EXCEPTIONALITY, *11*(4), 209–221

Null Hypothesis Significance Testing: Philosophical and Practical Considerations of a Statistical Controversy

Paul L. Morgan
George Peabody College
Vanderbilt University

This article first outlines the underlying logic of null hypothesis testing and the philosophical and practical problems associated with using it to evaluate special education research. The article then presents 3 alternative metrics—a binomial effect size display, a relative risk ratio, and an odds ratio—that can better aid researchers and practitioners in identifying important treatment effects. Each metric is illustrated using data from recently evaluated special education interventions. The article justifies interpreting a research result as significant when the practical importance of the sample differences is evident and when chance fluctuations due to sampling can be shown to be an unlikely explanation for the differences.

No statistical method has been as strongly condemned for as long as null hypothesis significance testing (NHST). Bakan called NHST an exercise in "mindlessness in the conduct of research" (Bakan, 1966, p. 436). Carver condemned it as a "corrupt form of the scientific method" (Carver, 1978, p. 397). Schmidt and Hunter dismissed it as "disastrous" (Schmidt & Hunter, 2002, p. 66).

Despite a long history of criticism, however, NHST continues to be both widely employed and misunderstood. For example, researchers employed significance tests in 97% of the sampled studies published between 1940 and 1999 in the *Journal of Applied Psychology* (Finch, Cumming, & Thomason, 2001). Yet researchers consistently fail to grasp even the fundamentals of the logic behind NHST (e.g., Mittag & Thompson, 2000; N. Nelson, Rosenthal, & Rosnow, 1986; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). This situation lead Tyron (1998) to decry "the fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress towards correcting this situation" (p. 796).

Requests for reprints should be sent to Paul L. Morgan, Box 328, Peabody College, Vanderbilt University, Nashville, TN 37203. E-mail: paul.l.morgan@vanderbilt.edu

Contemporary efforts at education may even be contributing to this lack of progress. For example, Gliner, Leech, and Morgan (2002) found that many graduate-level research and statistics texts rarely addressed the major problems with using NHST. The texts were often unclear about alternative methods for judging the practical importance of a result. Given that researchers publishing in peer-reviewed journals represent one of our best "consumer protections" against educational fads (Stanovich, 1993/1994, p. 288), the continued widespread confusion about NHST argues for a more detailed understanding of its limitations by both researchers and practitioners.

This article makes three specific contributions toward understanding the utility of NHST in special education research. First, because many of the misunderstandings of NHST result from confusion about its basic logic (e.g., Zuckerman et al., 1993), this article presents the underlying philosophical framework of NHST. Second, this article outlines the main philosophical and practical problems associated with using NHST in special education research. Third, this article advocates for three alternative metrics—a binomial effect size display (BESD), a relative risk ratio, and an odds ratio—that can help special education researchers avoid over-reliance on NHST in identifying treatment effects. Each metric is especially useful and informative in identifying the practical importance of an effect. Together, these three contributions should help advance much-needed progress (e.g., Gliner et al., 2002) in properly understanding the utility of one of the field's most widely used and abused statistical tools.

WHAT IS NULL HYPOTHESIS TESTING?

Today's NHST is a hybrid of two different sets of statistical methods (Harlow, 1997). Ronald Fisher put forth the first a set of methods in the early half of the 20th century (McClure & Suen, 1994). Fisher established a single null hypothesis with a known distribution, against which the probability of a particular result was to be compared (Gill, 1999). Jerzy Neyman and Egon Pearson put forth the second set of methods in reaction to Fisher's methods (McClure & Suen, 1994). Neyman and Pearson outlined a decision process for evaluating alternative hypotheses, rather than one null hypothesis (Gill, 1999). Together these two sets of methods evolved into today's NHST (Harlow, 1997).

NHST formulates an artificial, statistical hypothesis to evaluate a research hypothesis. The statistical hypothesis, termed the *null hypothesis*, states that there are no systematic differences between the populations from which the two samples are drawn (Oakes, 1986). Any observed differences between the two samples result only from chance. The collected data are then tested in light of this null hypothesis. Results may lead to either a rejection of the null hypothesis or a failure to reject the null hypothesis.

