PsycCRITIQUES 
October 12, 2004 Vol. 49, Supplement 3 .

# Paradigm Shift in Social Science Research
# A Significance Testing and Effect Size Estimation Rapprochement?

**A Review of**

Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research, by Rex B. Kline, Washington, DC: American Psychological Association, 2004. 336 pp. ISBN 1–59147–118–4.

Reviewed by
**Thomas D. Parsons**
**Nathaniel W. Nelson**

---

During the past 40 years, the vast majority of students trained in the social sciences have been taught the "received view" of null-hypothesis significance-testing (NHST). Social science researchers following the received review have approached their work with a heavy reliance on statistical tests, which have been strongly identified with probability or "significance" testing. The process of "rejecting" or "retaining" the null hypothesis has dictated this research, leaving the hope of accepting an alternative, explanatory hypothesis to the mercies of somewhat arbitrary designations of chance probabilities (e.g., $p < .05$). However, over the years a growing number of researchers have argued against the received view of statistical inference (e.g., Carver, 1978; Cohen, 1994; Hunter, 1997; Loftus, 1996). In fact, the American Psychological Association's (APA) Board of Scientific Affairs has recently considered prohibiting the reporting of significance tests in APA journals (Shrout, 1997). However, there are strong supporters of NHST, who state that critiques of NHST tend to be metatheoretical assertions that do not take into account the fact that rejection of $H_0$ is only one factor in the practice of testing a theoretical hypothesis (Abelson, 1997; Chow, 1996; Cortina, & Dunlap, 1997; Frick, 1996).

Rex B. Kline, in his recent book *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*, takes issue with the received view and foresees a future in which the NHST in social science research will diminish. Kline is calling for a "paradigm shift," in which social science will more closely reflect the hard sciences. Kline concisely explicates his vision:

Most studies in the future will not use statistical tests as the primary decision criterion, and those that do will concern only very specific problems for which variations of NHST may be appropriate, such as equivalence testing or inferential confidence intervals. It is also envisioned that the social sciences will become more like the natural sciences. That is, one will report the directions and magnitudes of one's effects; determine whether they

replicate; and evaluate them for their theoretical, clinical, or practical significance, not just their statistical significance (p. 15).

Matters raised here are essential and engaging. After identifying and examining the central theses of Kline's book, we intend to estimate the accomplishment of his overall position.

In Part 1: Introductory Concepts (Chapters 1–3), Kline summarizes proposed limitations of NHST found in literature reviews of multiple disciplines. He argues that critiques of NHST give reason for change in data-analytic praxes. He also suggests that the increase in meta-analyses further attests to the increasing demand for alternatives to traditional significance testing. Additional support for the notion that the milieu of research in behavioral science is changing may be found in the APA-appointed Task Force on Statistical Inference report (APA, 1996) and the emphasis on alternative research methods in the fifth edition of the *Publication Manual of the American Psychological Association* (2001). Consistent with these changes, Kline recommends that readers recognize that a statistically significant result is not explicitly revealing and a statistically nonsignificant result should not be readily discounted. Further, he calls for a pervasive reporting of effect sizes and the construction of confidence intervals whenever possible. Next, Kline reviews sampling issues in comparative studies (e.g., balanced vs. unbalanced samples) and related estimations (e.g., point and interval estimation, confidence intervals). He also provides a review of basic parametric statistics (*t*, *F*, and $?^2$) and the assumptions that underlie these statistics.

Part 1 also includes numerous critiques of NHST, including appraisals of the interpretation of *p* values and false conclusions that follow rejecting or retaining the null hypothesis. Kline examines some of the ways in which researchers do not meet NHST assumptions and the limitations of NHST itself, such as its overconcern for groups rather than individuals. Kline concludes that if researchers must use NHST, they should do so sparingly, such as "in very exploratory research where it is unknown whether effects exist" (p. 86).

In Part 2: Effect Size Estimation in Comparative Studies (Chapters 4–7), Kline endeavors to supply readers with the skills necessary to perform effect size estimation and interval estimation for each of the following: parametric effect size indexes, nonparametric effect size indexes, effect sizes in one-way designs, and effect size estimation in multifactor designs. Throughout Part 2, Kline emphasizes the uniqueness of designs for each of these respective types of estimation, the contexts for use and limitations of effect size estimations, the role of confidence intervals, error rates, and research examples.

In Part 3: Alternatives to Statistical Tests (Chapters 8 and 9), Kline suggests bootstrapping and Bayesian statistics as possible supplementary alternatives to NHST. Here, Kline is concerned with replication and argues that it is critical to social science research. Different types of replication are presented, including the various types of internal and external replication. Accordingly, Kline emphasizes meta-analyses as advantageous for both qualitative and quantitative research synthesis. He also contends

that researchers should make use of resampling techniques because they are forms of internal replication. He describes three such methods, including bootstrapping, jackknife, and randomization. Finally, Kline offers a very clear, though somewhat abbreviated, description of Bayesian estimation. He presents the general principles of Bayesian analysis and its rationale, and then gives examples with discrete hypotheses and continuous random variables.

