CE Online

THE CONTROVERSY OF SIGNIFICANCE TESTING: MISCONCEPTIONS AND ALTERNATIVES

By Dale N. Glaser, PhD. From Pacific Science & Engineering Group, San Diego, Calif.

The current debate about the merits of null hypothesis significance testing, even though provocative, is not particularly novel. The significance testing approach has had defenders and opponents for decades, especially within the social sciences, where reliance on the use of significance testing has historically been heavy. The primary concerns have been (1) the misuse of significance testing, (2) the misinterpretation of P values, and (3) the lack of accompanying statistics, such as effect sizes and confidence intervals, that would provide a broader picture into the researcher's data analysis and interpretation. This article presents the current thinking, both in favor and against, on significance testing, the virtually unanimous support for reporting effect sizes alongside P values, and the overall implications for practice and application. (American Journal of Critical Care. 1999;8:291-296)

In the history of science, specifically since 1900, many researchers have resorted to tests of significance when testing hypotheses. Karl Pearson laid the foundation for significance testing as early as 1901. Ronald Fisher, as well as the team of Jerzy Neyman and Egon Pearson, formulated the practice of null hypothesis significance testing, involving such crucial components as type I and type II errors, as well as power.[1] Many refinements of significance testing occurred during the 1950s (for a brief history of significance testing, see Gigerenzer[2] and Huberty[3]). Even Fisher made clear that the *P* value of .05 was at best a convention, with many advising against the rigid adherence to ".05" (some rebukes cropped up as early as the late 1930s). For the most part, however, the echoes of alarm have been summarily dismissed or minimized.[2,4] Debates about the misapplication of significance testing have again arisen within the social sciences, with the psychological sciences particularly sounding the alarm. Part of this increased debate may be due to the long-lasting reliance on and misuse of significance testing within the behavioral and educational sciences, in conjunction with the documented litany of complaints about significance testing.[4] In brief, the controversy involves the sole use (and misinterpretation) of the *P* value without taking into account other descriptive statistics, such as effect sizes and confidence intervals, statistics that provide a broader glimpse into the data analysis. Hence, a primary objective of this article is to summarize the controversy surrounding the use of significance testing, solutions that are being proffered, and what implications the suggestions have for applied researchers in the health sciences. This article is geared for both consumers of clinical research and applied researchers; thus, statistical arguments are kept to a minimum. Even though significance testing has many critics, it still has a place for clinical researchers so long as it is used judiciously.

CE Online

To receive CE credit for this article, visit the American Association of Critical-Care Nurses' (AACN) Web site at http://www.aacn.org and click on "Earn CEs" from the main menu, or call AACN's Fax On Demand at (800) 222-6329 and request item No.1120.

Reprint requests: InnoVision Communications, 101 Columbia, Aliso Viejo, CA 92656. Phone, (800) 899-1712 or (949) 362-2050 (ext 515); fax, (949) 362-2049; e-mail, ivcReprint@aol.com.

A Brief History

As we know from Kuhn,[5] paradigmatic changes in science take place slowly, especially when the change involves modifying or eliminating a practice that has been firmly entrenched, such as significance testing. In an oft-cited paper from 1966, Bakan[6(p423)] asserted that "the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; and that, furthermore, a great deal of mischief has been associated with its use." In 1960, Rozeboom[7] vigorously attacked the faulty inferences used in the service of significance testing and in 1997 exhorted: "Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students."[8(p336)] In even more terse terms, Frank Schmidt and John Hunter, two leading proponents of meta-analysis, have claimed that the use of significance testing actually retards the ongoing development of the research enterprise.[9,10]

Especially instrumental in the resurgence of interest in the topic of the misuse of significance testing was the 1994 article "The Earth Is Round ($p<.05$)" by Jacob Cohen,[4] which appeared in the *American Psychologist.* Some readers will recognize Cohen as the author of a text on power analysis[11] that is a mainstay of many researchers when determining sample sizes for their studies. In the *American Psychologist* article, Cohen cited errors that many researchers were still committing in the service of significance testing; most had to do with the flawed interpretation of the *P* value. Partially because of that provocative article, as well as Schmidt's address at the 1994 American Psychological Association (APA), debates on the misapplication of significance testing were reinvigorated, with many articles and convention forums focusing on this topic.[12] Because of the controversy, the APA Board of Scientific Affairs appointed the Task Force on Statistical Inference, comprising some of the most influential researchers in psychology, statistics, and the social sciences. The task force first met in December 1996, and their initial report can be located at the following Web site: http://www.apa.org/science/tfsi.html. To the relief of many, the task force did not recommend the abolishment of significance testing, but rather advised that supplementary tools (eg, effect sizes, confidence intervals) should be used with the same frequency that researchers currently use significance testing.[13] Following in the footsteps of the task force, the 1997 text *What If There Were No Significance Tests?* accelerated the discussion (both for and against significance testing).[14]

