Essays in Education

Spring 3-1-2005

Statistical Significance and Evidenced-Based Policies: A Realistic View

Tara Stevens
Texas Tech University

Arturo Olivarez Jr.
Texas Tech University

CALL FOR SUBMISSIONS!

Essays in Education (EIE) is a professional, peer-reviewed journal intended to promote practitioner and academic

dialogue on current and relevant issues across human services professions. The editors of *EIE* encourage both

novice and experienced educators to submit manuscripts that share their thoughts and insights. Visit

https://openriver.winona.edu/eie for more information on submitting your manuscript for possible publication.

Follow this and additional works at: https://openriver.winona.edu/eie

Part of the Education Commons

Statistical Significance and Evidenced-Based Policies:
A Realistic View

Tara Stevens
Arturo Olivárez Jr.

Texas Tech University

Abstract

The call for evidenced-based policies in the field of education has prompted the encouragement of a greater reliance on experimental designs. Without an awareness of the limitations of Null Hypothesis Statistical Testing (NHST), these changes will likely fail in making educational research more influential. NHST does not yield practical information. Researchers and educators must work together to develop theory that is meaningful and practical that can be tested and replicated over time. The purpose of this article is to revisit the limitations of NHST while explaining how these limitations are important to practitioners.

Introduction

The implementation of evidence-based polices has challenged educators to evaluate their teaching and related outcomes using experimental or quasiexperimental designs (U.S. Congress, 2001). Legislators have encouraged administrators, teachers, and school psychologists to base their educational decisions on "proven" models (U.S. Department of Education, 1999) and have described the field of education as "incapable of the cumulative progress that follows from the application of the scientific method and from the systematic collection and use of objective information in policy making" (U.S. Department of Education, 2002, p. 48). Slavin (2002) expressed hope that this emphasis on causal designs and research-based practices will do nothing short of revolutionize education, much like the transformation experienced by other fields, such as medicine and agriculture. However, he also acknowledged that sudden breakthroughs in the field of education comparable to the Salk polio vaccine, for example, are not likely to occur. Instead, Slavin (2002) advocated for "dozens or hundreds of randomized or carefully matched experiments going on each year on all aspects of educational practice" so that "steady irreversible progress" will result to improve education (p. 19).

Certainly, we agree with Slavin (2002) that education can be revolutionized through the use of evidenced-based policies when randomized or carefully matched experiments that are replicated are employed. However, we believe that not all legislators and educators will understand the importance of the latter part of that statement, and instead will exaggerate the importance of results yielded from single experimental studies and Null Hypothesis Statistical Testing (NHST). Equating the word "proven" with "experimental-control comparisons on standards-based measures" (U.S. Department of Education, 1999) when social sciences and

education researchers avoid this word due to its inappropriateness considering the inevitable statistical and methodological weaknesses associated with NHST warrants concern.

Although NHST "is surely the most bone-headedly misguided procedure ever institutionalized" (Rozeboom, 1997, p. 335), many researchers continue to place their trust in its results. NHST is employed to determine if a result, such as a difference between control and treatment group means, is so extreme that it is not likely this difference would occur given the null hypothesis is true. A continued reliance on NHST will be necessary to meet the goals of legislators to base educational practice on research. According to Mulaik, Raju, and Harshman (1997), eliminating NHST is unlikely as long as researchers have a need to differentiate results that occur by chance and those that are systematic, which is often the case in educational research. Ironically, Chow (1998) argued that researchers ask too much of NHST, assuming that it provides practical information when it clearly does not. Although Chow (1998) posited that NHST is still a valuable statistical technique, others have accused NHST of impeding the advancement of psychology and other social sciences (Carver, 1978, 1993; Cohen, 1994; Schmidt & Hunter, 1997). "Even if properly used in the scientific method, educational research would still be better off without statistical significance testing" (Carver, 1978, p. 398).