Rejection of the null hypothesis means that the claim that group differences are unsystematic is untenable. Thus, rejection of the null hypothesis is indirect evidence for the research hypothesis because the statistical test helps eliminate chance as a plausible explanation for the differences between the samples (Fan, 2001). Failure to reject the null hypothesis means that chance cannot be discounted as a reason for observed differences between the two samples.

THE LOGIC OF NULL HYPOTHESIS TESTING

The logic of the NHST is based on inductive inference (Fisher, 1942; Kreuger, 2001). Early proponents of inductive inference, such as David Hume in the mid-1700s and Karl Pearson in the early 1900s, claimed that expectations of future events could be justified based only on the sequence or frequency of previous experiences (Alexander, 1972; MacNabb, 1972). For example, one's expectation that a falling teacup will break on hitting the floor results from prior observations that, when dropped, fragile things often break. To Hume and Pearson, cause and effect could be justified only as a series of extended coincidences, which one begins associating with an expectation (Black, 1972).

One implication of this philosophical framework is that, although it is possible to prove something false, it becomes impossible to prove something true (Howell, 1997). There can be no a priori basis for asserting why a particular event has to occur if the idea of cause and effect is based only on the previous occurrence of like events. For example, one might hypothesize that "all dogs bark" after hearing hundreds of dogs do so. Yet one cannot deny the possibility, however improbable based on previous observation, of discovering a dog that, say, talks. Likewise, experiments relying on NHST cannot attempt to prove the presence of a treatment effect, but rather can only falsify a hypothesis that a treatment effect is not present.

Fisher, like Hume and Pearson, thought that inductive inference was the only process that allowed essentially new knowledge to come into the world (Fisher, 1942). Therefore, the focus using Fisher's set of methods is to challenge, or falsify, a hypothesis that a particular treatment led to sample differences (Mulaik, Raju, & Harshman, 1997). Accordingly, a phenomenon to Fisher was "experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" (Fisher, 1942, p. 14). When differences between groups are unlikely after statistical testing, researchers can attribute these differences to manipulated causes and thus can expect to observe such differences again (Krueger, 2001). Thus, the basis of NHST is falsification based on probability (Mulaik et al., 1997).

PROBLEMS WITH NULL HYPOTHESIS TESTING

Many philosophical and practical problems are associated with relying on NHST to falsify a hypothesis. Researchers should keep these in mind when employing NHST to evaluate research findings. First, as a philosophical problem, the logic of null hypothesis testing relies on an inappropriate combination of syllogistic reasoning with probability testing (Cohen, 1994). Proving the null effect thus becomes a logical impossibility. Second, as a set of practical problems, rejection of the null hypothesis does not allow experimenters to make claims about the truth of any alternative hypotheses (Carver, 1978), while retaining the null hypothesis does not allow experimenters to confirm chance as the cause of observed sample differences (Cohen, 1994). Third, as an additional set of practical problems, NHST may yield either statistically significant results of little or no practical value (Thompson, 1999) or statistically nonsignificant

results that nevertheless remain highly important (Gliner, Morgan, Leech, & Harmon, 2001).

Philosophical Problems

NHST mixes logical reasoning with probabilistic interpretation (Cohen, 1994; Hofmann, 2002). Consequently, NHST does not allow categorical decisions about whether a claim is true or false. This makes proving the null effect an inherently impossible undertaking (Cohen, 1994). Instead, NHST only provides experimenters with something of a "reasonable doubt" that chance fluctuation explains differences between two samples.

The logic of hypothesis testing is based on a form of syllogistic reasoning termed a *modus tollens* (Hofmann, 2002; Kreuger, 2001; Martinez-Pons, 1999). A proof using a modus tollens is obtained through denial of a consequence (Brody, 1972). Modus tollens arguments take the form: "If A then not B: B; therefore, not A." Or, "if Sarah has homework tonight, she will not go to the game. She went to the game, therefore, she did not have homework."