For now, the preceding summary provides a brief glimpse of the events that have led up to the resurgence of interest in significance testing. The rest of the article explores and discusses some of the uses and misuses of significance testing, alternative solutions, and what implications these factors have for the clinical researcher.

Defining the *P* Value

Regardless of the statistical technique used (eg, parametric vs nonparametric) or type of design (eg, experimental vs quasi-experimental), *P* values are often reported in clinical research. The *P* value (the applied researcher will recognize this as "Sig." in most statistical printouts) is compared with the a priori alpha level (α), which serves as the basis for rejecting or failing to reject the null hypothesis. However, misinterpretations of the *P* value and what it entails are not unusual. Part of this misinterpretation may stem from the lack of uniformity across texts in the definition of the *P* value. Even though the exact terminology may differ,[15-18] one generally agreed upon definition of the *P* value is that we are ultimately testing the null hypothesis against a level of significance (α) designated by the researcher.

Significance and Importance

Indoctrination of certain statistical principles occurs early in the training of clinical researchers. An example in statistics is the frequent exhortation that "correlation does not imply causality" (for an interesting discussion of correlation and causality in the context of structural equation modeling, see Bullock et al[19]). Another principle that has been part of the statistical training of applied researchers is that "statistical significance does not mean the same as importance." Even though many authors have clearly differentiated statistical significance from practical significance, it is still not unusual for researchers to herald the *P* value as being synonymous with importance.[20-23] Bruce Thompson,[23(p39)] editor of the influential *Educational and Psychological Measurement,* states in regard to the issue of practical versus statistical importance: "Improbable events are not intrinsically interesting. Some highly improbable events, in fact, are completely inconsequential." The following example illustrates this problematic misconstruing of significance testing.

A dietary change tested in a randomized clinical trial results in a mean decrease in total cholesterol level from 243 mg/dL to 238 mg/dL in the experimental group and no change in the control group. Is

this mean difference of 5 mg/dL enough to modify clinical practice? Most clinicians would judge the difference inconsequential. But what if the difference is statistically significant (ie, $P<.05$)? Many might be tempted to view this difference as more important. If the pooled standard deviation of this example was 25 and the groups had 400 patients each, a difference of 5 mg/dL would indeed be statistically significant.

At this juncture, it becomes critical to disentangle the phenomenon of statistical significance (determined in part on the basis of the a priori α level) from clinical significance. If obtaining significance (ie, rejecting the null hypothesis) is the predominant objective of the researcher, then diminishing the probability of type II error (ie, failing to reject the null hypothesis when it is false, or saying no difference exists when a difference may exist) is warranted.[24] Hence, many clinical researchers are savvy enough to do a power analysis before the start of the study. Given a postulated effect size, directionality of effect (ie, 1-tailed vs 2-tailed), α level, and desired power (power of a statistical test being the probability of yielding statistically significant results), a sample size for the study at hand can be determined.[11] What is clearly evident, as illustrated in the cholesterol example, is that, given a large enough sample size, it becomes increasingly easy to reject the null hypothesis. But rejecting the null hypothesis is not synonymous with import of effect. Hence, there has been a strong movement of late (even though Cohen has been discussing the issue of examining effect sizes for years) to report not only P values but also effect sizes.[4]