Berkson (1938) may be the first credited with attacking the overall utility of NHST, followed by the strong criticisms identified by Rozeboom (1960), Meehl (1967), Bakan (1966), and Lykken (1968). More recently, Cohen (1990, 1994), Hunter (1997), Pollard (1993), and Schmidt (1996) have helped to revive the argument with their poignant concerns. This issue apparently central to the future of scientific inquiry into the disciplines of psychology, also captured the attention of the American Psychological Association, which appointed a task force to examine the problem (Wilkinson and Task Force on Statistical Inference APA Board of Scientific Affairs, 1999).

Although some expected the APA Task Force on Statistical Inference would recommend an outright NHST ban (Kirk, 1996), the Task Force's findings instead urged researchers to focus on basic methodological issues and the appropriate interpretation of results. These recommendations supported the view that the criticisms of NHST are most likely related to the misinterpretation of data by the researcher rather than an inherent theoretical or philosophical flaw in the statistical method itself (Wainer, 1999). Others have echoed the emphasis upon the individual researcher's responsibility in utilizing informed judgment and sensibility in his or her use of inferential statistical methods (Cohen, 1994; Grayson, 1998).

This is the opportune time to again consider this issue in the field of education. As greater numbers of researchers are being called upon to share their findings with those in practice, researchers should be reminded of their responsibility in ensuring that not only they recognize the limitations of NHST, but that consumers of their research will as well. Furthermore, we believe that educational practitioners should be an informed public, especially when the decisions they make will directly affect the education and lives of many children. Turning to textbooks will not provide a remedy for this problem as many textbooks fail to mention that a controversy even exists (Gliner, Leech, & Morgan, 2002). The purpose of this article is to revisit the theoretical and applied problems of NHST so that researchers can most appropriately

interpret and convey results to avoid misunderstanding and practitioners can understand the following practical implications.

1. A statistically significant finding does not warrant an immediate and palpable change in practice.
2. A statistically nonsignificant finding does not warrant an immediate change in practice.
3. A statistically significant finding does not indicate that the effect is strong or even important.
4. A statistically significant finding should be evaluated using one's own experience, knowledge of replicated studies, and collaborative experiences with educational researchers.

Misunderstanding the Null Hypothesis

Cohen (1994) addressed the question of what is wrong with NHST rather eloquently when he wrote, "Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997). Researchers commonly fail to understand that when a significant result does not occur, this simply means that the null hypothesis is probably true, not that it is true or that no difference is present. Numerous researchers have made incorrect statements referring to the presence of no differences or no relationships in their published work (Cohen, 1990).

Consider the example of an educational researcher interested in investigating the difference in science interest between two groups of students; one that has been able to remotely employ a university telescope to learn about science and another that has learned about science through traditional classroom activities. The researcher suspects that students who have been exposed to the novel learning opportunity, the remotely operated telescope, will subsequently report a greater interest in science. However, instead of evaluating whether or not a difference exists, NHST only allows the researcher to assess if the level of science interest is the same for both students exposed to the telescope and those not exposed to the telescope. First, consider the interpretation if the researcher completes the analysis and discovers that there is no statistically significant difference in the level of science interest between the two groups of students. Instead of indicating that no difference exists between the two groups, probabilistic reasoning states that if the null hypothesis is true, a difference observed between the two groups of students would probably not occur by chance or is not rare enough.

Second, consider the possibility in which the researcher's findings indicate a significant difference is present between the level of science interest of students who worked with the telescope and those who did not work with the telescope, with those students working with the telescope reporting higher levels of subsequent science interest. Again, the reasoning is probabilistic and indicates that in the event that no difference exists in the level of science interest reported, then the discovery that a difference was found between the two groups is unlikely. Therefore, what is actually accepted is the "unlikelihood" of the null hypothesis, not the likelihood of a difference occurring.