## Analysis of Kline's Argument

Kline's *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research* offers a comprehensive, up-to-date, and insightful overview of the current literature on alternatives to statistical significance testing. His firm grasps of both statistics and social science research makes this an exceptional book that is important reading for researchers, teachers, and practitioners in the field of social science research. Further, this work is accessible to readers, regardless of experience and is indispensable for those interested in the significance testing debate.

A primary goal of this book is to argue that NHST is an inadequate tool for advancement of social sciences research. Although the lack of convenient alternatives to significance testing have historically forced social science researchers to endure the inadequacies of NHST, Kline hopes to make available viable alternatives. For example, the researcher may make use of effect magnitude measures, replication measures, and meta-analyses. Kline's goals are as follows: (a) offer social science researchers the skills necessary to perform effect size estimation and interval estimation for effect sizes; (b) suggest bootstrapping and Bayesian statistics as possible supplementary alternatives to NHST; and (c) show that critiques of NHST give reason for alteration in data-analytic praxes.

Kline does an exceptional job of providing readers with an elucidation of the methods necessary for the praxes of effect size estimation and interval estimation for effect sizes. Further, he supplies the reader with additional tools for improving statistical praxes by including bootstrapping and Bayesian statistics as possible supplementary alternatives to NHST.

Unfortunately, Kline does not spend a great deal of time responding to claims, radical as he may feel that they are, made by NHST supporters who call into question the assumed significance of the effect size. NHST supporters maintain that effect size is not an index of the evidential support for the substantive hypothesis offered by the data, and that effect size alone cannot be the index of the practicable significance of the research result in the case of the functional experiment (Chow, 1996). Further, some significance testing supporters make the controversial claim that power analysis is limited in that, being a conditional probability; it cannot be the probability of obtaining statistical significance. Further, according to said supporters, power analysis has a Bayesian flavor, which makes power analysis subject to the critiques against Bayesian assumptions about research methodology. A more fully developed response to these claims would have been helpful to Kline's overall argument contra significance testing.

Kline argues that critiques of NHST give reason for the minimization or elimination of NHST from data-analytic praxes. According to Kline, "These shortcomings are so serious that it is recommended that the continued use of statistical tests as the primary inference tool in behavioral sciences is not acceptable" (p. 44). Yet, Kline does little to answer the significance testing supporters' claim that many of the problems inherent in current NHST are relative to researchers confusing the null-hypothesis statistical testing procedure (NHSTP) with theory corroboration. Rejection of $H_0$ is only one factor in the practice of testing a theoretical hypothesis. These supporters argue that the NHSTP is better viewed as proffering an objective methodology that may remove chance influences from researchers' description of data. Hence, the NHSTP is argued by significance testing supporters to be a statistical decision that supplies a minor premise that will be used for inductive decisions relative to experimental design.

Kline's argument would have been aided by a response to arguments for NHST as a statistical procedure (NHSTP) used for theory collaboration, and not a theory (NHST) in itself. Although Kline is correct in stating that NHST is lacking when size of effect is important, NHST can be argued to be appropriate for examining "ordinal" assertions linking the order of conditions. Further, although Kline is correct in his argument that NHST is inadequate for determining epistemic justification, it is appropriate for support of ordinal claims (Frick, 1996). Statistical significance advocates remind critics that the soundness of research decisions is relative to the quality of the research design in general and the research questions in particular. Although Kline may argue that statistics and research design are two mutually dependent areas, NHST supporters contend that one should not criticize the NHSTP (statistical measure) for a failure to supply answers to questions of research design (theory support). Again, rejection of $H_0$ is only one factor in the practice of testing a theoretical hypothesis. Significance testing advocates have presented a number of important distinctions existing among conceptual, research, and statistical hypotheses. For example, Chow (1996) makes a distinction between theory corroboration and testing practical utility. Unfortunately, in Kline's analysis there is a dearth of analysis of arguments that emphasize the theoretical precepts and underpinnings that support the use of the NHSTP for theory corroboration.