Effect sizes, which can be calculated for virtually any type of statistical analysis, assist researchers in arriving at a measure of magnitude or strength.[25] In the cholesterol example, if the absolute value of the difference between the means is divided by the pooled standard deviation (5/25), I arrive at Cohen's d, an effect size that is expressed in units of a standard deviation. So, the "significant" finding in the cholesterol analysis reveals an effect size that is a standardized score of only .20. To put it in context, a small d is considered to be .20, according to Cohen's guidelines of small, medium, and large effect sizes. However, and as Cohen warns, these effect sizes should be used only as guidelines and not as blind convention, because sometimes small effects may indeed be clinically important (eg, decreased mortality).[11] Reporting the effect size, in conjunction with the result of significance testing, provides a broader perspective as to what the data mean.[26] It is not enough to state that "a significant and positive correlation was obtained, $P=.001$"; perchance, this resultant P value may be associated with $r=0.20$, accounting for just 4% of the variation. Now, 4% may be impressive in some research contexts, especially if the manipulation of the independent variable was weak or the researcher was investigating unexplored terrain, but it still leaves 96% of the variation unexplained.[27] Hence, the onus should be placed on researchers to report effect sizes with the same regularity as they report P values.[28]

P as an Indicator of Strength

In those cases in which the clinical researcher obtains a P value that is about .05 (eg, $P=.064$), he or she may choose to suspend judgment, invoking the paraphrased witticism: "Surely, God loves .064 as much as .049!!" The researcher makes a predetermined decision about the level of significance (α) necessary to reject the null hypothesis. This α level (generally set at .05) is also the type I error (ie, probability of wrongly rejecting the null hypothesis when it is true—saying a difference exists when it does not).[16,24] It is essential to keep in mind that significance testing is inherently a dichotomous process: the ultimate decision is to reject the null hypothesis or to fail to reject the null hypothesis.

When criticisms are put aside, significance testing still has an important role in the sciences for making ordinal claims (reject vs fail to reject).[29] Significance testing "provid[es] us with the criteria by which provisionally to distinguish results due to chance variation from results that represent systematic effects in data available to us."[20(p81)] However, what is most egregious is when the researcher interprets the P value in 1 of the 2 following ways: (1) "the obtained results ($P=.064$) approached significance" or "the results were marginally significant" and (2) "the obtained P value of .00001 indicates that the results were very significant." Both of those depictions are fallacious and misleading. Regarding the first interpretation, Thompson[21] provides an example of an apt retort: "How did you know your results were not trying to avoid being statistically significant?" In an effort to understand (or embellish) their findings, researchers turn to P values as indices of importance, a logic that is deeply flawed. It is specious to suggest that a study with a P value of .001 is "more important" than a study with a resultant P value of .049.[20] P values say nothing about magnitude or import. Those types of apple-and-orange contrasts lead researchers and their audience to draw conclusions about a magnitude of effect that may be mathematically infinitesimal.

Many editors require that authors provide exact P

values in the results section, but the process of significance testing is dichotomous: you either reject or fail to reject the null hypothesis, given your sample data, when comparing the resultant P value with the nominal α level. So does the P value of .064 in the first interpretation warrant the author's vote of support for "marginal significance," given the inherently dichotomous nature of significance testing? It is true that with a larger sample size (or a stronger manipulation—an issue generally not factored in when power analysis is done), the obtained P might be less than .05. So, instead of claiming marginal significance on an a posteriori basis, researchers should report the effect size and let the results carry their own weight. Unfortunately, and as many researchers are painfully aware, null results rarely see the light of day in the published literature, a phenomenon that has been referred to as the "file drawer problem."[30] Thus, the incentive is great to claim marginal significance in a study that is void of "significant" effects; again, this situation is one in which the publishing of effect sizes can be most beneficial.

Now, what about the case of the ubiquitous asterisks? In the following scenario, the researcher sets the α at a predetermined level (generally at .05), obtains a P value of .001, and then, with unintentional trickery, uses asterisks: $*P<.05$; $**P<.01$; $***P<.001$. This practice goads researchers into conclusions that their findings are "more significant" and hence "more powerful" than the findings really are, a conclusion that is deceptive and misleading. Cronbach and Snow,[31] as cited in Pedhazur and Schmelkin,[16(p201)] assert: "A p value reached by classical methods is not a summary of the data. Nor does the p attached to a result tell how strong or dependable the particular result is. . . . Writers and readers are all too likely to read .05 as p (H:E), 'the probability that the Hypothesis is true given the Evidence.' As textbooks on statistics reiterate almost in vain, p is p (E:H), the probability that this Evidence would arise if the [null] Hypothesis is true." Cohen called this type of confounding the inverse probability error.[4] It is incumbent that researchers realize the dubious practice, as well as faulty logic, of ascribing strength or magnitude to a P value. Unfortunately, the requirement of many journals to report asterisked P values, absent of effect sizes, perpetuates the myth that the P value denotes importance.

