The Journal of Experimental Education, 2002, 71(1), 83–92

Problems With Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say?

JEFFREY A. GLINER
NANCY L. LEECH
GEORGE A. MORGAN
Colorado State University

ABSTRACT. The first of 3 objectives in this study was to address the major problem with Null Hypothesis Significance Testing (NHST) and 2 common misconceptions related to NHST that cause confusion for students and researchers. The misconceptions are (a) a smaller *p* indicates a stronger relationship and (b) statistical significance indicates practical importance. The second objective was to determine how this problem and the misconceptions were treated in 12 recent textbooks used in education research methods and statistics classes. The third objective was to examine how the textbooks' presentations relate to current best practices and how much help they provide for students. The results show that almost all of the textbooks fail to acknowledge that there is controversy surrounding NHST. Most of the textbooks dealt, at least minimally, with the alleged misconceptions of interest, but they provided relatively little help for students.

Key words: effect size, NHST, practical importance, research and statistics textbooks

THERE HAS BEEN AN INCREASE in resistance to null hypothesis significance testing (NHST) in the social sciences during recent years. The intensity of these objections to NHST has increased, especially within the disciplines of psychology (Cohen, 1990, 1994; Schmidt, 1996) and education (Robinson & Levin, 1997; Thompson, 1996). In response to a recent survey of American Educational Research Association (AERA) members' perceptions of statistical significance tests and other statistical issues published in *Educational Researcher*, Mittag and

Address correspondence to Jeffrey A. Gliner, 206 Occupational Therapy Building, Colorado State University, Fort Collins, CO 80523-1573. E-mail: Gliner@cahs.colostate.edu

Thompson (2000) concluded that "Further movement of the field as regards the use of statistical tests may require elaboration of more informed editorial policies" (p. 19).

The American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson & the APA Task Force on Statistical Inference, 1999) initially considered suggesting a ban on the use of NHST, but decided not to, stating instead, "Always provide some effect size estimate when reporting a *p* value" (p. 399). The new APA (2001) publication manual states, "The general principle to be followed . . . is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship" (p. 26).

Although informed editorial policies are one key method to increase awareness of changes in data analysis practices, another important practice concerns the education of students through the texts that are used in research methods and statistics classes. Such texts are the focus of this article.

We have three objectives in this article. First, we address the major problem involved with NHST and two common misconceptions related to NHST that cause confusion for students and researchers (Cohen, 1994; Kirk, 1996; Nickerson, 2000). These two misconceptions are (a) that the size of the *p* value indicates the strength of the relationship and (b) that statistical significance implies theoretical or practical significance. Second, we determine how this problem and these two misconceptions are treated in textbooks used in education research methods and statistics classes. Finally, we examine how these textbook presentations relate to current best practices and how much help they provide for students.

The Major Problem With NHST

Kirk (1996) had major criticisms of NHST. According to Kirk, the procedure does not tell researchers what they want to know:

> In scientific inference, what we want to know is the probability that the null hypothesis (H_0) is true given that we have obtained a set of data (D); that is, $p(H_0|D)$. What null hypothesis significance testing tells us is the probability of obtaining these data or more extreme data if the null hypothesis is true, $p(D|H_0)$. (p. 747)

Kirk (1996) went on to explain that NHST was a trivial exercise because the null hypothesis is always false, and rejecting it is merely a matter of having enough power. In this study, we investigated how textbooks treated this major problem of NHST.

Current best practice in this area is open to debate (e.g., see Harlow, Mulaik, & Steiger, 1997). A number of prominent researchers advocate the use of confidence intervals in place of NHST on grounds that, for the most part, confidence intervals provide more information than a significance test and still include information necessary to determine statistical significance (Cohen,

1994; Kirk, 1996). For those who advocate the use of NHST, the null hypothesis of no difference (nil hypothesis) should be replaced by a null hypothesis specifying some nonzero value based on previous research (Cohen, 1994; Mulaik, Raju, & Harshman, 1997). Thus, there would be less chance that a trivial difference between intervention and control groups would result in a rejection of the null hypothesis.