Although a modus tollens argument structure is used in NHST, the test presents its premises probabilistically rather than categorically. The argument structure of NHST is therefore: "If A, then B is highly unlikely. B has occurred, therefore A is highly unlikely." Or, more technically, but no less probabilistically: "If it were true that no systematic differences exist between the means of the populations from which these samples came, then the probability that observed means would be as different as they are is less than five in one hundred." NHST therefore does not evaluate the logical nature of an event, but instead evaluates the rareness of an event (Carver, 1978). This makes it a conceptually flawed tool for researchers wanting to evaluate a hypothesis using falsification (Martinez-Pons, 1999). That is, NHST dilutes the logical rigor and value of the syllogism with probabilistic wording (Cohen, 1994; Hofmann, 2002), making it impossible to prove or disprove a null effect.

Practical Problems

Attempting to prove the null effect is also practically problematic. First, because it is based on falsification through probability, NHST does not allow experimenters to confirm a true lack of differences between the samples' populations (Cohen, 1994). All that can be concluded from a nonsignificant result is that "it cannot be concluded that the null hypothesis is false" (Nicholls, 2001, p. 983). The tendency to confuse this latter statement with "proving the null effect" causes misinterpretations by many researchers (Glass & Hopkins, 1995). For example, Finch et al. (2001) found that 38% of the articles published in the *Journal of Applied Psychology* over the last 60 years reporting a statistically nonsignificant result interpreted it as demonstrating that the null hypothesis was true. In fact, Cohen (1994) pointed out that the null hypothesis can never be proved true, because, with a sufficiently large sample, any effect could be declared statistically significant.

Second, NHST does not allow researchers to make claims about alternative hypotheses. NHST only answers the question: "What is the probability of these data, given that the null hypothesis is true?" Answering this question does not allow researchers to confirm the research hypothesis. Moreover, NHST does not allow a researcher to evaluate the likelihood that hypotheses different than the research hypothesis lead to sample differences. Instead, NHST often provides experimenters with the illusion that the hypothesis has been confirmed, because rejection of the null hypothesis is often mistakenly seen as direct, rather than indirect, evidence of validity of the research hypothesis (Gill, 1999).

Third, use of NHST is problematic in that it may yield statistically significant results of little or no practical value (Thompson, 1999) or statistically nonsignificant results that nevertheless remain highly important (Gliner et al., 2001). The results from a NHST do not reflect the magnitude of a treatment effect (Fan, 2001; Rennie, 1998). NHST only assesses the likelihood of obtaining as large a difference as that obtained, given chance fluctuations and a true lack of difference between the samples' two populations (Glass & Hopkins, 1995; Martinez-Pons, 1999). Thus, relying only on NHST to evaluate research hypotheses is problematic in that the practical effects of some treatments are disregarded due to a nonsignificant test result whereas others having little practical value are declared statistically significant (Carver, 1978; Thompson, 1999). These problems limit the utility of NHST and suggest that, under certain conditions, a nonsignificant result still indicates indirect support for a treatment effect.

KEEPING NULL HYPOTHESIS TESTING IN PERSPECTIVE

Given NHST's conceptual and practical limitations, how should special education researchers interpret results from studies relying on significance tests? Foremost, we must remember the proper role of NHST. Too often the question "Were the results statistically significant?" is mistaken as an evaluation of the merits for an intervention. Instead, researchers should evaluate whether the effects of an intervention are of practical importance. That is, do the results of the study have significant implications for how researchers can improve the lives of individuals with disabilities? NHST only serves the relatively minor function of providing criteria for judging results trustworthy, based on the fact that they are unlikely. NHST is silent on whether the results are important (Anderson, 2001).