Although the belief is widespread that the level of significance at which the null is rejected is the actual probability that the null is correct, this is certainly not the case (Cohen, 1994) and has been coined the "permanent illusion" by Gigerenzer (1993). "Yet how often do we see such a result to be taken to mean, at least implicitly, that the effect is *significant*, that is, *important, large*. If a result is *highly* significant, say $p < .001$, the temptation to make this interpretation is all but irresistible" (Cohen, 1990, p. 1307). Because the *p* value is not a probability concerning the parameters, but a probability about the data (Grayson, 1998) our educational researcher knows nothing about the truth of the alternate or research hypothesis and cannot say that students who work with telescopes are significantly more likely than those who do not work with telescopes to have a high level of science interest. In addition, our researcher cannot estimate the probability of rejecting the null hypothesis in future replications.

Schools that have access to the university telescope may serve students who possess higher levels of socioeconomic status than those schools without access. These students may have greater opportunities to experience science outside of the school day than those students who attend the school without access to the telescope who may also come from low socioeconomic backgrounds. Because of the former group's greater overall exposure to science, these students may have a greater opportunity to develop an interest in science than those students of the latter group. In classical probability theory, an event has an equal chance or independent chance of occurring (i.e., the roll of an unbiased die), whereas certainty is lost when this equal likelihood is not present (Mood, Graybill, & Boes, 1974). The researcher's scenario can be equated to throwing biased dice, which means that each side is not equally likely to be rolled. As a result, frequency probability or a posteriori probability must be utilized. "Note that although the relative frequency of the different outcomes are predictable, the actual outcome of an individual throw is unpredictable" (Mood, Graybill, & Boes, 1974, p. 6). This reductionist method that condenses numerous individual observations to a single number, the mean, does not provide the researcher with information concerning the science interest of any individual student; the very information that concerned parents will want to know before taking on hardships to enroll their children in a science program that employs a large telescope to promote science interest.

Results from NHST tell us nothing about the individual case, and some would extend this concern to the researcher's ability to tell the research consumer about the population of students under investigation. NHST provides information about what to do with the null hypothesis; reject it or fail to reject it. However, further consideration about the meaning of the null hypothesis, or the statement that the test parameter is equal to zero, reveals yet another problem. The considerable overlap that exists among variables measured in the social sciences and education make the idea that the null hypothesis is equivalent to zero unrealistic (Cohen, 1990). Therefore, Cohen (1994) referred to the null hypothesis as the "nil" hypothesis (p. 1000). If this logic is difficult to follow, again think back to the example of the researcher investigating differences in the level of science interest subsequently reported by students who learned about science using a telescope and those who learned about science in a traditional manner. The null hypothesis indicates that the subsequent level of science interest reported by both groups of students would be exactly the same, hence no difference. In reality, this lack of a difference is highly unlikely,

not because the two groups are actually different, but because the variables will not be measured without error. Simply by chance, the two groups will be different, at least to a small degree.

Suppose the researcher uses self-report to determine the subsequent level of science interest of both groups of students. One student may overestimate her interest in science, a second student might underestimate his, and a third student might simply randomly respond to the interest items. As a result, differences in the level of science interest reported by students will simply differ due to these issues, all of which are considered error. When the null hypothesis is false, which is always the case in the real world, even to the smallest degree, a large enough sample size will result in its rejection (Cohen, 1994). "If the probability of a point hypothesis is indeterminate, the empirical discovery that such a hypothesis is false is no discovery at all, and thus adds nothing to what is already known" (Krueger, 2001). This returns us to the initial argument that NHST does not tell us what we want to know, and we are not always certain what it is really telling us. Therefore, practitioners may very well be utterly confused concerning how to utilize such flawed information, especially when they are being called upon to do so.

To best convey practical implications, revisiting our interest researcher is necessary. First, consider that the interest researcher fails to reject the null hypothesis. In other words, the researcher has not found a significant difference between the science interest scores of the group receiving science instruction using a telescope and the group receiving traditional science instruction. Does this mean that a school interested in employing novel tasks, such as working with a remotely operated telescope to increase students' science interest should save its resources and continue utilizing traditional science instruction? Furthermore, should a school that has been investing resources in novel science activities cease in this expenditure? Finally, should parents trying to increase their children's interest in science stop spending their time visiting planetariums and money purchasing amateur telescopes to view the night sky? The answer to these questions is a resounding "no" based on statistical limitations alone.