The Size of the p Value Indicates the Strength of the Treatment

Outcomes with lower *p* values are sometimes interpreted by students as having stronger treatment effects than those with higher *p* values; for example, an outcome of $p < .01$ is interpreted as having a stronger treatment effect than an outcome of $p < .05$. The *p* value indicates the probability that the outcome could happen, assuming a true null hypothesis. It does not indicate the strength of the relationship because although *p* values do not provide information about the size or strength of the effect, smaller *p* values, given a constant sample size, are correlated with larger effect sizes. This fact may contribute to the misconception that this article is designed to clarify.

How prevalent is this misinterpretation? Oakes (1986) suggested,

> It is difficult, however, to estimate the extent of this abuse because the identification of statistical significance with substantive significance is usually implicit rather than explicit. Furthermore, even when an author makes no claim as to an effect size underlying a significant statistic, the reader can hardly avoid making an implicit judgment as to that effect size. (p. 86)

Oakes found that researchers in psychology grossly overestimate the size of the effect based on a significance level change from .05 to .01. On the other hand, in the AERA survey provided by Mittag and Thompson (2000), respondents strongly disagreed with the statement that *p* values directly measure study effect size. One explanation for the difference between the two studies is that the Mittag and Thompson (2000) survey question asked for a weighting of agreement with a statement on a 1–5 scale, whereas Oakes embedded his question in a more complex problem.

The current best practice is to report the effect size (i.e., the strength of the relationship between the independent variable and the dependent variable). However, Robinson and Levin (1997) and Levin and Robinson (2000) brought up two issues related to the reporting of effect size. Is it most appropriate to use effect sizes, confidence intervals, or both? We agree with Kirk (1996), who suggested that when the measurement is in meaningful units, a confidence interval should be used. However, when the measurement is in unfamiliar units, effect sizes should be reported. Currently there is a move to construct confidence intervals around effect sizes (Steiger & Fouladi, 1997; Special Section of *Educational and Psychological Measurement, 61*(4), 2001). Computing these confidence intervals

involves use of a noncentral distribution that can be addressed with proper statistical software (see Cumming & Finch, 2001).

Should effect size information accompany only statistically significant outcomes? This is the second issue introduced by Robinson and Levin (1997) and Levin and Robinson (2000). The APA Task Force on Statistical Inference (Wilkinson et al., 1999) recommended always presenting effect sizes for primary outcomes. The Task Force further stated that "reporting effect sizes also informs power analyses and meta-analyses needed in future research" (p. 599). On the other hand, Levin and Robinson (2000) were adamant about not presenting effect sizes after nonsignificant outcomes. They noted a number of instances of single-study investigations in which educational researchers have interpreted effect sizes in the absence of statistically significant outcomes. Our opinion is that effect sizes should accompany all reported *p* values for possible future meta-analytic use, but they should not be presented as findings in a single study in the absence of statistical significance.

Statistical Significance Implies Theoretical or Practical Significance

A common misuse of NHST is the implication that statistical significance means theoretical or practical significance. This misconception involves interpreting a statistically significant difference as a difference that has practical or clinical implications. Although there is nothing in the definition of statistical significance indicating that a significant finding is practically important, such a finding may be of sufficient magnitude to be judged to have practical significance.

Some recommendations to facilitate the proper interpretation of practical importance include Thompson's (1996) suggestion that the term "significant" be replaced by the phrase "statistically significant" to describe results that reject the null hypothesis and to distinguish them from practical significance or importance. The AERA members survey (Mittag & Thompson, 2000) strongly agreed with this statement.