This is a critical distinction for special education researchers, given that sample sizes in special education research tend to be small. Regardless of the magnitude of the treatment effect, results from studies using small samples are likely to be nonsignificant due to a lack of statistical power (Lipsey, 1998). The likelihood of a nonsignificant result is inversely proportional to an experiment's sample size: The smaller the sample size, the larger the difference necessary to reject the null hypothesis (Cohen, 1992). With a large enough sample, exceedingly small and unimportant differences between two samples will likely yield a statistically significant result (Martinez-Pons, 1999). Noting this, Tukey (1991) wrote, "It is foolish to ask 'Are the effects of A and B different?' They are always different—for some decimal place" (p. 100).

Conversely, with a small-enough sample size, exceedingly large and important differences will likely yield statistically nonsignificant results. That is, with smaller samples, a test of a null hypothesis will more often yield a nonsignificant result (Finch et al., 2001), regardless of the magnitude of the treatment effect. Relying on a small sample quickly decreases the statistical power needed to detect an effect that, with a larger sample, would be labeled significant, even when the strength of the effect remains the same (Lipsey, 1998). For example, Thompson (1999) reported that experimenters using a sample size of 16 would need a variance-accounted-for effect size of 25% to reach statistical significance, whereas a sample size of 342 would reach statistical significance with a variance-accounted-for effect size of just 1%. Rosenthal and DiMatteo (2001) offered a helpful formula (for a more detailed account, see Lipsey, 1998) for keeping NHST in perspective:

$$\text{NHST} = \text{Effect Size} \times \text{Sample Size}$$

INTERPRETING A NONSIGNIFICANT RESULT

To avoid the problems associated with NHST, many statisticians have called for increased reliance on effect sizes, which are standardized measures of relative differences between samples. They argue that special education researchers may interpret a statistically nonsignificant result as significant when the relative difference between samples is evident (e.g., Cohen, 1994; Kirk, 2001; Schmidt, 1996; Thompson, 1999). A variety of methods have been advocated for deciding if a relative difference is important (e.g., Cohen, 1992; Hedges & Olkin, 1985; Wolf, 1986). For example, the importance of a treatment effect may be evaluated by effect size estimates such as adjusted variance-accounted-for measures or standardized mean differences (Rennie, 1998; Thompson, 1999). A standardized mean difference of .50, for example, means that the intervention raised the scores of the experimental group by one half of a standard deviation, so that the percentage of overlap between the distributions of scores from the two groups would only be 67% (Howell, 1997).

Unfortunately, as with NHST, over-reliance on effect size measures can be problematic. First, even experienced researchers often lack a good understanding of effect size measures (Oakes, 1982). Second, each of the more common types of effect size estimates—mean differences, percentage of variance, and mathematical models that use a value parameter to represent the size of an effect and a weight to represent its importance—have their own statistical shortcomings (see Anderson, 2001, for a discussion).

Third, small effect sizes, even when statistically significant, can lead researchers to ignore promising results. For example, a meta-analysis by Smith and Glass (1977) reported an *r* of .32 for the effect of psychotherapy, leading Rimland (1979) to question whether such a modest effect size was the "death knell for psychotherapy" (p. 192). Compare Rimland's reaction to that of the Steering Committee of the Physicians Health Study Research Group (1988). The research group was tracking the results of a study on the use of aspirin to reduce the occurrence of heart attacks. The research group stopped

the study because the value of $r = .04$ was so encouraging that it was considered unethical to continue withholding the aspirin treatment from the control group. Whereas most special education researchers would not consider an $r = .04$ of any practical importance (accounting as it does for only .0016 of the variance), the magnitude of the finding is evident when "we can count ourselves among the 4 per 100 who manage to survive" a heart attack (Rosenthal, 1990, p. 775). One solution to both the limitations of NHST and the drawbacks of effect sizes is to employ metrics that identify and communicate the practical importance of a study's treatment effect, but that are more readily resistant to misunderstanding or misuse (Hallahan & Rosenthal, 2000).