Although no significant difference was found between the mean science interest scores of the two groups, a difference was likely present. Certainly, this difference could just be a result of measurement and experimental error; however, if the researcher failed to employ a large enough sample size, the experiment may have lacked the power to actually discover the difference. In other words, the use of the telescope may have resulted in higher levels of science interest. This logic also creates concerns when considering the second circumstance in which the null hypothesis is rejected or the researcher did find significantly higher levels of interest in the group exposed to the telescope. In this case, the researcher could have employed a sample size large enough to detect even the smallest difference. As a result, differences resulting from only error and not from the actual treatment (i.e., exposure to the telescope) would be viewed as rare and not likely to occur if the null were true. This example illustrates that while schools should not eliminate programs based on a failure to reject the null, they should also not spend resources on purchasing programs or interventions based on a significant result or rejection of the null.

NHST is a technique that employs dichotomous reasoning alone. The only information yielded is whether or not the null should be rejected for the data analyzed. The research consumer must be aware that the NHST result does not provide practical information, such as whether a school should employ the treatment tested. In fact, some have pointed out that

statistical hypotheses may be reconsidered once further evidence from practical sources has been evaluated (Mulaik, Raju, & Harshman, 1997). Although schools may be presented with "significant" results and differences, such findings are only one piece of evidence that should be evaluated when making practical decisions. Furthermore, educators should recognize that significance levels only indicate the probability that the result occurred by chance, not the probability that the effect will happen in the population or the probability of the effect being replicated. Practitioners should be wary when "highly significant" results are touted for a certain treatment, intervention, or program. Instead, they should be interested in finding information concerning the effect size or issues of estimation using confidence interval.

Power Analysis, Effect Sizes, and Confidence Intervals

After discussing the problems associated with the null hypothesis, one might be surprised that a statistical method plagued with so many fundamental problems would still be employed by researchers. Interestingly, arguments over the appropriateness of NHST have been ongoing (Harlow, 1997). In addition, some who have expressed considerable opposition to the appropriateness of the method have continued to employ NHST in their published work (Greenwald, Gonzalez, Harris, & Guthrie, 1996). Even the APA Task Force on Statistical Inference did not recommend an outright ban on NHST (Wilkinson and Task Force on Statistical Inference APA Board of Scientific Affairs, 1999). Instead, the Task Force encouraged that researchers employ effect sizes, confidence intervals, and power analyses to provide the information that is really desired. A number of major journals in the field of education adopted new guidelines that required authors to supply this information as it might actually answer the questions concerning what we really want to know.

Fisher (1935), the founding father of NHST, only endorsed the advancement of science through the use of inductive inference, which is essentially related to the rejection of the null hypothesis. As aforementioned, the weaknesses related to the null hypothesis have been duly identified; however the reader might wonder about the role of the alternate or research hypothesis. Fisher did not consider the alternate hypothesis in NHST. The idea of having two hypotheses was instead introduced by Neyman and Pearson (1928). With this creation, one can estimate an alpha risk, the rejection of the null when it is true, as well as a beta risk, or the failure to reject the null when it is false. Consequently, this information allows researchers to estimate the sample size necessary to find a significant result. Researchers can also find the probability of rejecting the null hypothesis based on the effect size, alpha, and the sample size by calculating beta, which is referred to as the statistical power of the test (Cohen 1994).