It is important to note that Kline does do a nice job of incorporating theory into his formulation of replication and meta-analyses. Kline makes use of ideas gleaned from Kuhn's (1996) paradigms and the progression of normal science, as well as Hedges's (1987) "theoretical cumulativeness." For Kline, these ideas help explain the ways in which data analytic methods that move beyond NHST extend earlier work and expand our knowledge. However, Kline does not address the arguments made by significance testing supporters who claim that the NHSTP is something used for theory collaboration. NHST advocates defend their acceptance of the NHSTP with principles from Karl Popper's (1962/1968) falsificationism and hypothetical-deductive perspective (Chow, 1990). For the NHSTP, the goal is to remove chance as a confound within an inductive program of experimental investigations. As a result of these theoretical formulations, the NHSTP advocate may differentiate between inductive logic and statistics. Further, they are able to present a more robust agenda for the use of significance testing as one aspect of "theoretical cumulativeness."

## In Summary

Kline's *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*, offers an insightful discussion of the critiques of statistical significance testing. Kline has written a clear and comprehensive text that will be an invaluable guide for both students and researchers. This work offers an engaging, informative, and up-to-date discussion of both the historical development and current practice of significance testing. We agree with Kline that the theory and praxes of NHST by many researchers today are inadequate for the advancement of the social sciences. Further, Kline's clear and careful provision of practicable alternatives to significance testing will aid social science researchers to move beyond mere application of the NHSTP. Kline has successfully made available the information necessary to perform effect size estimation and interval estimation for effect sizes. He has also done an excellent job of illuminating and arguing for bootstrapping and Bayesian statistics as possible supplementary alternatives to NHST. However, we felt that his argument that critiques of NHST give reason for alteration in data-analytic praxes, would have been aided by a thorough discussion of the claims made by supporters of significance testing on which the soundness of research decisions is argued to be relative to the quality of the research design in general and the research questions in particular. Although Kline may argue that statistics and research design are two mutually dependent areas, NHST supporters contend that one should not criticize the NHSTP (statistical measure) for a failure to supply answers to questions of research design (theory support). Again, rejection of $H_0$ is only one factor in the practice of testing a theoretical hypothesis. Hence, we feel that Kline's overall argument would have been stronger had he answered the significance testing supporters' claim that many of the problems inherent in current NHST are relative to researchers confusing the null-hypothesis statistical testing procedure (NHSTP) with theory corroboration.

In general, we are in agreement with Kline that a paradigm shift appears imminent. The ubiquity of controversy surrounding NHST reflects Kuhn's contention that the emergence of new scientific ideas requires a decision process that allows researchers to disagree. According to Kuhn the existence of differential preferences and values among researchers allows new theories to develop. Advocates of NHST and effect size estimation often subscribe to different methodological standards and have nonidentical sets of cognitive values. Nevertheless, we are not sure that NHST and effect size estimation are radically incommensurable. For example, it does not appear to be impossible for translation to occur between these rival methodologies.

In our opinion the NHSTP should be viewed as one of the many tools available to the researcher within a multistage research process. We do not believe that the incorporation of the NHTP into our data analytic arsenal results in what Kuhn would call the "underdetermination of data," in which the rules or evaluative criteria of social science research should ascribe to one theory uniquely or unambiguously (effect size estimation) to the exclusion of all its competitors (significance testing). Instead, we feel that the effective use of research design, deductive and inductive logic, as well as the insurance of adequate experimental controls will furnish the researcher with essential information from which the researcher may make inferences and develop knowledge. Although we

feel that the NHSTP means little more than a decision that chance influences may be ruled out as an explanation of the data, we do not agree with critics' claims that it should be banned from the journals. One reason for this is that the provisional nature of the logic-based conclusions of an experiment are not made less tentative by the use of effect sizes, confidence intervals, power analysis, or meta-analysis. Hence, we feel that greater emphasis should be placed on both increased rigor of experimental design, as well as the inclusion of NHSTP, effect size indices, confidence intervals, meta-analysis, and power analysis. Although we agree with Kline that a paradigm shift appears imminent, we do not believe that it will necessarily result in the exclusion of NHST. Instead, we envision a significance testing and effect size estimation rapprochement that takes into account the benefits and limitations of each methodology.

## References


Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12-15.

American Psychological Association, Board of Scientific Affairs (1996, December). *Task Force on Statistical Inference report.* Retrieved August 18, 2004, from http://www.apa.org/science/tfsi.html

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

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

Chow, S. L.. In defense of Popperian falsification. *Psychological Inquiry, 1,* 147-149. l990

Chow, S. L. (1996). *Statistical significance: Rationale, validity, and utility.* Thousand Oaks, CA: Sage.

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

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

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

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42,* 443-455.

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

Kuhn, T. (1996). *The structure of scientific revolutions* (3rd ed.) Chicago: University of Chicago Press.

Loftus, G. R. (1996). Psychology will be a much better science when the authors change the way the authors analyze data. *Current Directions in Psychological Science, 5,* 161-171.

Popper, K. R. (1968). *Conjectures and refutations: The growth of scientific knowledge.* New York: Harper & Row. (Original work published 1962)

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8,* 1-2.