The Null and the Nil

The language of significance testing, when introduced at the undergraduate level, can indeed be forbidding, if not somewhat contrary. When we reject the null hypothesis, we discourage language such as "prove" or "confirm," when referring to the alternative hypothesis. Given methodological artifacts such as sample specificity and measurement error, in the spirit of the Popperian notion of falsification, we can only hope to disconfirm the null hypothesis.[16] We use double negatives when we assert that no difference exists—that is, we fail to reject the null hypothesis—or we claim that the results are "inconclusive."[32] We do not encourage the student to "accept" the null hypothesis in the case of a decision to fail to reject the null hypothesis, for failing to reject the null hypothesis does not prove that the null hypothesis is true.[20] Even though the intrepid student may see this wordplay as a barometer of the researcher's unwillingness to take a theoretical stand, this line of reasoning, tenuous as it may appear, characterizes the scientific pursuit. Claiming a significant result does not prove the veracity or the durability of the theory. In fact, one misconception often associated with significance testing is that $1 - P$ denotes the probability that the findings will be replicated in future studies. This misconception has been termed the replicability fallacy.[16,20] Every study is unique, with all its attendant idiosyncrasies: sample characteristics, environmental variables, and other factors that may contribute to error variance. Hence, rejecting the null hypothesis ($P<.05$) for 1 study, even though that rejection provides a guidepost to future research, does not prove anything.

Furthermore, failing to reject the null hypothesis ($P>.05$) does not provide corroborating evidence for the nonexistence of the phenomenon. As stated earlier, failing to reject the null hypothesis does not prove the null hypothesis. For any given study, countless factors may have culminated in a fail-to-reject decision: faulty manipulation of the independent variable, insufficient power, unreliable measures, and threats to internal validity. Accepting the null hypothesis has led many a researcher to conclude that the test statistic equals the parameter, that is, the sample mean minus the population mean equals zero. Part of this logic may be attributed to the default settings in many of the popular statistical software programs and to the discussion of significance testing in introductory statistics texts. For example, the test of significance for the Pearson product moment correlation coefficient (r) is compared with that of the population correlation (ρ), with the latter assumed to be zero. However, how realistic is this point estimate of zero? If a researcher were to correlate nursing satisfaction and job performance, and on the basis of meta-analytic findings the average

effect size is $r=0.32$, does comparing this sample estimate to a population parameter of $\rho = 0$ seem at all logical? A more defensible tactic might be to compare the sample estimate against a nonzero effect size—an effect size that makes sense given the context of the study. Cohen, influenced by Meehl, has had provocative discussions about the problematic application of testing a point estimate against a parameter of zero.[4,20,33,34] This mode of significance testing has been termed the nil hypothesis, referring to the postulate that an effect size for a given null hypothesis is zero. The argument is that there will always be some difference between an estimate and a parameter (or difference between 2 samples in the context of a *t* test, in which the null hypothesis revolves around a zero difference), so failing to reject the null hypothesis (or by software settings, the nil hypothesis) does not confirm that the parameter equals the hypothesized value. In fact, some extreme factions have asserted that the null hypothesis is always false, at least to some decimal point.[20] Even though this line of reasoning can reach the depths of reductio ad absurdum, the prevailing message is that we do not accept the null hypothesis or prove the null hypothesis. Some difference will always be present; only by investigating the effect sizes can researchers answer the question, How much difference and does it matter?

Suggestions and Alternatives

Positioning oneself as the naysayer is easy; it does not take great facility to degrade practices that have been ingrained in the mainstream. The larger challenge is to provide suitable alternatives that can withstand reasonable debate and be instituted either alongside or in place of hitherto conventional practices.

