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.