Kirk (1996) suggested reporting confidence intervals about a mean for familiar measures and reporting effect sizes for unfamiliar measures. However, as more researchers advocate the reporting of effect sizes to accompany statistically significant outcomes, we caution that effect size is not necessarily synonymous with practical significance. For example, a treatment could have a large effect size according to Cohen's (1988) guidelines and yet have little practical importance (e.g., because of the cost of implementation). On the other hand, Rosnow and Rosenthal (1996) studied aspirin's effect on heart attacks. They demonstrated that those who took aspirin had a statistically significant lower probability of having a heart attack than those in the placebo condition, but the effect size was only $\phi = .034$. One might argue that phi is not the best measure of effect size here because when the split on one dichotomous variable is extreme compared with the other

dichotomous variable, the size of phi is constricted (Lipsey & Wilson, 2001). However, the odds-ratio from these data was only 1.8, which is not considered strong (Kraemer, 1992). The point here is that one can have a small effect size that is practically important, and vice versa. Although this effect size is considered to be small, the practical importance was high, because of both the low cost of taking aspirin and the importance of reducing myocardial infarction. Cohen emphasized that context matters and that his guidelines (e.g., $d = 0.8$ is large) were arbitrary. Thus, what is a large effect in one context or study may be small in another.

Perhaps the biggest problem associated with the practical significance issue is the lack of good measures. Cohen (1994) pointed out that researchers probably were not reporting confidence intervals because they were so large. He went on to say, "their sheer size should move us toward improving our measurement by seeking to reduce the unreliable and invalid part of the variance in our measures" (p. 1002).

Method

Six textbooks used in graduate-level research classes in education and six textbooks used in graduate-level statistics classes in education were selected for this study. We tried to select a diverse set of popular, commonly used textbooks, almost all of which were in at least the second edition. We consulted with colleagues at a range of universities (from comprehensive research universities to those specializing in teacher training to smaller, private institutions) about the textbooks they used, and we included these books in our sample. The statistics textbooks either referred to education in the title or the author was in a school of education; they covered at a minimum through analysis of variance (ANOVA) and multiple regression. The textbooks used for this study are listed in the references and are identified with one asterisk for research books and two asterisks for statistics books.

We made judgments about the textbooks for each of the issues and examined all the relevant passages for each topic. Each author independently rated two thirds of the textbooks, yielding two raters per textbook. Table 1 shows the rating system with criteria for points and an example of how the criteria were used for one of the issues.

Table 2 shows the interrater reliability among the three judges. Although exact agreement within an issue was quite variable, from a high of 100% to a low of 42%, there was much less variability (from 92% to 100% agreement) among raters for close agreement (i.e., $\pm$ 1 point). The strongest agreement was for issue 3, which posits that statistical significance does not mean practical importance, on which there was 100% agreement for all texts. This issue also had the highest average rating among the three (see Table 3), indicating that the issue was typically presented under a separate heading so that it was easy for the raters to find and evaluate. If the raters disagreed, they met and came to a consensus score, which was used for Table 3.

TABLE 1
The Criteria and Examples of How the Ratings Were Used

Rating	Degree of emphasis	Example (from statistical vs. practical performance)
0	None	No mention of the issue of practical vs. statistical significance.
1	Indirect	This book discussed briefly whether a difference was real. No specific information included in the index or text about practical importance.
2	Direct statement but brief; easily missed (no heading, box, examples, etc.)	These books had only a few statements about this issue, no headings contrasting statistical and practical significance, and nothing in the index about this issue. Thus, the relatively isolated statements could easily be missed.
3	More than brief statement, but not not very helpful in forms of examples of best practice	Although these books had several statements such as "results can be statistically significant without being important," they were usually an isolated point in a broader section (e.g., on the level of significance). The examples (e.g., about sample size and correlation) did not provide much help in terms of deciding what is practically important.
4	Clear statement with emphasis, examples, or both and with help about best practice	In these books, there was discussion of the issue in the sections on several or all of the major statistics. There were headings such as "statistical versus practical significance." In addition to repeated statements that not all statistically significant results have practical importance, there were helpful examples.