THREE METRICS FOR ASSESSING PRACTICAL IMPORTANCE

Many statisticians now advocate alternative metrics to NHST and standard effect sizes when evaluating treatment effects (e.g., Hallahan & Rosenthal, 2000; Howell, 1997; Rosenthal, Rosnow, & Rubin, 2000; Tinsley & Brown, 2000). Binomial effect size displays, relative risk ratios, and odds ratios are examples of such alternative metrics. These measures differ from NHST in that they can help to better describe the data from a particular study, rather than draw conclusions or make inferences based on a particular study's data. Unlike NHST or standard effect sizes, each metric is intuitively understandable. Furthermore, because each alternative takes into account the context of the particular study, they allow researchers to better assess the practical importance of an obtained effect size. As such, metrics such as BESDs, relative risk ratios, and odds ratios hold great promise for evaluating special education interventions.

Each of these metrics is detailed in the following paragraphs. All are based on effect size *r*, or the set of statistics based on correlation, which has several advantages over effect size *d*, or the set of statistics based on mean difference (see Rosenthal & DiMatteo, 2001). Each metric is illustrated using data from interventions reported in the recent special education research literature. I do so by (a) translating a significance test into *r*; (b) graphing this *r* using one metric, a BESD; and then (c) "standardizing" it using two other metrics, a relative risk ratio and an odds ratio (Rosenthal et al., 2000).

Effects of Self-Evaluating Teacher Praise

Sutherland and Wehby (2001) studied the effects of teacher self-evaluation on the rates of praise given in classrooms for students with emotional and behavior disorders (EBD), while also measuring the number of correct responses given by the students. Praise is important in that it can help motivate students with EBD (Walker, Colvin, & Ramsey, 1995), leading to more on-task behavior (Sutherland, Wehby, & Copeland, 2000) and, hopefully, increased correct responding. Sutherland and Wehby reported an overall analysis of variance (ANOVA) that indicated no significant main effect for the treatment group (i.e., the group of students whose teacher was using self-evaluation to monitor use of praise) for total correct responses, $F(1, 18) = 1.78$, *ns*. This *F* translates into an *r* of .30, a medium-low effect size (Cohen, 1988).

A BESD translates an effect size based on *r* into a difference in outcomes. In essence, a BESD provides a graphic for assessing the difference in success rates for two groups (i.e., treatment vs. control). A BESD is obtained from *r* by computing the treatment group's success rate as 0.50 plus *r*/2, and the control condition success rate as 0.50 minus *r*/2. The results are then put into a 2×2 table, with cells labeled as A, B, C, and D (Rosenthal & DiMatteo, 2001). As shown in the Appendix, an *r* of .30 means that the success rate (i.e., students giving a correct response) increased from 35% to 65% when teachers used self-evaluation to monitor their use of praise.

After displaying the BESD, researchers can then, quite simply, compute either a relative risk ratio or an odds ratio to assess the practical effects of an intervention. A relative risk ratio is (a) the proportion of those in the control group at risk for a bad outcome divided by (b) the proportion of those in the treatment group at risk for a bad outcome (Rostenthal & DiMatteo, 2001). In the BESD, relative risk is computed by D/B. Thus, in the Sutherland and Wehby study (2001), the relative risk is 65/35 = 1.9; students in the control group were almost twice as likely to suffer a bad outcome (i.e., failure to supply a correct response) than students in the treatment group.

Like the BESD, the odds ratio is a way for researchers to assess the practical importance of a treatment. An odds ratio is the ratio of (a) bad outcomes to good outcomes in the control group divided by (b) the ratio of bad outcomes to good outcomes in the treatment group (Howell, 1997). In the Sutherland and Wehby (2001) study, this would be the odds of giving an incorrect response if in the control group (D/C, or 65/35) to the odds of giving an incorrect response in the treatment group (B/A, or 35/65). Here, the odds ratio is 5.5. Thus, the odds of giving an incorrect response if the teacher was not using self-evaluation to monitor use of praise were almost four times greater than the odds of giving an incorrect response if the teacher was using self-evaluation to give praise. Put in reverse, students in the treatment group classrooms were only one fourth as likely to supply an incorrect response as the students in the control classrooms.

