreat to his/her original position on the subject.

Comments and Questions Related to the Present Articles

A similar structure and sequence of events are witnessed in the present collection of three essays. The "bad guy, good guy" script is closely followed, with each essay providing informative backgrounding, coherent evidence, and a convincing closing argument in the form of practical suggestions and proposed solutions. At the same time, even though the Editors of *RESEARCH IN THE SCHOOLS* have striven to be impartial and maintain a balance of perspectives here, the fact that two of the essays are clearly hypothesis-testing indictments whereas only one supports the process indicates that the present scales of justice were tipped a priori toward conviction. Given this unfair state of affairs and not knowing in advance the substance of the other critics' critiques, I can be "up front" in my admission of evening out the imbalance with the comments I am about to make.[2]

All the authors of the present articles cite relevant literature in a scholarly fashion and then proceed to make their case. As a reminder of what those cases are: (a) Nix and Barnette (1998) nix hypothesis testing in favor of a number of more thought-to-be informative alternatives to it (the provision of effect sizes, confidence intervals, replication, meta-analyses); (b) Daniel (1998) basically concurs and then goes on to recommend specific journal editorial-policy measures that could be implemented to effect those changes; and (c) McLean and Ernest (1998) disagree with the fundamental assertion about hypothesis testing's inutility, arguing essentially that it has an important "time and place" (Mulaik, Raju, & Harshman, 1997) in the scientist's analytic arsenal.

Although I have found it unwise to argue with people on matters of politics, religion, and their convictions about hypothesis testing, I will nonetheless attempt to do so by commenting on selected specifics in the three focal articles, in no particular order. Included in my comments are a number of questions that the articles evoked, the responses to which I look forward to reading in the authors' rejoinders. With the exception of Nix and Barnette's discussion of "research registries" (which I found to be a useful notion that should be given serious consideration by social scientists), the case against hypothesis testing introduces all the usual suspects. In that the present authors have examined these suspects in a generally commendable fashion, I will do my best to cross-examine them. In addition to being invited to serve as a commentator on these articles, I was encouraged to get in my own "two bits worth." And so I shall, beginning with a confession: Because of my previously professed "pro" position in the hypothesis-testing debates, I apolo-gize in advance for disproportionately carping and sniping more at the "con" positions of Nix-Barnette and Daniel.

Hypothesis-Testing Fever/Furor

Considerable issue can be taken with something that Nix and Barnette claim early on, namely, that "the informed stakeholders in the social sciences seem to be abandoning NHST . . ." As one who considers himself to be an informed stockholder, I'd be curious to learn to whom Nix and Barnette are referring, on what survey or other supporting reference their claim is made, and exactly how prevalent this abandonment is. One has to wonder: If the perniciousness of hypothesis testing is so pervasive, then why has APA's elite task force recommended that the practice *not* be abandoned, but rather supplemented and improved by many of the same enhancements that are mentioned in the present exchange (*viz*., effect magnitude measures, confidence intervals, replications, and meta-analysis)?

It is understandable that much, if not most, of what Daniel decries and prescribes has been decried and prescribed before. It is understandable because: (a) Daniel draws heavily from the words and work of Bruce Thompson (11 references and counting); and (b) Daniel, as a Thompson collaborator (Thompson & Daniel, 1996a, 1996b), is undoubtedly quite familiar with that corpus. Prominent in Daniel's list of hypothesis-testing do's and don'ts are Thompson's (e.g., 1996) "big three" recommended editorial policy "requirements" for authors of empirical studies – namely, that authors must always: (a) modify the word "significant" with "statistically," in reference to hypothesis tests; (b) include explicit effect-size information; and (c) provide some form of outcome "replicability" evidence.

"Significance" Testiness

Such proposed editorial policy changes are sensible enough and I clearly support the spirit – though not the letter – of them (e.g., Levin & Robinson, in press; Robinson & Levin, 1997). What is difficult to support are requirements that take away certain freedoms of author style and expression; in particular, when editorial policy is only half a vowel away from turning into editorial police. For example, when addressing a professional audience with a shared understanding of technical terminology, why should an author be forced into using stilted, reader-unfriendly, language (e.g., "The two correlations are each statistically significant but not statistically significantly different from one another.")? In a Results section where statistical hypotheses are being tested, there can be no misunderstanding what the word "significant" does or does not mean; the context disambiguates the concept. On the other hand, if an author who detects an effect that is significant statistically (e.g., a significance probability of $p = .01$) but insignificant practically (e.g., a standardized difference in means represented by a Cohen's d of .01) goes on to talk about the effect with reckless hyperbole, then, yes, that author should be shot at sunrise – or at least appropriately chastised.[3]

Effect-Size Defects?

Speaking of talking, the just-mentioned confusion represents a profound mismatch between an author's evi-dence and his/her words, stemming from a preoccupation with statistical significance at the expense of taking into account the magnitude of the obtained effect (which in the $d = .01$ case was minuscule). However, I have problems with the other side of the "nouveau" editorial-policy- recommendations coin regarding effect-size reporting as well. I will mention a few such problems, none of which is noted either by Daniel or by Nix and Barnette.

First, and even though I am all for including effect sizes as ancillary evidence of outcome importance, it has been pointed out previously (Levin & Robinson, in press; Robinson & Levin, 1997) that there are extremists in the mandatory effect-size camp (including journal reviewers and editors) who advocate reporting and concentrating on effect sizes *only* (i.e., without accompanying statistical/probabilistic support). This practice is absurdly pseudoscientific and opens the door to encouraging researchers to make something of an outcome that may be nothing more than a "fluke," a chance occurrence. Without an operationally replicable screening device such as statistical hypothesis testing, there is no way of separating the wheat (statistically "real" relationships or effects) from the chaff (statistically "chance" ones), where "real" and "chance" are anchored in reference to either conventional or researcher-established risks or "confidence levels." McLean and Ernest's description of Suen's (1992) "overbearing guest" analogy is especially apt in this context.[4]

Examples of the seductive power of large observed effect sizes that are more than likely the result of chance outcomes are provided by Levin (1993) and Robinson and Levin (1997). In its extreme form, effect-size-only reporting degenerates to strong conclusions about differential treatment efficacy that are based on comparing a single score of one participant in one treatment condition with that of another participant in a different condition. Or, even more conveniently and economically (i.e., in situations where time and money are limited), how about conclusions from a "what if" meta-experiment in which scores of two *imaginary* participants are compared ($N = 0$ studies)? The latter tongue-in-cheek situation aside, consider the following proposition:

> Suppose that Aladdin's genie (Robin Williams?!) pops out of the lamp to grant you only *one* forced-choice wish in relation to summarized reports of empirical research that you will read for the rest of

> your lifetime: You can have access to either a statistical-significance indicator of the reported findings or a practical-significance index of them, but not both (and no sample-size information can be divulged). Which would you choose?

Personally speaking, it would be painful to have to choose only one of these mutually exclusive alternatives. Based on the aforementioned "chance" and "seductive effect size" arguments, however, I think that a strong case can be made for statistical over practical significance. McLean and Ernest's chance-importance-replicability trichotomy represents a nice way of thinking about the problem, with an assessment of the findings' nonchanceness and replicability each given priority over importance. At the same time, l heartily endorse Nix and Barnette's statement, "We would like to see a situation where all studies that were adequately designed, controlled and measured would be reported, regardless of statistical significance." In fact, I am quite sympathetic with others who have called for manuscript reviews and editorial decisions based on just a study's rationale, literature review, and methods and procedures, in the form of a research proposal – with the associated outcomes and data analyses not included until an editorial decision has been reached (e.g., Kupfersmid, 1988; Levin, 1997; Walster & Cleary, 1970a).

So you want to change the world? Nix and Barnette, as well as Daniel, make it sound as though the research world will be a far better place when the hypothesis- testing devil is ousted by the effect-size angel. In my opinion, that is not a fair assumption, as effect-size calculating and reporting are subject to the same "bias" criticisms inherent in familiar "how to lie with statistics" treatises. How to lie with effect sizes? Levin and Robinson (in press) have noted how researchers can select from any number of conventional effect-size measures (including both more and less conservative variants of the indices listed in Nix and Barnette's Table 1, among others) to make the preferred case for their own data. Another problem associated with relying on commonly calculated effect sizes alone is illustrated in the following hypothetical example.