TABLE 2
Interrater Reliability

Issue	Exact agreement (%)	Close agreement (%)
Controversy about NHST	83	92
p does not indicate the strength of the relationship	42	92
Statistical significance does not mean practical importance	100	100

Results

Table 3 shows the percentage of books that covered each of the topics at a rating of at least 2 (direct but brief statement) and the average rating for each of the issues.

TABLE 3
Percentage of Texts (Rating 2 or More) and Average Rating for Each of the Three Issues

	Research texts		Statistics texts		Combined research and statistics texts	
Issue	%	*M*	%	*M*	%	*M*
Controversy about NHST	0	0.00	33	1.17	17	.58
p does not indicate the strength of the relationship	67	1.50	67	1.67	67	1.58
Statistical significance does not mean practical importance	100	3.33	67	3.17	84	3.25

The Major Problem With NHST

The issue of null hypothesis testing had the lowest average overall rating (0.58) and was covered at a level of 2 or more in only 17% (2 out of 12) of the texts. It was rarely addressed in any of the research texts, and only mentioned indirectly then. At least 1 text, however, made an effort to address NHST, citing the following example:

> Fisher opposed the idea of an alternative hypothesis. This was the creation of . . . (Neyman & Pearson) . . . whose views Fisher vehemently opposed . . . nevertheless, it became standard practice that when rejecting the null hypothesis, one accepts the alternative. It must be emphasized, however, that a *p*-value does not give the probability of either the null or the alternative hypothesis as being true. (Minium, King, & Bear, 1993, p. 294)

This quote was found in a large box titled "Point of Controversy—Dichotomous Hypothesis-Testing Decisions."

The Size of the p Value Indicates the Strength of the Treatment

This issue also had a low average overall rating (1.58) but was stated directly (i.e., covered at a level of 2 or more) in two thirds of the texts. There appeared to be no difference in the percentage of texts or depth of coverage between research and statistics texts. When covered at a level of 3 or 4, a typical excerpt is as follows: "Recommendation 14: Do not use tests of statistical significance to evaluate the magnitude of a relationship" (Fraenkel & Wallen, 2000, p. 273). The recommendations are indented and in italics, making them obvious and easy for the student to see and realize that it was important. Some of the texts mentioned the use of effect size measures to indicate the strength of the relationship, but many did not.

Statistical Significance Versus Practical Significance

The issue of practical significance had the most coverage (84% of textbooks had direct statements) and the highest average overall rating (3.25). Most often, this issue was covered with more than a brief statement, with emphasis, and was frequently presented under a separate heading. Typical statements for this issue were "Just because a result is statistically significant (not due to chance) does not mean that it has any practical or educational value in the real world in which we all work and live" (Fraenkel & Wallen, 2000, p. 254). Or, "Remember that a 'reject' decision, by itself, is not indicative of a useful finding" (Huck, 2000, p. 199). However, we judged the discussions that followed these and similar statements to be less helpful on the basis of an examination of the examples and the context in which the statement was included. It was clear that statistical significance is not the same as practical significance or importance, but it was usually less clear how to know whether a result has practical importance.

Discussion

Our interpretation is similar to that of Mittag and Thompson (2000), who noted "Our results contain some heartening and disheartening findings" (p. 19). Most disheartening was the failure of almost all of these recent texts to acknowledge that there is controversy surrounding NHST. Although many of the texts provided detailed information on confidence intervals and effect sizes, few related this information to hypothesis testing. This was especially true for effect sizes; many textbooks discussed effect sizes only in the contexts of computing power or of meta-analysis. On the positive side, most of the texts dealt, at least minimally, with the two misconceptions of interest for our article. A few of these textbooks went into detail and provided recommendations similar to those suggested by Kirk (1996).

