Feature Article

Methodological Issues Related to the Use of *P*<0.05 in Health Behavior Research

Elias Duryea, Stephen P. Graner, and Jeremy Becker

ABSTRACT

This paper reviews methodological issues related to the use of P<0.05 in health behavior research and suggests how application and presentation of statistical significance may be improved. Assessment of sample size and P<0.05, the file drawer problem, the Law of Large Numbers and the statistical significance arguments in epidemiology, health behavior, and psychology were examined. The reporting of confidence intervals (CI), effect sizes (ES), and use of non-statistical graphics can improve portrayal and understanding of findings. Health behavior literature has had some scholarly examination of how to improve analysis of findings but has not had an in-depth dialog on other concepts related to P<0.05. Attention to these concepts could improve clarity in how research outcomes are presented and thereby increase credibility of health behavior research.

Duryea E, Graner SP, Becker J. Methodological issues related to the use of P < 0.05 in health behavior research. Am J Health Educ. *2009;40(2):120-125. This paper was submitted to the Journal on April 23, 2008, revised and accepted for publication on November 24, 2008.*

INTRODUCTION

In the mid-1980s, the *American Journal of Public Health* called for a moratorium on significance tests in favor of confidence intervals for all journal submissions. The moratorium did not endure.

The quantitative testing of posed hypotheses has been central to scientific inquiry. Whereas various fields of science, including medicine and public health, have assessed findings using statistical methods for almost a century, health behavior researchers have had comparatively less time utilizing *P*<0.05 to assess outcomes. Because health behavior literature is much newer than that of its predecessors in psychology and medicine, the absence of dialog on the role and interpretation of statistical significance testing in our literature is not surprising. Recent publications from Buhi,[1] Watkins and associates,[2] and Zhang and colleagues[3] have been much needed and scientifically sound contributions in this area. Buhi[1] proposed that health education researchers increase the calculation and incorporation of effect sizes (ES) into their studies to assist meta-analysts who synthesize results from hundreds of studies. Researchers should consider the context of their study including the researchers' own judgment of the importance of a result. Finally, Buhi[1] recommends appropriately that researchers provide evidence of the results i.e., replicability, by either cross-validation or external replications with new samples and settings.

Watkins and associates make similar recommendations and call for researchers to use the guidelines set forth by Vacha-Haase and Thompson[4 (p. 107)] and the APA Task Force on Statistical Significance for reporting ES:

- Express what specific effect sizes are being reported
- Include confidence intervals (CI) for effect sizes, and
- Interpret the ES with regard to analytical assumptions and related limitations.

They also call for journals to provide instructions for calculating ES, CIs for them and graphics from the calculated CI estimates.

Elias Duryea is a professor in the Department of Health & Exercise Sciences, University of New Mexico, Albuquerque, NM 87131; E-mail: Duryea@unm.edu. Stephen P. Graner is in the Department of Emergency Medicine, Kapi'Olani Medical Center, Honolulu, HI. Jeremy Becker is in the Department of Orthopedics, Presbyterian Hospital, Albuquerque, NM.

Zhang and colleagues[3] address this last recommendation by Watkins et al.[2] by presenting clear and applicable techniques for how researchers may evaluate and facilitate the use of CIs in health education studies. They present software that can be used in calculating CIs and ways for investigators to interpret each CI accurately. Each of these reports has added to the growing momentum in the health education research field to build that Zhang et al.[3(p. 35)] describe as the "building blocks of meta-analytic thinking" among its researchers.

The convention of setting the Type I error rate (i.e., alpha) at < 5%, or stated differently, at 1 in 20 chance of rejecting the null hypothesis given that it actually is true, has a fascinating, yet problematic history.[5] The fields of psychology, medicine, and epidemiology for the past three decades have debated the use of statistical significance testing as a standard for assessing results.[6] At the center of this debate has been the contention that scientists have routinely not only misapplied, but also misinterpreted statistical significance tests.[7,8] The class paper by Cohen[9(p. 997)] stated *it is nearly universal that alpha is misinterpreted to mean the probability the null hypothesis is false.*

This report highlights some of the classic arguments for and against the use and interpretation of statistical significance testing in quantitative research in general, and specifically, within the context of health behavior research. The issues of sample size and the *Law of Large Numbers*, the "file drawer problem," and the academic tendency to publish mostly "significant" results are also discussed. Finally, there is examination of how interpretation of results may be improved when $P<0.05$ is presented in conjunction with other estimates such as effect sizes (ES), confidence intervals (CI) and raw graphic displays.