Suppose that an investigator wants to help older adults remember an ordered set of ten important daily tasks that must be performed (insert and turn on a hearing aid, take certain pills, make a telephone call to a caregiver, etc.). In a sample of six elderly adults, three are randomly assigned to each of two experimental conditions. In one condition (A), no special task instruction is given; and in the other (B_1), participants are instructed in the use of self-monitoring strategies. Following training, the participants are observed with respect to their success in performing the ten tasks. As can be seen in the first two columns of Table 1, the average number of tasks the participants correctly remembered to perform was 1.33 and 3.33 for the no-instruction (A) and self-monitoring (B_1) conditions, respectively. For the data provided in Table 1, it can be determined that the "conditions" factor accounts for a hefty 82% of the total variation in task performance (i.e., the squared point-biserial correlation is .82, which for the two-sample case, is equivalent to the sample ς^2). Alternatively, the self-monitoring mean is 3-½ within-group standard deviations higher than the no-instruction mean (i.e., Cohen's *d* is 3.5). From either effect-size perspective (ς^2 or *d*), certainly this represents an impressive treatment effect, doesn't it? Or does it?

Table 1
Hypothetical Data Illustrating Equivalent Standardized Effect Sizes (Condition B Versus Condition A) With Vastly Different Practical Implications

	Condition A	Condition B_1	Condition B_2
	1	3	5
	1	3	8
	2	4	10
M	1.333	3.333	7.667
SD	.577	.577	2.517

Suppose that instead of self-monitoring training, participants were taught how to employ "mnemonic" (systematic memory-enhancing) techniques (B_2) – see, for example, Carney & Levin (1998) – with the results as indicated in the third column of Table 1. The corre-sponding B_2 mean is 7.67 correctly remembered tasks and a comparison with no-instruction Condition A surprisingly reveals that once again, the conditions factor accounts for 82% of the total variation in task performance (equivalently, *d* again equals 3.5).[5] Thus, when expressed in standardized/relative terms (either ς^2 or *d*), the effect sizes associated with the two instructional conditions (B_1 and B_2) are exactly the same, and substantial in magnitude. Yet, when expressed in absolute terms and with respect to the task's maximum, there are important differences in the "effects" of B_1 and B_2: Increasing participants' average performance from 1.33 to 3.33 tasks remembered seems much less impressive than does increasing it from 1.33 to 7.67. Helping these adults remember an average of only 3 of their 10 critical tasks might be regarded as a dismal failure, whereas helping them remember an average of almost 8 out of 10 tasks would be a stunning accomplishment. Yet, the conventional effect-size measures are the same in each case.[6]

How, then, not to lie with effect sizes? To borrow from Cuba Gooding, Jr.'s character in the film, *Jerry Maguire*: Show me the data! Show me, the reader, "sufficient" data (American

Psychological Association, 1994, p. 16) either in raw (preferably) or in summary form. Then, let me, the reader, decide for myself whether a researcher's particular finding is educationally "significant" or "important," with respect to the standards that I regard as "significant" or "important" (see also Prentice & Miller, 1992).

Lack-of-confidence intervals. Briefly noted here are other suggested alternatives to hypothesis testing that are briefly noted by Daniel, as well as by Nix and Barnette. These include the inclusion of confidence intervals and meta-analyses, both of which are signature recom-mendations of Schmidt and Hunter (e.g., 1997). As far as the former are concerned, it is well known that one can simply slap a standard error and degree of confidence on an effect size and build a confidence interval *that is equivalent to testing a statistical hypothesis* (but see McGrath, 1998). Schmidt, Hunter, and their disciples, however, eschew that particular application and instead encourage researchers to select two or three or four or five degrees of confidence (e.g., 99%, 95%, 90%, 80%, 70%) and then build/display two or three or four or five corresponding confidence intervals. Well and good, but how is the researcher or reader to interpret these varying-degrees-of-confidence intervals, and what is one to conclude on the basis of them (e.g., when a 95% interval includes a zero treatment difference but a 90% interval does not)? How much confidence can one have in such subjective nonsense?

I never met a meta-analysis . . . Concerning meta-analyses: I have nothing against them. They can be extremely valuable literature-synthesis supplements, in fact. Yet, their purpose is surely quite different than that of an individual investigator reporting the results of an individual empirical study, especially when the number of related studies that have been previously conducted are few or none. Alas, what's a poor (graduate-student or otherwise) single-experiment researcher to do (Thompson, 1996)? Of course, if the logical corollary to the meta-analysis argument is that no single-experiment reports should be published in empirical journals as we currently know them, then count me in! I strongly endorse the recommendation that replications and multiple-experiment "packages" comprise an essential aspect of a researcher's LPU ("least publishable unit") – see, for example, Levin (1991, p. 6).

Robust Conclusions Versus Replicated Outcomes

There's something about "replication" in two of the present articles with which I take issue. That something is a restatement of Thompson's (1993) view that data-analysis strategies such as cross-validation, boot-strapping, and jackknifing "indicate the likelihood of replication" (Nix and Barnette) or "may provide an estimate of replicability" (Daniel). For readers not in the know and who might be misled by such semantic twists, allow me to elaborate briefly. A "replication" defined by corroborating analyses based on alternative slices or samples of the same data – which applications of the just-mentioned statistical procedures attempt to do (see, for example, Efron & Gong, 1983) – is nice for establishing *the robustness of a single study's conclusions* (Thompson's "internal" replication). However, that type of "replication" is neither as impressive nor as imperative for the accumulation of scientific knowledge as is a "replication" defined by *an independently conducted study* (i.e., a study conducted at different sites or times, with different specific participants and operations) *that yields outcomes highly similar to those of the original study* (Thompson's "external" replication) – see, for example, Neuliep (1993) and Stanovich (1998). Even to suggest that researchers should be satisfied with the former, by rationalizing about researchers' diminished physical or fiscal resources (as both Thompson and Nix and Barnette do), is not in the best interest of anyone or anything, and especially not in the best interest of educational research's credibility within the larger scientific community.

What if there were no more bickering about significance tests? Conclusion robustness itself is a matter of no small concern for researchers, for outcome "credibility" (Levin, 1994) and generalizability depend on it. Yet, because of the excessive "heat" (Thompson, in press) being generated by hypothesis-testing bickerers, little time is left for shedding "light" on how to enhance the conclusion robustness of educational and psychological research. In addition to the methodological adequacy of an empirical study (e.g., Levin, 1985; Levin & Levin, 1993; Stanovich, 1998), the credibility of its findings is a function of the study's "statistical conclusion validity" (Cook & Campbell, 1979), which in turn encompasses a consideration of the congruence between the statistical tools applied and their associated distributional assump-tions. Reviews of the literature indicate that precious little attention is being paid by researchers and journal referees alike to that congruence: Statistical tests are being mindlessly applied or approved even in situations where fundamental assumptions underlying them are likely grossly violated (e.g., Keselman et al., in press; Wilcox, 1997).[7] Bickering time spent on significance testing is also time away from considering other critical conclusion-robustness issues, including particularly those associated with the pervasive educational research realities of: nonindependent sampling, treatment, and testing units; random (as opposed to fixed) treatment factors; longitudinal and other multivariate designs, among others (e.g., Clark, 1973; Kratochwill & Levin, 1992; Levin, 1992a; Raudenbush & Bryk, 1988; Willett & Sayer, 1994). Accompanied or not by

significance testing per se, such statistical issues remain properly "significant."

That concludes my comments on the "big issues" addressed by the three focal articles in this issue of *RESEARCH IN THE SCHOOLS*. Before concluding with a few additional big issues of my own, I will address several misleading and erroneous statements that appear in the present articles. Though not of the magnitude of the issues just discussed, these statements are nonetheless sufficiently distressing that they should not go unmentioned.

Misleading and Erroneous Assertions in the Present Articles

It is bad enough when *consumers* of research reports are uninformed with respect to the methods and meanings of the data analyses reported (as has been claimed, for example, with respect to the hypothesis-testing term "significant"). Even worse is when *researchers/authors* are misinformed with respect to those methods or meanings. But worst of all is when *critics* of data-analytic practices dangerously mislead or make erroneous assertions regarding those practices – and particularly when the words "misuse and misinterpretation" are featured in the title of a critic's critique (as in Daniel's article, for example).