Researchers should recognize that inherent differences are most always present between groups under study. Certainly these differences are often related to measurement error, such as in the example of the educational researcher studying students' science interest who found that some students overestimated their interest, whereas others underestimated it. With a large enough sample size this meager difference could result in a significant result or a Type I error. That is, the null was rejected when, in fact, it was true or true as one would expect it to be. However, sometimes a real difference can exist between groups as a result of the treatment (i.e., learning science using a telescope) that is not statistically significant. This may occur when the measurement error is great, which results in a mean difference that is not extreme enough to be

considered rare. In other words, so much variation already exists between the two group means as a result of many undetermined errors that the real difference is not identified. In this circumstance, the null is actually false, but not rejected. If the researcher would have conducted a power analysis before NHST he or she would have been able to calculate the sample size necessary to detect even a negligible effect at a prespecified level of probability. In this case, if the researcher found that the results were not significant and the effect was not detected when utilizing the calculated sample size, then the researcher could conclude that no nontrivial effect was present. Despite the method's apparent benefits, power analysis is rarely utilized by researchers, which may be related to the extremely large sample sizes required to put into practice such reasoning (Cohen, 1990).

In spite of the obvious drawback of Cohen's (1990) suggested employment of power analysis, he identified the advantage that researchers would be encouraged to take the magnitude of effects into account. NHST seems to discourage this incremental understanding of phenomena as it works to advance science in a binary fashion. Following NHST the researcher only has to decide if the null can be rejected or not. The determination is based on a cutoff criterion, often set at .05, which raises the concern of the meaning of one's findings when the *p* value only reaches .055. Information concerning to what degree the effect is supporting the null hypothesis is not available. "Because science is inevitably about magnitudes, it is not surprising how frequently *p* values are treated as surrogates for effect sizes" (Cohen, 1990, p. 1309).

Instead of relying upon such misinterpretation, the use of effect sizes, or mean differences, correlations, and squared correlations, to report the magnitude of the results has been recommended (Cohen, 1990, 1994; Harlow, 1997; Wright, 2003). Effect sizes are often reported in the context of a confidence interval, which provides a range of values with a given probability. Cohen (1994) strongly advocated for reporting the confidence intervals as they reveal information about both the "nil hypothesis" and the "non-nil hypothesis" (p.1002). Thus, the researcher who reports confidence intervals supplies information concerning whether the effect size is significant while also providing an estimate concerning what range of values it might have.

For example, suppose our interest researcher rejects the null hypothesis because a significant difference was found between the levels of science interest across two groups, with the group that learned science through interacting with a telescope reporting higher science interest than a group who learned science in the traditional manner. Rejecting the null is only a binary decision; that is, reject or fail to reject. This decision lends no insight into how large a difference in science interest exists between these two groups of students. The researcher can calculate a point biserial correlation coefficient to evaluate the strength of the relationship between the group membership and science interest. If the relationship is weak, then the researcher can assume that other factors must be explored to fully understand the development of science interest and if the relationship is strong, then the researcher may continue this line of inquiry in order to better understand how novel tasks develop science interest. Furthermore, the researcher may also calculate confidence intervals, which are determined by specified limits. The researcher may select a 95% confidence level indicating that the researcher wishes to be 95% confident that the derived mean estimate for the population will fall within the interval. Thus the

interest researcher could say that he or she is 95% confident that the mean difference between two groups in the population would fall between the calculated two limits.

Despite the apparent appeal of power analysis and effect sizes, many researchers continue to omit them from their research. This omission for some is due to their failure to understand the concerns associated with the use of NHST; however for others, this omission is related to the belief that these techniques do not lend valuable information to the researcher (Chow, 1998). Chow (1998) asserted that statistical power is not capable of revealing the probability of obtaining statistical significance. "A nonstatistical theoretical justification is required if an efficacious capability is attributed to statistical power. Because no such justification is offered, it is only proper *not* to attach any extra-statistical meaning to the term *statistical power*" (Chow, 1998, p. 183). The manipulation of sample size suggested by power analysis fails to take into account factors that may only be addressed through considerations that take place before research is conducted. Chow (1998) argued that the issue of sample size has much to do with the stability of the data, which is not determined by the sample size alone. Instead, factors such as the nature of the experimental task, the amount of practice that subjects have before the collection of data, measurement error, and the experimental design utilized also affect data stability in many unpredictable ways. The effects caused by such factors on the stability of the data cannot be detected by the power analysis and must be addressed prior to the start of the study.