To set a context for this discussion, it should be noted that critics of using $P<0.05$ have proposed that such procedures be excluded *completely* from journal articles and grant applications.[10,11] Conversely, proponents of $P<0.05$ argue that $P<0.05$ is just one of many useful pieces of data that can be used to assess statistical significance and that without such cut-offs scientists may drift toward subjectivity, and thus, weaken their public credibility.[12,13]

Regardless of how $P<0.05$ is viewed, critics have suggested that its use and interpretation is consistently misguided. To understand these concerns, the classic definition and central purpose of "*P*" are examined.

Classic Definition and Purpose of P <

Thompson[14(p. 191)] proposes that "*there are few certainties in the conduct of epidemiologic research. One thing that can be counted upon, however, is that chance will play some role in the pattern of results.*" The possibility that findings resulted from "chance" is not a welcome thought to researchers. In contemporary graduate-level research texts, students are taught that if the difference between experimental and control group outcomes is significant at $P<0.05$, then the likelihood that this difference is due to chance is less than five percent.[15,16] Students are instructed to recognize $P<0.05$ as alpha or the probability of making a Type I error, i.e., rejecting a true null hypothesis.[17]

Regardless of language, the concept of null hypothesis significance testing is a mainstay in much of science.[18] Glantz[19(p. 108)] defines the *P* level as "*the probability of being wrong when concluding that a true difference exists between groups.*" Since $P<0.05$ has been so entrenched, or as Mckinlay and Marceau[20] refer to it - "sanctified," from years in research literature, most researchers feel a sense of accomplishment if their test statistics reach $P<0.05$.

Investigators have raised numerous troubling, and ultimately divisive questions, about $P<0.05$ being used as the cut-off for designating supposedly important differences.

Traditionally, scientists have set the alpha before the experiment as a dichotomous point from which results are either designated significant or not. Historically, not all scholars have agreed that such a firm and inflexible cut-off point should be used in statistical testing.[21] Some researchers have proposed that statistical significance estimates actually be viewed without rigid borders.[22] Arguments have been made, for example, that the meaningful difference between a $P=0.049$ and a $P=0.060$ is virtually negligible.

Beyond the debate of whether $P<0.05$ should be a firmly adhered to cut-off point or whether it should be used as a flexible guide, a common misconception exists that statistical significance equates with importance of the result. As will be shown, this is neither true conceptually, nor operationally. The major purpose of setting $P<0.05$ or 0.01 *a priori* is to convey to readers what predetermined criterion is being used to define statistical significance. This also allows the researcher to present the probability that a significant finding could be due to chance.[22] One situation where results do not reach the $P<0.05$ criterion and a journal decides not to allow publication has been referred to as the "file drawer problem."

The Ethical Dilemma of the File-Drawer Problem

Many research studies with non-significant findings, i.e., $P>0.05$, are never published in nationally recognized journals. It is not known if journals deliberately refuse to publish such manuscripts, or if authors decide not to pursue such outlets because their results are not statistically significant. Results from such studies are sometimes disseminated in alternative or open source venues referred to as "gray literature."[23] This publication trend is what many term the "file-drawer" problem in research.[24] Essentially, it refers to the dilemma that emerges when a researcher perceives that the likelihood of getting a manuscript accepted is low due to non-significant findings and thus alternative venues for publication are sought. Failure to reach statistical significance may result from a wide array of factors: poor theory, use of instruments with poor psychometric properties, poor study implementation, and inadequate sample size. It should also be noted that failure to find a statistically significant effect between groups may occur simply because there *was* no effect.[25] It can be argued, however, that rigor in theory, methodology, and contribution to the field should be paramount; but

Table 1. By Increasing N, the Same Result Eventually Produces a Statistically Significant P-Value

Study 1: N = 30
Outcome = 4 point drop in blood pressure, P = 0.27 (no statistical significance)

Study 2: N = 60
Outcome = 4 point drop in blood pressure, P = 0.12 (no statistical significance)

Study 3: N = 120
Outcome = 4 point drop in blood pressure, P = 0.03 (statistically significant)

also that excluding research simply because $P<0.05$ is not reached be avoided. Certainly, numerous advances in science have resulted from publishing and disseminating results that did not immediately attain $P<0.05$. If a study's theory, rigor, and methodology is sound, but $P<0.05$ is not found, perhaps readers should be informed of this fact so they can know what has "not" worked.