Sample size and statistical power. To wit, consider Daniel's comments about the components of an *F*-test of mean differences, which I quote [with numbers added for convenience in referencing]:

> . . . the mean square for the error term will get smaller as the sample size is increased [1] and will, in turn, serve as a smaller divisor for the mean square for the effect [2], yielding a larger value for the *F* statistic [3]. In the present example (a two-group, one-way ANOVA), a sample of 302 would be five times as likely to yield a statistically significant result as a sample of 62 simply due to a larger number of error degrees of freedom (300 versus 60) [4].

What a misrepresentation of the *F*-test and its operating characteristics! The error mean square (*MSE*) is an unbiased estimator of the population variance ($ó^2$) that is not systematically affected by sample size. What increasing sample size does is to reduce the sampling variability associated with each condition's mean, which results in increased variability *among* those means, which in turn increases the mean square between conditions (*MSB*) in the *F*-test's *numerator*. Propositions [1] and [2] are therefore false, which invalidates proposition [3]. Proposition [4] is not true as a result of the preceding illogic.

It is also false as a consequence of Daniel's stated "larger number of error degrees of freedom." Again, larger sample sizes increase statistical power by decreasing the sampling variability associated with each condition's mean, which operates to increase the variability among those means. None of this works auto-matically to increase the *F*-statistic by a constant amount, however, as is asserted by Daniel (e.g., "by five times"), *unless* it is also stated that *all else (except sample size) is held constant* – which includes the value of *MSE* and the means for each condition (all of which are statistics that will vary unsystematically with changes in sample size). To give the impression that merely increasing sample size *guarantees* a larger *F*-ratio, as Daniel and others imply, is unfortunate because it simply is not true.

Show you the data? Don't press the issue. I could come up with dozens – if not hundreds, thousands, or zillions, if I had the time and temperament – of examples from actual empirical studies, many from my own substantive research program, where an *F*-ratio based on small sample sizes (calculated, for example, early in the data-collection process) becomes *smaller* when based on larger or final sample sizes.

Some of Nix and Barnette's assertions about statistical power and a study's publishability are similarly misleading. First, the authors state that the problem is of special concern in educational research, where ". . . effect sizes may be subtle, but at the same time, may indicate meritorious improvements in instruction and other class-room methods." If instructional improvements are indeed "meritorious," then: (a) effect sizes will not be "subtle;" and (b) even with modest sample sizes, statistical signifi-cance will follow. Second, many readers are likely to be misled by the authors' statements that "reliability . . . can be controlled by reducing . . . sampling error" and "the most common way of increasing reliability . . . is to increase sample size." Reducing *sampling* error or increasing *sample* size (the number of participants) does not increase reliability. Reducing *measurement* error or increasing *test* size (the number of items) does. Increasing sample size increases the power or sensitivity of a statistical test, however.

Errors and effect sizes. Nix and Barnette also state that in a hypothesis-testing context, "errors can be due to treatment differences." This statement will come as news to many and deserves some elaboration. In the section entitled "Misunderstanding of *p* values," the authors caution that "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) . . ." How can a difference be simultaneously "trivial" and "large?" Read that sentence again. Later in the same section, the authors argue that researchers should "con-tinue to determine if the statistically significant result is due to sampling error or due to effect size." The imprecisely worded statement may lead an uninitiated reader to believe that it is actually possible for a researcher to make such a precise either-or determination, when it is not. In Nix and

Barnette's section, "Interpreting effect size," the impression is given that the various U measures are separate/unrelated, when in fact they are alternative ways of thinking about the same outcome – just as is converting *d* (a standardized difference in means) to *r* (the correlation between treatment and outcome), something that was left unsaid. Omitted from a subsequent paragraph is the caution that comparing single-study effect sizes with composite effect sizes can be grossly misleading unless all treatments in question are evaluated relative to functionally equivalent "control" groups (see also Levin, 1994).

Hypothesis Testing as a Meaningful, Memorable Process

In this section I will provide a few personal thoughts about statistical hypothesis testing and its rightful role in the analysis and reporting of empirical research in education and psychology.

Dump the Bathwater, Not the Baby...

No, statistical hypothesis testing, as is generally practiced, is not without sin. I too oppose mindless (e.g., Cohen's, 1994, "rare disease" scenario; Thompson's, 1997, "reliability/validity coefficient testing" criticism) and multiple (e.g., testing the statistical significance of all correlations in a 20 x 20 matrix) manifestations of it. Such manifestations surely portray the practice of hypoth-esis testing at its worst. More forethought and restraint on the part of researchers would likely help to deflect much of the criticism concerning its misapplication.

Absent in each of the present articles' proposed replacement therapies for traditional statistical hypothesis testing are *alternative hypothesis-testing therapies themselves* – which I have referred to generically as "intelligent" hypothesis-testing practices (Levin, 1995) and which have been articulated in a set of ideal principles (Levin, 1998). The overarching premise is that statistical hypothesis testing can be a valuable decision-making tool, if implemented in conjunction with a researcher's a priori (i.e., prior to data collection) planning, specification, or determination of:

- a select number of carefully developed (prefer-ably, theory-based) hypotheses or predictions
- a statistical test or tests that validly and parsimoniously assess those hypotheses
- Type I error probabilities that are adequately controlled
- magnitudes of effects that are regarded as substantively "important," along with their associated probabilities of detection
- magnitudes of effects that are regarded as substantively "trivial," along with their associated probabilities of nondetection
- sample sizes that directly follow from these specifications.

The more of these ingredients that are incorporated into the hypothesis-testing process, the more intelligent and informative is that process.

Effects that emerge as statistically significant as a result of intelligent hypothesis testing should be supple-mented by ancillary "practical significance" information, including effect sizes (based on relative and/or absolute metrics), confidence intervals, and even – heaven forbid! – more "qualitative" assessments of treatment efficacy (e.g., experimenter observations and participant self-reports). *The* most important supplement to this statistical basis for scientific hypothesis confirmation is evidence accumulation, initially through empirical replications (Levin's, 1995, "A replication is worth a thousandth *p*-value.") and ultimately through literature syntheses (which include the tools of meta-analysis).

In contrast to the anti-hypothesis-testing reforms in the graduate-level statistics courses taught at Michigan State (alluded to by Nix and Barnette), UW-Madison colleague Ron Serlin and I attempt to impart intelligent hypothesis-testing practices to our students. In addition to simply teaching and writing about the potential of such improvements to statistical hypothesis testing (e.g., Levin, 1985, 1997; Seaman & Serlin, in press; Serlin & Lapsley, 1993), we also attempt to practice these preachings in our substantive research investigations. For example, Ghatala and Levin (1976, Exp. 2) adapted Walster and Cleary's (1970b) procedure for determining "optimal" sample sizes to distinguish between substantively important and trivial effects based on acceptable Type I error control and statistical power. Similarly, I convinced a former student to incorporate components of "predicted pattern testing" (Levin & Neumann, in press) to provide stronger, more sensible, tests of his theoretically based predictions – see Neumann and DeSchepper (1991, Exp. 3).

To present a case for a place for intelligent statistical hypothesis testing in educational research, I invite you to imagine the following seemingly far-from-educational- research situation:

> Suppose that you are a medical doctor, whose life work is to keep people alive. A particular patient fits a profile for being "at risk" for developing some dangerous abnormality. You need to make a decision, based on a simple screening test, whether or not to proceed to more extensive/expensive testing. For patients with this kind of "at risk" profile, the screening test is known to have a 90% chance of identifying those who have the abnormality to some substantial degree, a 5% chance of identifying those who have the abnormality only to some very minimal

> degree, and a 1% chance of identifying those who do not have the abnormality at all.[8]

Based on the preceding information, does it seem reasonable to you, as a responsible doctor, to use the screening test as a basis for making a decision about whether or not to proceed to the next phase of evaluation? It does to me.