Effects of Contextualized Math Instruction

Bottge (1999) studied the effects of contextualized math instruction on the problem-solving performance of middle school students in remedial and prealgebra classes. Contextualized math instruction is a promising means of fostering skills generalization, in that it may help students to better "explore semantically rich learning environments with the knowledge they bring to the learning situation" (Bottge, 1999, p. 82). The effectiveness of contextualized math instruction was assessed using measures of fractions computation, word problems, and a contextualized problem (a transfer measure was also given). Bottge reported mixed results using two-way analysis of covariance (i.e., with instruction and class as the two factors and pretest scores entered as the covariate) and with alpha set at .10. For example, there was a significant main effect on the computation for class, $F(1, 61) = 4.86$, $p = .03$, but not for instruction, $F(1, 61) = 1.50$, $p = .23$, nor for the class by instruction interaction, $F(1, 61) = 0.90$, $p = .35$. The *F* for instruction translates into an *r* of .15, which is considered a small effect size (Cohen, 1988).

Using a BESD, this *r* of .15 means that the success rate (i.e., students giving a correct response on the computation task) increased from 42.5% to 57.5% when teachers provided students with contextualized math instruction. The relative risk ratio indicates that students in the contextualized math condition were about 1.4 times more likely to correctly solve computation problems than students in the control condition. The odds ratio indicates that students in the control condition were almost twice as likely to solve a computation problem incorrectly as students in the contextualized math condition. Put differently, students in the treatment group were about half as likely to supply an incorrect response as the students in the control classrooms.

Effects of Cooperative Learning

Gillies and Ashman (2000) investigated the effects of providing cooperative learning training to students with learning disabilities on their behaviors and learning outcomes. Although cooperative learning is a promising method for students with disabilities as it can improve both their academic skills (e.g., Fuchs, Fuchs, Mathes, & Simmons, 1997; Stevens & Slavin, 1995) and social acceptance by nondisabled peers (e.g., Fuchs, Fuchs, Mathes, & Martinez, 2002), some students with disabilities continue to display problem behaviors during cooperative learning activities (J. R. Nelson, Johnson, & Marchand-Martella, 1996). Gillies and Ashman reported that a repeated measures ANOVA showed no significant main effect for group (i.e., the group of children who received training in cooperative learning behaviors and skills vs. the group of children who participated in identical activities but who did not receive the training) for task-oriented behavior, $F(1, 20) = 2.87$, *ns*. This *F* translates into an *r* of .35, a medium-low effect size (Cohen, 1988).

This *r* means that the success rate (i.e., students showing a more task-oriented behavior) increased from 32.5% to 67.5% when teachers provided students with disabilities training in cooperative learning activities. This difference between the success rates of the treatment and control group is, of course, expressed in the *r* of .35. The relative risk ratio shows that students receiving training in cooperative learning activities were twice as likely to display task-oriented behavior than students in the control condition. The odds of displaying non-task-oriented behavior were four and one-third times greater (i.e., 67.5/32.5 vs. 32.5/67.5) when students did not receive the cooperative learning training.

The odds ratio, as well as the relative risk ratio and the BESD, all suggest the presence of promising treatment effects in the aforementioned studies. For example, students in the Sutherland and Wehby (2001) study supplied correct responses much more often when teachers used self-evaluation to monitor their use of praise. Relying only on NHST might lead special education researchers (or journal-reading consumers) to ignore this noteworthy finding because the effect would be classified as statistically nonsignificant. Moreover, the aforementioned metrics help researchers consider the context of the particular study's effect size. Whereas a standard effect size indicates that the Sutherland and Wehby treatment accounted for only 5% of the variance in rates of correct responding, a BESD shows that students supplied correct answers 30% more often

in classrooms when teachers monitored their use of praise than when the treatment was unavailable. Recasting standard effect sizes into one of the aforementioned alternative metrics can help identify findings that, although not statistically significant, remain practically important for special educators.