As a first recommendation, and without exception, effect sizes should be reported by clinical researchers with the same frequency that *P* values are reported. As a matter of practice, all pertinent descriptive statistics should be reported (eg, means, standard deviations, and any graphical displays such as box plots). Even the *Publication Manual of the American Psychological Association,*[35(p18)] acknowledging the integral role of effect sizes, encourages researchers to "provide effect-size information." Whereas some editors have maintained rather emphatic points of view about the reporting of effect sizes, with Bruce Thompson, editor of *Psychological and Educational Measurement,* requiring such reporting, the frequency of reporting effect sizes remains abysmally low.[17,36,37] Perhaps instead of "encouragement," the APA should "require" the reporting of effect sizes.[36,37] Commensurate with this philosophy, the editor of the *Journal of Applied Psychology,* Kevin Murphy,[38(p4)] has stipulated the following: "If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against reporting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit." The preceding guideline did receive approval from the APA Publications Committee.[37]

A second recommendation is that the reporting of point estimates and confidence intervals should be a regular practice.[4,15,36] Confidence intervals, which provide a margin of error around a sample estimate, not only are helpful for assessing the extent of error around the estimate but also reveal the same information that significance testing yields.[4] As Cohen has implied, however, the lack of reporting of confidence intervals may be due to the existence of the "embarrassingly" large intervals that transpire in many researchers' results.[4] As for point estimates, it may serve researchers well to consider testing the null hypothesis against an estimate that is logically defensible (eg, a nonzero correlation between nurses' satisfaction and performance), as opposed to the default nil (ie, zero effect) hypothesis.

Another option, even though relatively ambitious in application, is the use of meta-analysis. Meta-analysis integrates findings across multiple studies (eg, studies focusing on nursing interventions), with the effect size of the individual study generally being the unit of analysis. Frank Schmidt and John Hunter are longtime proponents of the use of meta-analysis, largely motivated by their concern about the deficient power associated with most individual studies.[9,10,39] If one considers their assertion that the average individual study has a power of .50, hence, a probability of type II error of .50, then one must be astounded that 50% of studies are not sufficiently sensitive to the experimental or correlational effect of interest. If that is indeed the plight of most individual studies, then all these fail-to-reject decisions (ie, nonsignificant findings) will end up in Rosenthal's metaphorical "file drawer," never to see the light of day in published journals.[30] Meta-analysts have argued that the strength in their technique is that the process of compiling numerous studies, across various samples and settings, offsets the deficient power associated with most individual studies. This affords researchers the opportunity to draw more general conclusions that would, by necessity, be precluded in the context of any single study.[40,41] Conversely, some criticisms are

associated with meta-analysis (eg, only studies with nonnull findings are considered, possibly resulting in overestimation of the true effect size), so meta-analytic findings should be used as a guide rather than a data-driven panacea. Use of meta-analysis entails a potentially inordinate extension of resources that may not be realistic for clinical researchers.

Further, given the high number of underpowered studies, clinical researchers must pay more attention to the power of their studies. As frequently exhorted in many texts, the most sophisticated statistical machinations cannot salvage a flawed design. It behooves researchers not only to consider sample size in their effort to enhance the power of their study but also to evaluate the manipulation of their independent variable.

Despite the rumblings and ominous overtones of the proposed banning of the significance test, a more temperate solution has been offered by a wide array of researchers.[42,43] Significance testing will always have its advocates and opponents. At this time, however, more than any other, researchers are considering the import of instituting effect sizes, confidence intervals, and power analysis alongside the traditional mode of significance testing. The recommendation is not that clinical researchers disavow significance testing, but rather that they incorporate additional information that will supplement their findings.