> OK, now suppose that you are an educational researcher whose life work is to study ways of improving the academic performance of "at risk" students. You have developed a literature- guided intervention for "at risk" middle-school students and you want to assess its effectiveness by comparing the end-of-year educational achievement of students who receive the intervention and those who do not (randomly determined). If the intervention produces a substantial difference in average achievement between the two groups (operationalized as d = 1.00), you want to have a 90% chance of detect-ing it; if it produces a minimal difference (d = .25), you only want a 5% chance of detecting it; and if there is no difference at all (d = .00), you are willing to tolerate a risk of 1% of falsely detecting that. Adapting the Walster-Cleary (1990b) approach, for example, indicates that the just-specified parameters and probabilities are satisfied if 32 students are randomly assigned to each of the two conditions (intervention and no intervention).

Based on the preceding information, does it seem reasonable to you, as a responsible educational researcher, to perform a statistical test as a basis for making a decision about the intervention's potential? It does to me – and especially because the situation just described incorporates the earlier listed intelligent hypothesis-testing ingredients. I certainly do not claim this hypothetical educational hypothesis-testing example to represent a detail-by-detail correspondence with the equally hypothetical medical screening-test example. Rather, it constitutes a close enough analogy that takes us through a similarly sensible decision-making process.

. . . And Now the Rest of the Story

I conclude my remarks with a story relevant to our discussion of hypothesis testing's proper place on the empirical research plate.

> It is a dark and stormy night. A shot rings out in the presidential palace. A body slumps and falls to the ground, dead. A one-armed man is seen fleeing the scene. Inspectors Poirot and Clouseau are called in to investigate. Poirot determines that the deceased is the president's lover. Clouseau notices a charred sheet of paper in the fireplace. He picks it up. "Ooooohh, it's still hot!" he yelps, but is nonetheless able to discern some scribblings on the paper. "Zoot, alors, I have it! And I know precisely how it happened!" Clouseau crows. He continues: "The murderer is . . . [pause] . . . the president's men . . . [pause] . . . or possibly it's the one-armed man . . . [pause] . . . or perhaps it's even the president herself . . . [pause] . . . I haven't a clew!"

Hey, c'mon, who dunnit? Tell us the rest of the story. Inquiring minds want to know!

So you want to know the ending? Let me tell you a different story. Somewhere along the academic trail I had an epiphany about reports of empirical research in scholarly journals (at least those in the fields of psychology and education): In addition to describing what was done, how it was done, and what was found, a journal article should "tell a story." I'm not using "story" in the fictional sense here, but rather as true to life and justifiable on the basis of the study's specific operations and outcomes. Telling a story, with a clever "hook" and memorable take-home message, represents a key land-mark on the publication highway (e.g., Kiewra, 1994; Levin, 1992b; Sternberg, 1996). It is something that editors usually demand, reviewers seek, and readers require. A study without a meaningful, memorable story is generally a study not worth reporting. In certain situations, and in light of my earlier comments, incorporating one or more additional experiments into a one-experiment study often helps to breathe life into an otherwise moribund article.

Exactly what does any of this have to do with our current hypothesis-testing discussion? I believe that an invaluable, though heretofore overlooked, function of statistical hypothesis testing (especially if implemented intelligently) is to assist an author in developing an empirical study's story line and take-home message. Just as with the preceding Clouseauian fantasy with its inconclusive conclusion (or its invent-your-own ending), an empirical research article without an evidence-based conclusion is not likely to satisfy either the reader's affective (interest, enjoyment) or cognitive (under-standing, memory) processes. We human animals seek to extract some form of order from the chaos in the world around us; we are all "meaning makers." As consumers of scientific research, we seek to do the same from the jumble of theory, methods, and results that are provided in a journal article. In my opinion, selective, planful statistical hypothesis testing can help one extract order from chaos, not just in the "chance-finding filtering" sense, but in the sense of cementing as firm a conclusion as can be made from the evidence presented until a critical replication-attempting study comes along. I additionally believe that hypothesis testing is much better suited to that cementing task than are other

proposed individual alternatives for summarizing the results of single-study investigations, including the provision of effect sizes (are they real?) and multiple-confidence-level confidence intervals (which one do you prefer?).[9]

I could go on about the story-telling function of journal articles and hypothesis testing, but I think you get the idea. As for stories, what's the take-home message of *this* article? There are actually three take-home messages, each enumerated in the Abstract. If you're interested, go back and (re)read them. That, of course, is what journal abstracts are supposed to summarily convey: the "bottom line" of one's work.

References

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

Carney, R. N., & Levin, J. R. (1998). Mnemonic strategies for adult learners. In M. C. Smith & T. Pourchot (Eds.), *Adult learning and development: Perspectives from educational psychology* (pp. 159-175). Mahwah, NJ: Erlbaum.

Clark, H. H. (1973). The language-as-fixed-effects fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

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

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally.

Daniel, L.G. (1998). Statistical significance testing: A Historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *RESEARCH IN THE SCHOOLS*, 5(2), 23-32.

Derry, S., Levin, J. R., & Schauble, L. (1995). Stimulating statistical thinking through situated simulations. *Teaching of Psychology, 22*, 51-57.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician, 37*, 36-48.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23*, 89-105.

Frick, R. W. (1995). *Using statistics: Prescription versus practice*. Unpublished manuscript, Department of Psychology, State University of New York at Stony Brook.

Ghatala, E. S., & Levin, J. R. (1976). Phenomenal background frequency and the concreteness/imagery effect in verbal discrimination learning. *Memory & Cognition, 4*, 302-306.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), *Review of Research in Education* (Vol. 5, pp. 351-379). Itasca, IL: Peacock.

Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (in press). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*.

Kiewra, K. A. (1994). A slice of advice. *Educational Researcher, 23*(3), 31-33.

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

Kratochwill, T. R., & Levin, J. R. (Eds.). (1992). *Single-case research design and analysis: New developments for psychology and education*. Hillsdale, NJ: Erlbaum.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*, 635-642.

Levin, J. R. (1985). Some methodological and statistical "bugs" in research on children's learning. In M. Pressley & C. J. Brainerd (Eds.), *Cognitive learning and memory in children* (pp. 205-233). New York: Springer-Verlag.

Levin, J. R. (1991). Editorial. *Journal of Educational Psychology, 83*, 5-7.

Levin, J. R. (1992a). On research in classrooms. *Mid-Western Educational Researcher, 5*, 2-6, 16.

Levin, J. R. (1992b). Tips for publishing and professional writing. *Mid-Western Educational Researcher, 5*, 12-14.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*, 378-382.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review, 6*, 231-243.

Levin, J. R. (1995, April). *The consultant's manual of researchers' common stat-illogical disorders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Levin, J. R. (1997). Overcoming feelings of power-lessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging, 12*, 84-106.

Levin, J. R. (1998). To test or not to test H_0? *Educational and Psychological Measurement, 58*, 313-333.

Levin, J. R., & Levin, M. E. (1993). Methodological problems in research on academic retention programs for at-risk minority college students. *Journal of College Student Development, 34*, 118-124.

Levin, J. R., & Neumann, E. (in press). Testing for predicted patterns: When interest in the whole is greater than in some of its parts. *Psychological Methods.*

Levin, J. R., & Robinson, D. H. (in press). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist, 53*, 796-797.

McLean, J. E., & Ernest, J. M. (1998). The Role of statistical significance testing in educational research. *Research in the Schools, 5*(2), 15-22.

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

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

Neuliep, J. W. (Ed.). (1993). Replication research in the social sciences. Special issue of the *Journal of Social Behavior and Personality, 8*(6).

Neumann, E., & DeSchepper, B. G. (1991). Costs and benefits of target activation and distractor inhibition in selective attention. *Journal of Experimental Psychology: Learning, Memory & Cognition, 17*, 1136-1145.

Nix, T.W., & Barnette, J.J. (1998). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. *RESEARCH IN THE SCHOOLS, 5*(2), 3-14.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin, 92*, 766-777.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160-164.

Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, p. 423-475). Washington, DC: American Educational Research Association.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*, 21-26.

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

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of signifi-cance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. A. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Seaman, M. A., & Serlin, R. C. (in press). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods.*

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

Stanovich, K. E. (1998). *How to think straight about psychology* (5th ed.). New York: Longman.

Sternberg, R. J. (1996). *The psychologist's companion: A guide to scientific writing for students and researchers* (3rd ed.). Cambridge, UK: Cambridge University Press.