Related to this dilemma is the manner in which research teams describe non-significant findings. Thompson[14] has claimed that researchers have tried to subtly align results with significance despite results failing to meet the 0.05 criterion by describing the result as "*approaching*" significance or having achieved "*marginal*" significance. Because the tradition of setting alpha at <5% *a priori* has been a dichotomous one, proponents of $P<0.05$ are adamant that results cannot "*approach*" or be "*partially*" significant. Moreover, they also cannot be "*in the expected direction of*" significance. Such rules leave little room for flexibility with regard to how the researcher interprets quantitative results for their study.

The Law of Large Numbers

One area where researcher discretion in designing the study *is* more flexible rests in sample selection and size. It is here that a volatile argument has raged for years among scientists.[26] Whereas health behavior literature has not examined this issue in-depth, as a younger member of the scientific community, these concepts warrant discussion. To understand this argument, we must describe the historical foundation for the relationship between N and probability of results, or what is historically known as the Law of Large Numbers.

In 1713, Jakob Bernoulli published the Law of Large Numbers (LLN). This statistical rule has come to be known as the Law of Averages and states that as an experiment is repeated over and over, the observed probability of a result will get closer and closer to the actual or true probability.[27] Stated more plainly, as sample size (N) increases so does the probability of getting a $P<0.05$. The LLN was used historically to refer to the concept that even rare events occur when a sufficiently large number of observations are conducted. If one did not know the probability of rain, one could estimate that probability by making a sufficiently large number of observations over an extended period of time. This mathematical rule eventually helped scientists produce measurement theory and psychometrics.[28] Its relationship to the concept of reliability, for example, is readily evident: with increased numbers (i.e., items, observations) random error is reduced and reliability of measurement is enhanced. The relationship between N and $P<0.05$ was perhaps best captured by Berkson in 1938 when he stated "*if, then, we know in advance the p that will result from an application of a statistical test to a large sample, there would seem to be no use in doing it on a smaller one.*"[29 (p. 527)] Similarly, Maxwell and Delaney[30 (p. 96)] five decades later stated that "*the sample size problem is that, for any difference from the null, and for any level of significance, however small, the p value can be made as vanishingly small as desired by increasing the sample size.*" Plainly, as N increases so does the probability (P) that the difference between what is observed and what is true, will be small. The conceptual core for this phenomenon is the Law of Large Numbers.

Modern research texts provide tables that show at what N a given correlation coefficient will be significant. The same applies to quasi-experiements and true experiments, where statistical significance can ultimately be achieved if the sample size becomes large enough. Goodman[31] described a study (Table 1) where an investigator set $P<0.05$ and then tested a new drug for high blood pressure on one group of 30 patients. The first trial registered a four-point reduction ($P=0.27$). When the experiment was repeated with a new N = 60, the same four points were obtained and the P reduced to $P=0.12$.

At this point, a colleague mentions that the investigator has already "used up" the allotted alpha of 0.05 in the first trial. Undeterred, the investigator runs a third experiment with a doubling of N to 120 new participants. Again, the same four-point improvement results, but now the P value = 0.03 and $P<0.05$ is achieved. The author then submits the last trial to a reputable journal omitting discussion of the first two experiments. Such statistical "fishing" is clearly unethical and related to the need to obtain a sample size that provides enough power to detect specific differences when they exist. Moreover, such practice is probably also related to the pressure to publish, and ultimately, avoid having one's research end up in the proverbial file drawer.

Predictably, all three trials had the same finding but at increasing N produced different and descending P values. In the end, the same colleague questions the author on whether a four-point result is even meaningful. Some investigators propose that in many situations meaningfulness, as opposed to significance, may be the more relevant issue.[32] For example, a large intervention reports that a modest drop in a national mortality rate, even if not statistically significant, might be an important and meaningful

finding.[33] As Torabi [34] and others in health behavior research literature have stated, if the outcome lacks meaningfulness to practitioners, what value is significance?