A FEW NOTES OF CAUTION

It is important to note that the magnitude of difference between two groups will be influenced by factors beyond just the treatment effect. These additional factors include experimental error, measurement error, confounds or nuisance variables, and the variability caused by differences between individuals in the particular study's sample (Keppel, 1991; Shadish, Cook, & Campbell, 2002). Special education researchers can improve the likelihood of finding a treatment effect by employing a variety of different design procedures (see Gersten, Baker, & Lloyd, 2000; Lipsey, 1990). For example, certain statistical techniques (e.g., analysis of covariance) reduce error variance better, and thus can help researchers detect a real difference between a treatment and control group (Lipsey, 1990, 1998).

Researchers will be more likely to detect a treatment "signal" (e.g., a larger risk or odds ratio) if they isolate the treatment while minimizing any "noise" (e.g., sampling error). Thus, it is also important to note that measures used to demonstrate practical importance are themselves influenced by sampling error. Indeed, the aim of statistical significance tests is to guard against sampling error (Lipsey, 1998; Pedhazur & Schmelkin, 1991). Special education researchers should therefore complement measures of practical importance with additional measures that evaluate the likelihood of chance fluctuations causing sampling mean differences (Fan, 2001). Additional methods that control for the influence of random variability include the counternull statistic (Hallahan & Rosenthal, 2000), confidence intervals (Kirk, 2001), replication (Gall, Borg, & Gall, 1996), meta-analysis (Schmidt, 1996), or possibly even null hypothesis tests (Fan, 2001; Krueger, 2001; Mulaik et al., 1997). In particular, replication of a particular study or meta-analysis of a number of studies would allow for a claim that chance alone is an unlikely explanation for sample differences. Fisher himself seemed to support this position. Fisher (1954) argued that, when using significance tests, "although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance" (p. 99).

CONCLUSION

In the hopes of increasing "consumer protection" against educational fads (Stanovich, 1993/1994, p. 288), this article presented both the basic logic underlying NHST and three potential alternatives for better evaluating the presence of a treatment effect. Special education researchers can interpret a statistically nonsignificant result as practically significant when both the importance of the sample differences is evident and chance

fluctuations can be shown, through a number of recommended means, to be an unlikely explanation of the differences.

ACKNOWLEDGMENTS

I thank James Bednar, Kevin Sutherland, and Paul Yoder for their insightful suggestions regarding preparation of this article. Paul L. Morgan is a doctoral student in the Department of Special Education at George Peabody College, Vanderbilt University. Beginning January 2004, he will be an Assistant Professor in the Department of Educational and School Psychology and Special Education at Pennsylvania State University.

REFERENCES

Alexander, P. (1972). Karl Pearson. *The encyclopedia of philosophy* (Vol. 6, pp. 68–69). New York: Macmillian.

Anderson, N. H. (2001). *Empirical direction in design and analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Black, M. (1972). Induction. *The encyclopedia of philosophy* (Vol. 4, pp. 169–181). New York: Macmillian.

Bottge, B. A. (1999). Effects of contextualized math instruction on problem solving of average and below-average achieving students. *The Journal of Special Education 33,* 81–92.

Brody, B. A. (1972). Logical terms, glossary of. *The encyclopedia of philosophy* (Vol. 5, pp. 57–77). New York: Macmillian.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 55–159.

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

Fan, X. (2001). Statistical significance and effect size in educational research: Two sides of the same coin. *The Journal of Educational Research, 94,* 275–282.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210.

Fisher, R. A. (1942). *The design of experiments* (3rd ed.). London: Oliver & Boyd.

Fisher, R. A. (1954). *Statistical methods for research workers* (12th ed.). London: Oliver & Boyd.

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Martinez, E. A. (2002). Preliminary evidence on the social standing of students with learning disabilities in PALS and no-PALS classrooms. *Learning Disabilities Research & Practice, 17,* 204–215.

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. C. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34,* 174–206.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). London: Longman Group.

Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing high-quality research in special education: Group experimental design. *The Journal of Special Education, 34,* 2–18.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly, 52,* 647–674.

Gillies, R. M., & Ashman, A. F. (2000). The effects of cooperative learning on students with learning disabilities in the lower elementary school. *The Journal of Special Education, 34,* 19–27.