Suen, H. K. (1992). Significance testing: Necessary, but insufficient. *Topics in Early Childhood Special Education, 12*, 66-81.

Thompson, B. (1993). The use of statistical signifi-cance tests in research: Bootstrap and other alter-natives. *The Journal of Experimental Education*, 61(4), 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1997). Editorial policies regarding significance tests: Further comments. *Educational Researcher, 26*, 29-32.

Thompson, B. (1998). Review of *What if there were no significance tests*? *Educational and Psychological Measurement, 56*, 334-346.

Thompson, B. (in press). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*.

Thompson, B., & Daniel, L. G. (1996a). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

Thompson, B., & Daniel, L. G. (1996b). Seminal readings on reliability and validity: A "hit parade" bibliography. *Educational and Psychological Measurement, 56*, 741-745.

Walster, G. W., & Cleary, T. A. (1970a). A proposal for a new editorial policy in the social sciences. *The American Statistician, 24*, 16-19.

Walster, G. W., & Cleary, T. A. (1970b). Statistical significance as a decision-making rule. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology* (pp. 246-254). San Francisco: Jossey-Bass.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*, 363-381.

Footnotes

[1] The authors of the present exchange can certainly be excused for their limited reference to the Harlow et al. volume, as it likely was released only after earlier versions of the current articles had been written and submitted.

[2] Psst! It should be a secret to nobody that I am a staunch hypothesis-testing defender (e.g., Levin, 1993, 1998; Robinson & Levin, 1997) – although I do not defend the form in which it is generally practiced. That predilection obviously colors my reactions to the present articles.

[3] As an aside and as not accurately conveyed by McLean and Ernest, we (Levin & Robinson, in press; Robinson & Levin, 1997) do not argue *that* alternative language is needed in Results sections. Rather, we suggest that *if* better language is mandated, then descriptors such as "statistically real," "statistically nonchance," and "statistically different" could readily say what one means and mean what one says without a trace of "significance." [4] A primary function of statistical hypothesis testing has been analogized in even more colorful terms – a "crap detector" – by a distinguished scholar who shall unfortunately remain nameless in that I cannot locate the appropriate citation at the moment.

[5] In each case, the obtained treatment difference is statistically "real," or nonchance ($p \# .05$, one-tailed), on the basis of either a parametric or nonparametric hypothesis test.

[6] The major problem in this example arises from the conditions' differing variabilities. That problem could be accounted for by defining alternative *d*-like effect-size measures based on just the control condition's (Condition A's) standard deviation, as has been suggested by Glass (1977), Hedges and Olkin (1985), and others. Interpreting effect sizes, in the absence of raw data, remains a problem for $ç^2$ and Cohen's *d*, however. Concerns about effect sizes based on relative metrics, and a variety of other concerns, are detailed by O'Grady (1982), Frick (1995), and Fern and Monroe (1996).

[7] Note that assumptions violations also affect the validity of other inferential statistical alternatives, such as confidence intervals and meta-analyses. Interestingly and in contrast to the "replication" objectives misattributed to them, bootstrapping and jackknifing are methods that *do* possess either "distribution-free" or other robust qualities that could be exploited to circumvent assumption- violations problems.

[8] In this example, I have tried to mitigate the important "base-rate" problem (e.g., Derry, Levin, & Schauble, 1995) by restricting the population to patients with an "at risk" profile. Even so, the problem remains and would need to be taken into account should the screening test's results prove positive.

[9] On the other hand, if it can be documented that the major impediment to scientific progress lies in the value-lessness of reporting single- or few-study investigations (as some have accused), then why not simply discontinue the production of journals that publish primary-research articles and continue with only those that publish research syntheses? Imagine what a triumph that would be for meta-analysis enthusiasts!

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 55-57

A Review of Hypothesis Testing Revisited: Rejoinder to Thompson, Knapp, and Levin

Thomas W. Nix
University of Alabama

J. Jackson Barnette
University of Iowa

This rejoinder seeks to clarify the authors' position on NHST, to advocate the routine use of effect size, and to encourage reporting results in simple terms. It is concluded that the time for action, such as that advocated in Nix and Barnette's original article, is overdue.

Before we respond to the critiques of our colleagues, we would like to comment that discourse such as that exemplified in this journal issue is the type of debate that is necessary to lead us to more coherent methods of analyzing data. As Mark Twain said, "Loyalty to petrified opinion never yet broke a chain or freed a human soul." The situation we have described (Nix & Barnette,1998) is one that has the potential to mislead those not well versed in statistical methods, the enlightened practitioners who look to educational research for guidance in the most difficult and, in our opinion, the most important of professions, the education of fertile young minds.

Clarification of Our Position

First, we must clarify our position that has been somewhat distorted by the reviews. We do not agree with Schmidt (1996) that Sir Ronald Fisher led us to this point of confusion and chaos in the educational research endeavor (Sroufe, 1997). Fisher deserves praise for bringing to agronomy the methods that have helped agriculture achieve the productivity that we see today. However, Fisher and Pearson allowed their insecurities to seep into their professional lives. Instead of criticizing these great men, we should learn from their human frailties and not allow ourselves to repeat their mistakes.

We do agree with Schmidt that the advancement of knowledge, particularly in educational research, has been stymied by rote adherence to null hypothesis significance testing (NHST). The extensive literature outlining the shortcomings of NHST cannot be ignored; we must look to new methods that will bring more coherence to our field. Our position is not that a draconian ban on NHST should be imposed on the huddled scholarly masses. We agree with Thompson (1998) that NHST's are "largely irrelevant" (p. 5). This is why we have offered alternatives such as effect size measure, confidence intervals, measures of study replicability, meta-analytic studies, and research registries of studies, along with strategies for how we could move in an orderly fashion away from NHST without imposing bans or unnecessary rules.

We do believe that universal standards for social scientific endeavor are in the best interest of advancing knowledge. These standards, after thoughtful study, should apply to scholarly journal submissions, to human use institutional review boards, and to the conduct of meta-analytic studies. Standards, however, should not prohibit the use of any statistical technique. Bans of sacred cows usually only solidify the opposition to rational change. It is our belief that rational change can happen from the top-down through concerted action by the large professional organizations (the APA, AERA, ASA, etc.). What we advocate is not radical change, since models exist in the medical field and in Europe that simulate the actions that we have suggested. The only requirement is action.

Effect Size

Levin (1998) and Knapp (1998) have reported on our enchantment with effect size measures and the methods we advocate. In no way do we mean to imply that these methods are perfect, only better than the existing methods. Cohen (1988) has expressed some of the difficulty in explaining effect size in the multivariate case. Cohen has stated that, ". . . f^2 (the multivariate effect size index) is neither simple nor familiar . . ." (p. 477). Cooper and Hedges (1994) have reported that the early meta-analytic work was "at best an art, at worst a form of yellow journalism" (p.7). All methods have to go through a period of development and expansion. We

believe our recommendations would have been foolhardy in the 1970s or 1980s, since the methods we advocate had not gone through rigorous testing. At this point, we believe the period of development is far enough along to advocate the routine use of these methods as a means of advancing social science. In fact, we see further need for empirical research on the relationships among several indicators of treatment influence, including test statistics, *p*-values, confidence intervals, and with effect size measures including eta-squared and omega squared. I (JJB) am particularly interested in how these measures are related and how they are influenced by research design, number of groups and the number of subjects. Yes, effect size and meta-analytic techniques do have their limitations, and we should always remain vigilant to their shortcomings, just as some of our predecessors have with NHST.

Reporting of Results

We do agree with Levin (1998) that writing skill is a necessary prerequisite to good scholarship, but we do not agree that the ability to turn a clever phrase and tell a story should necessarily be part of good writing skill. We would like to see researchers, regardless of their inherent creativity, be able to report valid research results in the simplest terms possible. In this manner not only could researchers understand and appreciate the literature, but practitioners could also glean information from studies that could help in their everyday practice.