In other words, the educational researcher studying science interest must consider the amount of experience that students already had with telescopes and other novel tasks, the potential problems associated with measuring interest through one's self-report, and the benefits of randomly assigning students to the two groups. These concerns as well as many other issues related to the experimental design, such as the differences in teaching styles and curriculum across the two groups, can also have an effect on the amount of science interest students subsequently report in addition to the influence of the telescope. Power analyses, effect sizes, and confidence intervals cannot and will not correct for any of these critical problems.

Chow (1998) continued his defense of NHST by questioning the appropriateness of effect sizes. At the center of his criticisms, he accused the utilization of such statistical methods as an attempt to mix statistical hypothesis testing with theory corroboration. Corroborating theory involves more than testing the statistical hypothesis (Meehl, 1978). Chow further posited that the impressiveness of the effect size depends upon the interpretation of the individual. In other words, independent of judgment, the effect size offers little about the practical impact of the result and is outside the domain of statistics. Again, consider that the interest researcher rejects the null hypothesis and finds that a significant difference exists between the science interest of those learning with the benefits of a telescope and those learning in a traditional manner. Suppose the interest researcher is concerned about the problems associated with NHST and follows the recommendation to report an effect size. The point biserial correlation indicates a moderate effect size. Is a moderate effect size large enough to suggest that schools invest a large amount of money in commercial grade telescopes to teach science to students in order to improve students' subsequent levels interest toward science? Will this improvement in science interest be large enough to make a difference in students' choices, such as enrolling in advanced science courses or selecting a science related career? These questions cannot be answered by the

effect size, but by educators and researchers who have a well developed theory concerning science interest supported by longitudinal evidence based on randomized or matched designs.

Merging Improved Methodology and Practice

Interestingly, the apparently complicated search for the understanding of problems associated with NHST leads back to concepts of basic research. Our educational researcher studying science interest could have randomly selected and randomly assigned students to either the group receiving science instruction with the telescope or to the group receiving traditional science instruction. Randomization would address the presence of systematic differences between the two groups. Recall the possibility that students attending the school that utilized the telescope may have been of upper socioeconomic status, benefiting from greater novel science experiences outside of school that could have raised their levels of science interest rather than their experience with the school telescope. A randomized design would distribute these students throughout the two groups, resulting in the random distribution of differences that would likely be small enough to not drastically alter NHST results.

Well-trained researchers would prefer the use of a randomized design; however when studying educational phenomena, researchers must often work within the existing structure. In other words, limitations in financial resources and time often make randomized designs difficult to employ. Our interest researcher would have considerable difficulty finding several schools that would allow students to leave during their science class to attend the science class to which they were randomly assigned and only for the length of time of the study. In addition, the parents of students not assigned to the treatment condition, in this case the group receiving science instruction with the telescope, would likely be upset that their children were not being given the same opportunity to increase their science interest. The researcher, for ethical purposes, would need to ultimately provide all students with the opportunity to use the telescope. Clearly randomization can be a challenging issue when working in the educational system.

Beyond issues of sampling and randomization, researchers must also consider that even when randomly selecting and assigning individuals an unusual group can result. Suppose the interest researcher randomly selects and assigns students to the two science learning groups, one with the opportunity to work with the telescope and one without. The researcher's results lead to a rejection of the null hypothesis indicating that those students working with the telescope have higher levels of science interest. Even though the students were randomly selected and assigned, which should have distributed individual differences randomly across the two groups, the students in the treatment group could have entered the study with an already elevated interest in science. Although this scenario is not probable, it is not impossible. Groups possessing extreme characteristics can occur simply by chance, even though the chance is small. Therefore, randomization is helpful, but not a solution.