Cautions on Significance Testing

Over the last 50 years the fields of psychology, medicine, and other sciences have debated whether using statistical significance levels was even warranted for assessing results of research.[35] Much of the debate does not even specify whether the 0.05 or some other level was the issue. Rather the core of the arguments focused on how misused and misinterpreted statistically significant analyses had become in various journals, conferences, and texts.[36]

Does achieving results with $P<0.05$ and presenting such outcomes as important to various constituencies reflect conceptual and possibly ethical misuse? A recent report proposed that health behavior investigators take care in how they describe results.[37] The suggestion was that erroneous causal language (i.e., the curriculum *caused* the better scores in the treatment group) be avoided so as not to mislead readers on what the results truly meant. Watkins and associates[2] also have called for clearer and more precise language in how outcomes are presented, specifically with regard to assumptions and limitations that accompany ES and CIs. Some researchers may well misconstrue $P<0.05$ as implying cause and effect in findings. Two factors can, for example, be statistically associated but certainly not causal. A critical question, thus, emerges for researchers: has the use of $P<0.05$, its interpretation and communication to constituencies created a false impressions regarding what results mean? One strategy to help answer this question is to clarify the functions of calculating significance levels.

What Significance Testing Does Not Do

Epidemiology is not the only field with professional differences on this issue. More recently, educational psychologists, communication, and other researchers have argued against the use of significance testing.[38] Central to these arguments is the following: (1) $P<0.05$ does *not* mean clinical/practical meaningfulness - just because a finding was $P<0.05$ does not mean that the result was meaningful to the recipients or had utility for the researchers; (2) $P<0.05$ does *not* mean importance - what outcome is seen as important is many times subjective; and (3) $P<0.05$ does *not* equate with strength - a $P<0.01$ is not stronger than a $P<0.05$ only less likely to be false.[39] There are tests to quantify strength of the relationship, such as omega-squared [34] and ES, but as researchers have suggested these procedures have not been routinely reported in health behavior research.[1-3]

What Significance Testing Does Do

Fleiss[12] and other supporters of significance testing have argued that such tests have a number of functions for researchers, especially, in epidemiology. First, independent replications are known to be one valid way to either refute or support a study outcome. If the original study was significant at $P<0.05$ and repeated studies also produce the same conclusion, then the original study result gains credibility. If the same significance finding is not produced, the initial result becomes problematic and scrutiny on demographics, implementation (i.e., strength of the treatment), sampling, and measurement (i.e., psychometric qualities) is conducted.

Second, investigators regularly calculate statistical significance in order to identify possible confounding variables. In different types of studies many potential confounders may be evident. Researchers often use sophisticated, yet efficient, statistical programs (e.g., http://sciencessoftware.com) to isolate which confounders are independent from each other yet relate to the dependent variable. Significance tests, along with theory, permit investigators a guide on how to decide which factors to control in the study. Factors associated with the outcome at $P<0.05$ level and those critically linked by theory, are then controlled.

Relevance for Health Behavior Investigators

Are such issues of importance, and thus relevance, for the discipline of health behavior? We argue that all scientific discussions that improve the ability to design, conduct, and more clearly and accurately present research findings have merit. As recent reports across the country suggest, there are well-organized political attacks against science, including public health education.[40] Sponsors of these campaigns appear motivated by economics as well as philosophical and political values. Additionally, they routinely depict supposedly "controversial" science as inaccurate, unreliable, and even deceptive. Consistently and strategically they focus on published scientific findings that are susceptible because the researcher did not clearly describe them. Shermer[41(p. 32)] reported that a pharmaceutical corporation had stopped publishing data from clinical trials because they found such information confused the public and that "*consumers were not scientifically minded enough to understand.*"

If health behavior researchers have inadvertently added to this confusion by not clearly explaining our research evidence, then the discipline should examine options. One step in this direction is to critically discuss how significant $P<0.05$, as well as non-significant findings, are presented in our research literature.