A prerequisite to good scholarship and good science is consistency in language. In the world of statistics, this is not a small problem. Vogt (1993) has attempted to explain some of the problems in definitions and vagueness of terms. For example, the symbol β is used to symbolize both the regression coefficient and the probability of a type II error in NHST. Similarly, the intercept and slope in a regression equation are often referred to as constants, when in fact, both have variance and standard errors asso-ciated with them. Additionally, researchers often fail to tell readers if the assumptions of a statistical test have been satisfied, let alone even tested, when the lack of adherence to the assumptions confounds the results of many tests. Statisticians understand these problems, but if only statisticians understand research, is the research of any value? For research to be valuable it must be precise and as unambiguous as possible so that is can be compre-hended by practitioners as well as other researchers. In this light, as opposed to Levin's preference for statistical significance, we would opt for the practical significance of research over statistical significance.

Apologies and Defense

We must now apologize to Knapp (1998) for our lack of clarity in using the term "reliability" to describe a study (p. 40). We stand corrected on this point. We should have used the term "replicability." However, it should be pointed out that in meta-analytic studies the individual study is a data point. Therefore, in this sense a study could be said to have reliability, if it can be replicated.

Knapp (1998) has also stated that the null hypothesis "need not have zero in it anyplace" (p. 39). In fact the use of H_o: $\mu_1 = \mu_2$ implies that there is no difference in the two population means, or H_o: $\mu_1 - \mu_2 = 0$. As other writers (Bakan, 1966; Cohen, 1988; Hinkle, Wiersma, & Jurs, 1994) have claimed, the null hypothesis is the hypothesis of no difference or no relationship. Of course, it is the hypothesis that is tested, but to say the null hypothesis need not have a zero in it is puzzling.

We agree with Knapp that we used Thompson's (1989, 1993) work as the basis for our recommendation that jackknife and bootstrap methods be used to test (within the limitations of the original data) the replicability of a study without full-scale replication. We also suggested that power of the test could be used as a surrogate for replicability (p. 10) We will leave Thompson (1998) and Knapp (1998) to resolve their disagreement, but conceptually we still believe, no matter what method or indicator is used, that the likelihood of the replicability of a study is important information for the reader and is in the best interest of good science.

Knapp (1998) indicated that we did not reference the outstanding work on the significance testing controversy by Harlow, Mulaik, and Steiger (1997). This is not correct. We reference three chapters that appeared in this book. With regard to Levin's concern about who the "informed stakeholders who are abandoning NHST" are (p. 44), we cited evidence of the first indications of movement away from NHST. Thompson (1998) corrects this assertion by citing sources from 1998 that provide evidence that a shift from NHST to the use of effect size measures is not underway. We stand corrected on this point but must point out that the sources that Thompson cites were unavailable when we developed our arguments.

We appreciate the opportunity to voice our opinions on the state of social science research and the critique of our work. None of our critics have provided sufficient evidence that the advancement of social science would be hampered if authors were required to provide more relevant information in their publications; and we found support for the establishment of research registries to

mimic the success that the medical field has had in conducting meta-analyses. Although our ideas are neither unique nor revolutionary, we believe the time for concrete action, such as that we advocate, is long overdue.

References

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

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

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

Harlow, L.L., S.A. Mulaik, & Steiger, J.H. (Eds.). (1997). *What if there were no more significance tests?* Mahwah, N.J.; Lawrence Erlbaum Associates.

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

Knapp, T. R. (1998). Comments on the statistical significance testing articles. *RESEARCH IN THE SCHOOLS*, *5*(2), 39-41.

Levin, J. (1998). What if there were no more bickering about statistical significance tests? *RESEARCH IN THE SCHOOLS*, *5*(2), 43-53.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *RESEARCH IN THE SCHOOLS*, *5*(2), 3-14.

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

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

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development*, *22*, 2-6.

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

Thompson, B. (1998). Statistical significance testing and effect size reporting: Portrait of a possible future. *RESEARCH IN THE SCHOOLS*, *5*(2), 33-38.

Vogt, P. W., (1993). *Dictionary of statistics and methodology*. Newbury Park: Sage.

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 59-62

Fight the Good Fight: A Response to Thompson, Knapp, and Levin

James M. Ernest
State University of New York at Buffalo

James E. McLean
University of Alabama at Birmingham

After discussing common sentiments in the three papers in this special issue, the authors address concerns of omission expressed by one of the critiquers and provide recommendations for the role of SST.

After reading the three papers (Knapp, 1998b; Levin, 1998; & Thompson, 1998b) that reviewed the articles by Daniel (1998), McLean and Ernest (1998), and Nix and Barnette (1998), it occurred to us that we "got off" rather lightly. In preparing our response to the contents of the other papers in this special issue of *RESEARCH IN THE SCHOOLS*, we would first like to comment on the general sentiments shared throughout the papers. Secondly, we thought that most of the comments directed toward our paper were concerned with perceived omissions. As Knapp (1998b) pointed out, the controversy has been going on for many years now, and therefore it is impossible to cite every relevant source. However, in this response we will attempt to address Knapp's concerns of omission. Finally, we would like to provide our recommendations for the role of Statistical Significance Testing (SST), agreeing with Thompson (1998b) that the status quo is not "peachy-keen" and that changes are warranted.

Levin (1998) noted that this special issue of *RESEARCH IN THE SCHOOLS* has approached the SST debate as many other forums have regarded the issue. In simple terms (and to use Levin's legal analogy), the debates have cast SST as the "bad guy" of science, often with the hope that the good rational people of the world (or at least those people interested enough to read these journal articles and participate in conferences) may hold trials not so much for, but of, the accused. Unfortunately, the accountability system for SST has not been as favorable as many accountability systems in the world. In the SST accountability system, this accords the accused a status of presumed guilty, and innocence must be proved.

When the topic of SST is raised, it is usually raised in a negative light, the faults of the procedure are considered, and then the issue is opened for proponents of SST to justify the procedure's worth. The debate--before it starts--is stacked against its use. We do not think it would be remiss to say that all people with an interest in the SST debate know there are problems with the SST practice. These problems, according to Hagen (1998), are typically centered around three broad criticisms. The criticisms are concerned with: "(a) the logical foundations of NHST [Null Hypothesis Statistical Testing], (b) the interpretations of NHST, and (c) alternative and supplementary methods of inference" (p. 801). As Hagan (1998) noted, the responses to his 1997 article (Falk, 1998; Malgady, 1998; McGrath, 1998; Thompson, 1998a; & Tryon, 1998) were concerned with all three issues; however, the bulk of Hagan's (1998) response was directed at the logical validity of SST. Rather than our paper being a re-hash of the same arguments concerning the logic of the test, the purpose of our paper was to consider the value of SST as one "of several elements in a comprehensive interpretation of data" (McLean & Ernest, 1998, p. 15).

Our approach to the SST issue was to argue for the positive aspects of SST. We advocated for the use of SST (a limited but necessary use) and also for the necessary inclusion of information concerning the practical significance of the results supported with an index of replicability. As Thompson (1998b) noted, this was a "moderate approach." Also, it was interesting to see Knapp (1998b) refer to his beliefs within a middle-of-the-road position, and Thompson reflect "[m]y own views are fairly similar to those of McLean and Ernest (1998) and Daniel (1998)." When one considers that Levin (1998) confesses to be on the "pro" side in the hypothesis testing debate (with McLean & Ernest, 1998 as pro; Daniel, 1998 and Nix & Barnette, 1998 as con), one realizes that the division between pro and con is not great – one dares to say even "non-significant."

Levin's (1998) reference to the 1998 American Educational Research Association annual meeting session (titled: "A no-holds-barred, tag-team debate over the statistical significance testing controversy")

reinforces the idea that there are a number of similarities between those that consider themselves on two sides of a battle. During the debate we saw Tom Knapp and Joel Levin in the "pro" corner, and in the "con" corner were Ron Carver and Bruce Thompson. Yet, even with what seemed to be two diametrically opposed views represented by Carver and Levin, it was interesting to hear Thompson conclude his remarks by stating "I don't think anyone totally disagrees with anyone else."