Why was there a discrepancy between the many articles acknowledging problems with NHST and the failure to recognize these problems by these research and statistics textbooks? We suggest three explanations for this apparent discrepancy. The first concerns textbook revisions. Most of these texts, especially the research texts, were in their third to sixth edition, and they had originally been published before 1990 when NHST was less controversial. Our speculation is that the authors of research textbooks in which we found little or no statement of the NHST controversy, focused their revisions on updating the content literature about the studies and methods they cited. They may not have updated the statistics chapters much, perhaps assuming that statistics do not change. In addition, adding information about how to compute and report effect size and confidence intervals would change too many sections of the text. Authors have to be practical, considering publishing company deadlines and competition from other newer texts. Thus, textbook revisions in this area have generally been limited.

A second possible explanation for the failure to include NHST issues in textbooks concerns the level of depth, difficulty of concepts, and students' prior knowledge. The textbooks that we reviewed were intended for master's- or doctoral-level students in education, often as a first course in research or statistics or one taken many years after having an earlier such course. The logic of hypothesis testing is relatively difficult to understand, especially if students are not familiar with research design and statistics.

Our third possible explanation relates to best practice. Although there is general acknowledgement that each of the topics we explored should be covered in research and statistics textbooks, there is not general agreement about how it should be covered. This is especially the case with regard to how to decide whether a statistically significant finding has practical importance. There is also controversy about best practice for hypothesis significance testing (Harlow, Mulaik, & Steiger, 1997) and effect size reporting (Levin & Robinson, 2000; Robinson & Levin, 1997). Textbook authors (and publishers) are often reluctant to put best practices in print that may be changed in the next few years.

We have three recommendations to those who are in the process of writing or revising a research or statistics text. First, consider introducing a section on the NHST controversy. This section would, at the minimum, point out that there is currently debate about whether NHST is the best method for advancing research in the fields of education and the social sciences. Second, because the fifth edition of the APA (2001) publication manual includes information on the importance of reporting effect sizes and confidence intervals, an author should provide specific examples (as might be published in a journal) of what to do following a statistically significant outcome. We also recommend reporting effect size for nonsignificant outcomes. Third, provide more help (with examples) for deciding whether a result has practical significance or importance.

REFERENCES

American Psychological Association (APA). (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

*Ary, D., Jacobs, L. C., & Razavieh, A. (1996). *Introduction to research in education* (5th ed.). Fort Worth, TX: Holt, Rinehart and Winston.

**Cobb, G. W. (1998). *Design and analysis of experiments*. New York: Springer-Verlag.

Cohen, J. (1988). *Power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

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

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

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

*Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). Boston: McGraw-Hill.

*Gall, J. P., Gall, M. D., & Borg, W. (1999). *Applying educational research: A practical guide* (4th ed.). New York: Addison Wesley Longman.

*Gay, L. R., & Airasian, P. (2000). *Educational research: Competencies for analysis and application*

(6th ed.). Columbus, OH: Merrill.
**Glass, G. V, & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Boston: Allyn and Bacon.
Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
**Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston: Houghton Mifflin.
**Huck, S. W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.
Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.
Kraemer, H. C. (1992). *Evaluating medical tests.* Newbury Park, CA: Sage.
*Krathwohl, D. R. (1998). *Educational and social science research: An integrated approach* (2nd ed.). New York: Addison Wesley Longman.
Levin, J. R., & Robinson, D. H. (2000). Rejoinder: Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher, 29,* 34–36.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.
*McMillan, J. H., & Schumacher, S. (1997). *Research in education: A conceptual introduction* (4th ed.). New York: Addison Wesley Longman.
**Minium, E. W., King, B. M., & Bear, G. (1993). *Statistical reasoning in psychology and education* (3rd ed.). New York: Wiley.
Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29,* 14–20.
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. 65–116). Mahwah, NJ: Erlbaum.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.
Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences.
New York: Wiley.
Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26,* 21–26.
Rosnow, R. L., & Rosenthal, R. (1996). *Beginning behavioral research* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.
**Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Needham Heights, MA: Allyn & Bacon.
Steiger, J. H., & Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). *What if there were no significance tests?* (pp. 221–258). Mahwah, NJ: Erlbaum.
Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25,* 26–30.
Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

---

*Textbooks used in educational research classes.
**Textbooks used in educational statistics classes.