REFERENCES

1. Harlow LL. Significance testing introduction and overview. In: Harlow LL, Mulaik SA, Steiger JH, eds. *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum; 1997:1-17.
2. Gigerenzer G. The superego, the ego, and the id in statistical reasoning. In: Keren G, Lewis C, eds. *A Handbook for Data Analysis in the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum; 1993:311-339.
3. Huberty CJ. Historical origins of statistical testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks. *J Exp Educ.* 1993;61:317-333.
4. Cohen J. The earth is round ($p < .05$). *Am Psychol.* 1994;49:997-1003.
5. Kuhn TS. *The Structure of Scientific Revolutions.* 2nd ed. Chicago, Ill: University of Chicago Press; 1970.
6. Bakan D. The test of significance in psychological research. *Psychol Bull.* 1966;66:423-437.
7. Rozeboom WW. The fallacy of the null-hypothesis significance test. *Psychol Bull.* 1960;57:416-428.
8. Rozeboom WW. Good science is not abductive, but hypothetico-deductive. In: Harlow LL, Mulaik SA, Steiger JH, eds. *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum; 1997:335-391.
9. Schmidt FL, Hunter JE. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: Harlow LL, Mulaik SA, Steiger JH, eds. *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum; 1997:37-64.
10. Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol Methods.* 1996;1:115-129.
11. Cohen JC. *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum; 1988.
12. Schmidt F. Board of scientific affairs action on significance testing. *Ind Organ Psychol.* 1996;33:110-111.
13. Task Force on Statistical Inference. Initial report—draft. Available at: http://www.apa.org/science/tfsi.html. Accessed January 10, 1997.
14. Harlow LL, Mulaik SA, Steiger JH, eds. *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum; 1997.
15. Reichardt CS, Gollob HF. When confidence intervals should be used instead of statistical tests, and vice versa. In: Harlow LL, Mulaik SA, Steiger JH, eds. *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum: 1997:259-284.
16. Pedhazur EJ, Schmelkin LP. *Measurement, Design, and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum; 1991.
17. Thompson B. AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educ Res.* 1996;25:26-30.
18. Moore DS. *The Basic Practice of Statistics.* New York, NY: WH Freeman & Co; 1995.
19. Bullock HE, Harlow LL, Mulaik SA. Causation issues in structural equation modeling research. *Struct Equation Modeling.* 1994;1:253-267.
20. Mulaik SA, Raju NS, Harshman RA. There is a time and place for significance testing. In: Harlow LL, Mulaik SA, Steiger JH, eds. *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum; 1997:65-115.
21. Thompson B. The concept of statistical significance testing. *Meas Update.* 1994;4:5-6.
22. Stevens J. *Applied Multivariate Statistics for the Social Sciences.* 3rd ed. Mahwah, NJ: Lawrence Erlbaum; 1996.
23. Thompson B. Five methodology errors in educational research: the pantheon of statistical significance and other faux pas. Paper presented at: Annual Meeting of the American Educational Research Association; April 15, 1998; San Diego, Calif.
24. Keppel G. *Design and Analysis: A Researcher's Handbook.* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1982.
25. Sawyer AG, Ball AD. Statistical power and effect size in marketing research. *J Mark Res.* 1981;18:275-290.
26. Chow SL. Significance test or effect size? *Psychol Bull.* 1988;103:105-110.
27. Prentice DA, Miller DT. When small effects are impressive. *Psychol Bull.* 1992;112:160-164.
28. Kirk RE. Practical significance: a concept whose time has come. *Educ Psychol Meas.* 1996;56:746-759.
29. Frick RW. The appropriate use of null hypothesis testing. *Psychol Methods.* 1996;1:379-390.
30. Rosenthal R. The "file-drawer problem" and tolerance for null results. *Psychol Bull.* 1979;86:638-641.
31. Cronbach LJ, Snow RE. *Aptitudes and Instructional Methods: A Handbook for Research on Interactions.* New York, NY: Irvington; 1977.
32. Aron A, Aron EN. *Statistics for Psychology.* Upper Saddle River, NJ: Prentice-Hall; 1994.
33. Cohen J. Things I have learned (so far). *Am Psychol.* 1990;45:1304-1312.
34. Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J Consult Clin Psychol.* 1978;46:806-834.
35. American Psychological Association. *Publication Manual of the American Psychological Association.* 4th ed. Washington, DC: American Psychological Association; 1994.
36. Thompson B. If statistical significance tests are broken/misused, what practices should supplement or replace them? Paper presented at: Annual Meeting of the American Psychological Association; August 17, 1997; San Diego, Calif.
37. Thompson B. AERA debate on use of statistical significance tests: comments for the royal rumble. Paper presented at: Annual Meeting of the American Educational Research Association; April 15, 1998; San Diego, Calif.
38. Murphy K. Editorial. *J Appl Psychol.* 1997;82:3-5.
39. Schmidt FL. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *Am Psychol.* 1992;47:1173-1181.
40. Glass GV, McGaw B, Smith ML. *Meta-Analysis in Social Research.* Beverly Hills, Calif: Sage; 1981.
41. Petitti DB. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis.* New York, NY: Oxford University Press; 1994.
42. Shrout PE. Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychol Sci.* 1997;8:1-2.
43. Chow SL. *Statistical Significance: Rationale, Validity, and Utility.* Thousand Oaks, Calif: Sage Publications Inc; 1996.