With respect to Knapp's Comment 1 concerning the challenge to find the idea of "replicability" in the original writings of Mosteller and Tukey (1977), Efron and Gong (1983), Diaconis and Efron (1983), or Mooney and Duval (1993), whom he credits with developing the jackknife and bootstrap procedures: we did not claim that establish-ing replicability was part of the original purpose of these procedures. We drew the idea from current practice and the writings of Thompson (e.g., 1994). There have been many developments in science and mathematics that have gone far beyond their original purposes. For example, Bonferroni would never have guessed that his inequality would become the basis for numerous multiple compar-ison procedures. In addition, our recommendation of including an estimate of replicability was not limited to these two approaches. In fact, we believe firmly that the best method of producing support for the replicability of the findings is to replicate the study.

In response to Knapp, the comment that our manuscript omitted the Schmidt (1996) article was well received. However, it is our opinion that the addition of Schmidt's arguments do not add substantially to our original arguments. The main thrust of Schmidt's argument (1996) is to abandon SST and substitute "point estimates of effect sizes and confidence intervals around these point estimates" (p. 116). It should be noted, as Thompson (1998a) advised, that the mindless interpretation of whether the confidence interval subsumes zero is doing nothing more than null hypothesis testing. Thus, Schmidt's rationale for the use of confidence intervals was within the context of comparing multiple studies.

With reference to individual studies, Schmidt's recommendations do not address the possibility of making "something of an outcome that may be nothing more than a 'fluke,' a chance occurrence" (Levin, 1998, p. 45). Another of Schmidt's recommendations is the multiple constructions of confidence intervals, yet as Levin (1998) challenges us, "how is the researcher or reader to interpret these varying-degrees-of-confidence intervals, and what is one to conclude on the basis of them?" (p. 46).

In reflection, with the proliferation of recent articles that address the SST debate, there were many authors' articles omitted. However, within this rejoinder, we felt it appropriate to acknowledge the role of Schmidt within the history of the SST debate. Also, we felt it pertinent to note that we concur with Knapp's (1998a) final summary statement provided during the AERA tag-team debate. Specifically,

> Frank Schmidt, the prime mover in all of this fuss, advocates the discontinuation of ALL significance tests in favor of confidence intervals for single studies and the discontinuation of ALL narrative literature reviews in favor of meta-analyses for synthesizing results across studies. I am pleased to see that he appears to be losing both battles. (Emphasis in original)

Knapp's (1998) comment about Cohen was an interesting point but fails to challenge our initial comment. Knapp noted that we "claim[ed] that Cohen (1988) recommended a power of .80 as arbitrarily as Fisher recommended an alpha of .05." Knapp (1998) continued "[t]hat's not fair. He (Cohen) argued there, and elsewhere, that Type I errors are generally more serious than Type II errors and therefore beta (1 - power) should be chosen to be considerably larger than alpha." We concur, Cohen did argue this point. Also, we agree that Type I errors are generally more serious than Type II errors; however, our issue is that the choice of .80 is just as arbitrary as the choice of .05 for an alpha level. Choosing one number over another (the choice of .05 rather than .06) is an arbitrary matter; choosing .80 rather than .79 is just as arbitrary. These numbers are subjective, and although we agree that the choice of beta should be "considerably larger than alpha" whether one chooses .79 or .80 is arbitrary. With tongue-in-cheek, and in reference to Rosnow and Rosenthal's (1989) comment of "surely God loves the .06 nearly as much as the .05" (p. 1277), surely God loves the .79 nearly as much as the .80 recommendation for power.

In reviewing the research, we feel that a major problem with articles that discuss SST (such as the ones within this special issue of *RESEARCH IN THE SCHOOLS*) is that, more often than not, we are not even "preaching to the choir." It is as though we are preaching to a congregation of ministers. And, more often than not, we are not preaching, we are arguing (or debating what should be a consensus about how we report empirical information). Within our article (McLean & Ernest, 1998) and endorsed by Thompson (1998b), practices

have not appreciably affected actual research reporting. When an issue is debated for as long as this issue has been debated, consensus is rare. If an argument is made that statistical testing should be used intelligently (Levin, 1995) including other pertinent pieces of information (an estimate of practical significance, etc.), it would seem reasonable for people to discuss the pros and cons of the issue and come to some consensus.

When statements are made that attack a practice valued by others, such as that NHST "retards the growth of scientific knowledge" (Schmidt & Hunter, 1997, p. 37), nature predicts the initial reaction turning from fright, to flight, to fight. When authors come to the conclusion that "we must abandon the statistical significance test" (Schmidt, 1996, p. 115), or "educational research would be better off without statistical significance testing" (Carver, 1993, 287), researchers who place value in SST fight for the test's validity. Rather than setting up a situation where people "fight the good fight" for their particular beliefs, it would appear prudent to create a situation where it is possible to compromise beliefs. Thus, it is our recommendation that a compromise be made by accepting tests of significance (or not trying to abandon them) and requiring estimates of effect sizes (Thompson, 1998b) along with evidence of external replicability when possible.

Perhaps Suen (1992) said it best: The

> ultimate conclusion of any study and its importance is inherently a human judgement. Significance testing, being mathematical and incapable of making judgements, does not provide such answers. Its role is to filter out the sampling fluctuation hypothesis so that the observed information (difference, correlation) becomes slightly more clear and defined. Judgements can then be more definitive or conclusive. On the other hand, if significance testing fails to filter out the sampling fluctuation hypothesis (i.e., nonsignificance), we may still make our judgement based on the observed information. However, our judgement in this case can never be definitive. (p.79)

As Suen (1992) noted, the value that one may attribute to an empirical study is largely subjective and based on human judgements. Statistics should be viewed as subjective and not, as Abelson (1995) humorously noted, "a set of legal or moral imperatives, such as might be announced at a public swimming pool. (ABSOLUTELY NO DOGS OR FRISBEES ALLOWED. VIOLATORS WILL BE PROSECUTED.)" (p. 56). It is our belief (and in line with Levin's concept of story telling) that the interpretation of statistics should be an exercise of statistical detective work, using as many pieces of the puzzle as possible to inform our decisions.

As noted in our original paper (McLean & Ernest, 1998) and in Levin's response (1998), a case can be made for considering the chance-importance-replicability of empirical findings. This subjective judgement about the utility of the results should be made from as much information as possible. The art of making decisions is exactly that, an art. Ergo, information regarding SST should be included in a research report with at least one measure of practical significance, and if possible (and recommended), evidence of external replication.

Oh, and in reference to Thompson's (1998b) comment that for something to be "mainstream" it requires "a factual judgement as regards a moving target – our moving discipline" (p. 34), Webster's dictionary considers "mainstream" to be a prevailing current or direction of activity or influence. Maybe this was just our wishful thinking.

References

Abelson, R. P. (1995). *Statistics as Principled Argument.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *RESEARCH IN THE SCHOOLS, 5*(2), 23-32.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248*, 116-130.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*, 36-48.

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist, 53*, 798-799.

Hagan, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24.

Hagan, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist, 53*, 801-803.

Knapp, T. R. (1998a, April). *A summary of Tom Knapp's position regarding significance testing.* Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Knapp, T. R. (1998b). Comments on statistical significance testing articles. *RESEARCH IN THE SCHOOLS, 5*(2), 39-41.

Levin, J. R. (1995, April). *The consultant's manual of researchers' common statillogical disorders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Levin, J. R. (1998). What if there were no more bickering about significance tests? *RESEARCH IN THE SCHOOLS, 5*(2), 43-53.

Malgady, R. G. (1998). In praise of value judgements in null hypothesis testing . . . and of "accepting" the null hypothesis. *American Psychologist, 53*, 797-798.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist, 53*, 796-797.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *RESEARCH IN THE SCHOOLS, 5*(2), 15-22.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

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

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

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Suen, H. K. (1992). Significance testing: Necessary but insufficient. *Topics in Early Childhood Special Education, 12(1*), 66-81.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality, 62*, 157-176.

Thompson, B. (1998a). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Thompson, B. (1998b). Statistical significance and effect size reporting: Portrait of a possible future. *RESEARCH IN THE SCHOOLS, 5*(2), 33-38.

Tryon, W. W. (1998) The inscrutable null hypothesis. *American Psychologist, 53*, 796.

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 63-65

The Statistical Significance Controversy Is Definitely Not Over: A Rejoinder to Responses by Thompson, Knapp, and Levin

Larry G. Daniel
University of North Texas

A rejoinder is offered on the three reviews of Daniel's article (this issue) by Thompson, Knapp, and Levin. It is concluded that the controversy over statistical significance testing will no doubt continue. Nevertheless, the gradual movement of the field toward requiring additional information in the reporting of statistical results is viewed as evidence of a positive response to long-term criticisms of statistical significance testing.