This problem associated with probability must be addressed with theory development and replication. Both researchers and practitioners should consider what they already know about a phenomenon. When new results are considered in light of what is already known, new information is much easier to interpret. Although researchers may employ Bayesian methods to actually quantify what is already known, practitioners certainly have much to offer when they

analyze their experiences using qualitative guidelines. Collaboration between research and practice appears necessary in order to develop credible models that will inform education. Furthermore, this relationship will promote the continued study of developed theory so that educational outcomes can be evaluated and results replicated across schools.

Improvement in research methodology and replication will begin to provide the answers to the questions that educational researchers are seeking. In addition, a reliance on the development of theory, which involves collaboration between researchers and educators, will yield more meaningful results in spite of the flaws associated with NHST. To meet the goal of legislators to make education an evidence-based field (U.S. Department of Education, 2002), educational researchers must do more than develop experimental designs. Burkhardt and Schoenfeld (2003) argued that practitioners, including teachers and administrators, do not look to research when solving problems and likely do not even consider educational research credible. Merely conducting a greater number of experiments will likely make little difference to practitioners, especially when individual experimental results that depend upon NHST do not yield practical information. Although Burkhardt and Schoenfeld (2003) recommended structural changes to address the lack of credibility in educational research, they also emphasized the importance of appropriate theory development that stresses addressing practical problems. Without these changes, advancement in educational practice is questionable.

As the field of education is on the cusp of a revolution, researchers and educators must work together to develop theory that is meaningful and practical. However, researchers and educators must recognize that NHST is only a tool that can be employed to inform their theory development. Researchers must attend to sampling issues, random assignment, and measurement error when designing experiments and must stress the importance of power, effect size, and confidence intervals in their work so that practitioners can understand how best to utilize the information. Without an understanding of the limitations of NHST, educators are destined to make poor choices based on significant differences that have no practical value. This scenario will not revolutionize the field of education, but will certainly impede it.

References

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

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526-542.

Burkhardt, H., & Schoenfeld, A.H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32,* 3-14.

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

Chow, S. L. (1998). Precis of statistical significance: Rationale, validity, and utility. *Behavioral and brain sciences, 21,* 169-239.

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

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

Fisher, R.A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

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, 2002, 71,* 83-92.

Grayson, D. A. (1998). The frequentist façade and the flight from evidential inference. *British Journal of Psychology, 89,* 325-345.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175-183.

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. 1-17). Mahwah, NJ: Erlbaum.

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

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

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

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151-159.

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Vol. 1. Matters of public interest: Essays in honor of Paul Everett Meehl* (pp. 3-39). Minneapolis: University of Minnesota Press.

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

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology. 46,* 806-834.

Mood, A.M., Graybill, F.A., & Boes, D.C. (1974). *Introduction to the theory of statistics,* 3rd ed. New York: McGraw-Hill.

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. 65-116). Mahwah, NJ: Erlbaum.

Neyman, J., & Pearson, E. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika,* 20A, 175-240.

Neyman, J., & Pearson, E. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika. 20A,* 263-294.

Pollard, P. (1993). How significant is 'significance'? In G. Keren & C. Lewis (Eds), *A handbook for data analysis in the behavioral sciences.* Hillsdale, NJ: Erlbaum.

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

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance test?* (pp. 38-64). Mahwah, NJ: Erlbaum.

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

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance test?* (pp. 38-64). Mahwah, NJ: Erlbaum.

Slavin, R. E. (2002). Evidence-based education policies: transforming educational practice and research. *Educational Researcher, 31,* 15-21.

U. S. Congress. (2001). *No Child Left Behind Act of 2001.* Washington, DC: Author.

U. S. Department of Education. (1999). *Guidance on the comprehensive school reform program.* Washington, DC: Author.

U.S. Department of Education (2002). *Strategic plan for 2002-2007,* Retrieved January 20, 2004, from http://www.ed.gov/about/reports/strat/plan2002-07/index.html.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4,* 212-213.

Wright, D.B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology, 73,* 123-136.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.