Glass, G. V., & Hopkins, K. D. (1995). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 7,* 83–92.

Gliner, J. A., Morgan, G. A., Leech, N. L., & Harmon, R. J. (2001). Problems with null hypothesis significance tests. *Journal of the American Academy of Child and Adolescent Psychiatry, 40,* 250–252.

Hallahan, M., & Rosenthal, R. (2000). Interpreting and reporting results. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 125–149). New York: Academic.

Harlow, L. L. (1997). Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 1–21). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Hofmann, S. G. (2002). Fisher's fallacy and NHST's flawed logic. *American Psychologist, 57,* 69–70.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61,* 213–218.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56,* 16–26.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental researchers*. London: Sage.

Lipsey, M. W. (1998). Design sensitivity: Statistical power for applied researchers. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 39–68). London: Sage.

MacNabb, D. G. C. (1972). David Hume. *The encyclopedia of philosophy* (Vol. 4, pp. 74–90). New York: Macmillian.

Martinez-Pons, M. (1999). *Statistics in modern research: Applications in the social sciences and education*. New York: University.

McClure, J., & Suen, H. K. (1994). Interpretation of statistical significance testing: A matter of perspective. *Topics in Early Childhood Special Education, 14,* 89–100.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14–20.

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, Inc.

Nelson, J. R., Johnson, A., & Marchand-Martella, N. (1996). Effects of direct instruction, cooperative learning, and independent learning practices on the classroom behavior of students with behavior disorders: A comparative analysis. *Journal of Emotional and Behavioral Disorders, 4,* 53–62.

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

Nicholls, N. (2001). The insignificance of significance testing. *Bulletin of the American Meteorological Society, 82,* 981–986.

Oakes, M. (1982). Intuiting strength of association from a correlation coefficient. *British Journal of Psychology, 73,* 51–56.

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

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

Rennie, L. J. (1998). Improving the interpretation of quantitative research. *Journal of Research in Science Teaching, 35,* 237–248.

Rimland, B. (1979). Death knell for psychotherapy? *American Psychologist, 34,* 192.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775–777.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

Rosenthal, R., Rosnow, R., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York: Cambridge University Press.

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

Schmidt, F. L., & Hunter, J. E. (2002). Are there benefits from NHST? *American Psychologist, 57,* 65–66.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Stanovich, K. E. (1993/1994). Romance and reality. *The Reading Teacher, 47,* 280–291.

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians health study. *The New England Journal of Medicine, 318,* 262–264.

Stevens, R., & Slavin, R. (1995). Effects of a cooperative learning approach in reading and writing on academically handicapped and nonhandicapped students. *The Elementary School Journal, 32,* 321–351.

Sutherland, K. S., & Wehby, J. H. (2001). The effect of self-evaluation of teaching behavior in classrooms for students with emotional and behavioral disorders. *Journal of Special Education, 35,* 161–171.

Sutherland, K. E., Wehby, J. H., & Copeland, S. R. (2000). Effect of varying rates of behavior-specific praise on the on-task behavior of students with EBD. *Journal of Emotional and Behavioral Disorders, 8,* 2–8, 26.

Thompson, B. (1999). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children, 65,* 329–337.

Tinsley, H. E. A., & Brown, S. D. (Eds.). (2000). Multivariate statistics and mathematical modeling. *Handbook of applied multivariate statistics and mathematical modeling* (pp. 3–36). New York: Academic.

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

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

Walker, H. M., Colvin, G., & Ramsey, E. (1995). *Antisocial behavior in school: Strategies and best practices*. New York: Brooks/Cole.

Wolf, F. W. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.

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.

APPENDIX

Example Binomial Effect Size Display

Outcome	*Correct Response*		*Incorrect Response*	
Condition				
Treatment	A	65	B	35
Control	C	35	D	65

Note. Table adapted from Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Reviews of Psychology, 52,* 76. With permission, from the *Annual Review of Psychology*, Volume 52 © 2001 by Annual Reviews www.annualreviews.org