In this rejoinder, I would like to (a) respond to the critiques of Bruce Thompson, Tom Knapp, and Joel Levin of my earlier article in this issue and (b) provide additional commentary as to the future direction of statistical significance testing.

Response to Three Critics

I would like to express my appreciation to the three respondents for their insightful observations and for their comments casting further light on the issues raised by the authors of the three articles appearing in this issue of the journal. Each of the respondents is a premier scholar whose contributions to the debates on statistical significance testing have been most useful as the issue has come to the forefront of methdological discussions in recent years. In their critiques of the three articles included in this issue, the three respondents have offered very useful discussions of the topic along with helpful references for those readers who might wish to explore the controversy further. My specific comments in relation to the points made by each respondent follow in the order in which they appear in this issue of the journal.

Bruce Thompson (1998) provides a nice framework for understanding the ongoing dialogue regarding statisti-cal significance testing. Thompson's reminder of the context of the current literature in which much of the controversy has developed is useful in understanding the issue. This serves as a good follow up to the historical perspective that I provided. As Thompson noted, I have shared a long association with him and his work (he has been a mentor, research collaborator, and fellow editor); hence, I was not surprised that he was in agreement with many of the points I had raised and that a number of the opinions he expressed were consistent with my own. Further, I appreciate his citing the newly revamped editorial policies of several journals in addition to those that I had mentioned, lending evidence to the importance of editorial policies in shaping practice related to the reporting of results of statistical significance tests (SSTs). Further, Thompson (1998) reiterated nicely my discussion on the inappropriateness of using SSTs for the reporting of nil hypotheses about validity and reliability coef-ficients.

I am sure that Tom Knapp (1998) anticipated that the other authors and I would be eager to respond to his list of our various "errors of commission and omission." Obviously, determining what constitutes a sin is at least somewhat dependent upon the particular book of faith to which one prescribes. Although I prefer a slightly different statistician's book of faith than the one Knapp uses, I would have to say I am guilty as charged on at least a few points. First, I appreciate Knapp's (1998) comment on the distinction between the obtained and hypothesized effect sizes, an issue that often gets lost in the discussions of issues of this type. Second, I did indeed omit Levin's (1998a) excellent review of the *What If* book (Harlow, Mulaik, & Steiger, 1997) from my original discussion. This review is noteworthy not only because of Levin's excellent review of the content of the various chapters of the book, but also due to the concise list of recommended statistical significance practices that Levin offers. Third, I did not specifically mention the chapter in the *What If* book by Abelson (1997), which as Knapp (1998) indi-cated, is one of the more tightly written defenses of statistical significance testing.

Now that I have duly confessed, I would like to make a few citations from my own statistical book of faith on a couple of Knapp's other points. First, Knapp (1998) commented that resampling techniques such as jackknife and bootstrap analyses do not provide evidence of result replicability. (Levin [1998b] levels somewhat different but similarly focused criticisms at these procedures.) Even though the developers of jackknife and bootstrap techniques may not have specifically mentioned the usefulness of these procedures in providing evidence of

replicability, the procedures do indeed create varied resamplings for which results may be recomputed many times over. Clearly, the replications of results from these resamplings are somewhat biased and do not replace actual replications of the results with independent samples, but in newer areas of research, biased estimates of result replication are definitely better than no estimates of replication at all.

Knapp (1998) also questions the usefulness of "what if" analyses in which the results of SSTs are referenced to variations in sample size. Although I appreciate Knapp's concern that sample size should be carefully considered prior to the initiation of a study, it is often useful to determine at what sample size a statistically significant result would have become statistically nonsignificant and at what point a statistically nonsignificant result would have become statistically significant. These findings may advise researchers in selecting samples for *future* studies.

Knapp (1998) also splits hairs over the definition of the null hypothesis, apparently hinting at Cohen's distinc-tion between null hypotheses in their most "general sense" and "the nil hypothesis" that states that "the effect size (ES) is 0" (Cohen, 1994, p. 1000). Although this is an important distinction, Cohen (1994) reminded us that "as almost universally used, the null in H_0 is taken to mean nil, zero" (p. 1000); hence, my use of this conventional definition. Similarly, Knapp (as well as Levin, 1998b), commented on the technicalities of my example compar-ing SSTs with an *n* of 62 versus an *n* of 302. My intent was not to suggest that the relationship between *p* and F is linear, but rather to show with a fixed effect that results that were not statistically significant given a particular sample size would be much more likely to be statistically significant given a larger sample size.

Levin (1998b), in his predictably amusing style, provided some excellent comments on the several papers and the controversy. His comments on "statistical testiness" are especially interesting. As Thompson (1998) noted, not all scholars will have totally positive opinions about editorial policies, such as the ones I prescribed, that encourage specific practices in the reporting of the results of SSTs. Here, Levin voices at least one oft-heard complaint leveled at such editorial policies, namely, that regulation of specific verbiage transforms editors from being scholarly gatekeepers to statistical police. Although I am an ardent supporter of academic freedom, I do feel that regulation of vocabulary so as to avoid miscommunication is essential, and, as an editor, I have with some frequency felt it necessary to correct authors' verbiage so as to enhance their clarity of communication. Without a doubt, the term "significant" constitutes one of the more significant (pun intended) instances of miscommunication in social science literature, especially among readers who may not be familiar with the logic underlying SSTs. And, even though, as Levin (1998b) suggested, the specific written context may sometimes disambiguate the use of the term "significant," I would prefer to require routine use of "statistically" before "significant" so as to avoid overlooking instances in which the term should have been modified thusly but was not.

I feel that Levin somewhat overstated my position on statistical significance testing when he suggested I advocated that "the research world will be a far better place when the hypothesis-testing devil is ousted by the effect-size angel." Although I would clearly acknowledge the heavenliness of effect size reporting, I do not see hypothesis testing as the devil, but rather as an oft-tormented, though well-intended, soul who needs the demon of misinterpretation exorcized from him. In fact, in this regard, my position is not extremely dislike the one stated by Levin: report both effect size estimates and results of SSTs, then allow the readers of the research report to draw their own conclusions about result importance.

Comments on the Future of Statistical Significance Testing

Contrary to Levin's hopeful assertion that perhaps one day soon the bickering over statistical significance testing will be quelled, I do not see that happening very soon. Rather, I agree with Thompson (1998) that the status quo regarding the use of statistical significance testing is far from "peachy keen." Unfortunately, the literature is still rife with studies in which authors have misused and misinterpreted SSTs. As long as this remains the case, the voices of reformers as well as defenders of statistical significance testing will continue to be loudly heard. The battle will continue to rage for some time to come with perhaps an occasional quietus as other important methodological issues emerge followed by rekindling of the flames of debate as thoughtful research-ers continue to see errors in the reporting of SSTs.

Despite the slowness of progress in reforming practice relative to statistical significance testing, it is encouraging to see that an increasing number of social science journals are adopting editorial policies that call for better reporting of the results of SSTs (Thompson, 1998) following the suggestions found in the APA manual (APA, 1994). The adoption and enforcement of stricter editorial policies regarding the reporting of the

results of statistical significance testing by an increasing number of social science journals will perhaps eventually move the field toward improved practice. At the recent annual meeting of the Mid-South Educational Research Asso-ciation, Jim McLean, Co-Editor of this journal held a session in which he solicited input from the association members regarding the journal's potential adoption of an editorial policy on statistical significance testing. As a session participant, I was pleased to see that the group overwhelmingly favored such a policy. I look forward to seeing how Jim and Co-Editor Alan Kaufman handle the input gathered during that session.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would have to be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

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

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

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Knapp, T. R. (1998). Comments on the statistical significance articles. *RESEARCH IN THE SCHOOLS*, *5*(2), 39-41.

Levin, J. R. (1998a). To test or not to test H_0? *Educational and Psychological Measurement*, *58*, 313-333.

Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? *RESEARCH IN THE SCHOOLS*, *5*(2), 43-53.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *RESEARCH IN THE SCHOOLS*, *5*(2), 33-38.