Enhancment of P<: Effect Sizes, Confidence Intervals and Raw Graphics

The following options for enhancing $P<0.05$ in research reports exclude formula. Such equations are readily available in texts and in various software routinely employed by researchers.[3] In addition, recent power calculations, as accurately analyzed by Price and associates for survey studies, are also assumed to be crucial to improving health behavior research results.[42]

Effect Sizes (ES)

One important practice an academic community can employ to improve public understanding of results, is to clearly explain quantitative information. Health behavior research literature has made great strides in this area. To this aim, investigators should use all available appropriate procedures for evaluating, depicting, and explaining outcomes.

Cohen [9] has long recommended that ES

be calculated for all major experiments. Unlike *P* values, ES actually does estimate the strength of the association. Moreover, ES can be performed for virtually any type of statistical analysis. An ES reported along with the *P* value offers a more detailed and complete view of the meaning of the obtained result.

Whereas the *P* value allows one to estimate the probability that the effect could be due to chance, the ES allows one to quantify the strength of the relationship. As Cohen[9] notes, however, just as the p value must be assessed in context of its purpose, ES must be explained with the same due caution. A small ES of .20 (as per Cohen's guidelines) may be clinically important (e.g., decreased mortality). Editors of both the *Journal of Applied Psychology* and *Psychological and Educational Measurement* have called for ES to be part of any manuscripts submitted.[32] Vaughan[43] reiterated the need for ES, power, and "meaningfulness" in presenting health research to constituencies.

Confidence Intervals (CI)

A second data analytic enhancement is the calculation of confidence intervals (CI) along with point estimates. A 95%/99% CI corresponds to *P* levels of .05/.01, respectively. Such procedures allow readers to visualize the range of sampling error within which the point estimate (e.g., mean, quit rate) resides. Larger CIs correspond to less precision, and thus, more error surrounding the result. Poole[13] has proposed that the reason many investigators omit CIs is because the intervals are embarrassingly large. If means for treatment and comparison groups were, for example, 77 and 57, and the CIs were calculated to be 29, 99, 21, and 86 respectively, the amount of error surrounding the means make this result difficult to interpret. If the difference in means was, however, significant at $P<0.05$, then Poole's contention may be true - report the $P<0.05$ and omit the large CIs.

Thompson[14] has even proposed that the width of the CI gives a clear indication of just how uninformative a study result may actually be. Kocher[38] correctly stated that the CI, unlike the *P* value, can be presented in the units of the variable of interest. As Watkins and colleagues[2] have proposed, this helps readers better interpret the data and the range of sampling error. Additionally, if the CIs do not overlap for groups, it is evidence for significance because CIs will generally not overlap if a significant difference exists. Some researchers use $P<0.01$ instead of 0.05 to approximate the "non-overlapping CIs" method for determining statistical significance. Unlike *P* values that simply give the probability a favorable conclusion could be wrong, reporting point estimates and their corresponding CIs provide three additional pieces of information: (1) statistical significance shown by lack of overlap, (2) meaningful significance shown by the magnitude of the values, and (3) precision of measurement shown by the CIs' range.

Graphic Displays from Raw Data

Finally, health behavior researchers may consider an alternative strategy that proposes that research data be analyzed and then portrayed by nothing more than graphic depictions of group performance. Before and after frequency distributions with raw means and variances, according to McKinlay,[20] may be a more informative way for consumers to evaluate what the findings show. Inevitably, situations arise where N is small, randomization is not feasible, and data do not need complex statistical analyses.[20] Under such contexts, results could be more clearly portrayed using a descriptive and/or visual format.

DISCUSSION

Health behavior research literature can benefit from a constructive evaluation of its application and presentation of quantitative procedures. It is in this area that the field will ultimately generate clearer and more persuasive evidence documenting the social merit of its research. It is crucial to reiterate that the use of $P<0.05$ should *not* be abandoned as part of data analysis and presentation. To the contrary, researchers should retain an intellectual balance in examining the recommendations in this review. In a best case scenario, if outcomes consistently converge on the positive effects of new health behavior innovations (e.g., syringe exchange programs, comprehensive sexuality curricula), then the use of *multiple* analytic approaches to verify and clearly portray findings becomes crucial.

As a newer scientific discipline, health behavior has an opportunity to improve its professional image by taking steps to ensure that its research findings are presented with maximal clarity, accuracy, and meaningfulness. Simply reporting statistically significant outcomes may not accomplish this goal. As investigators in our discipline continue their research, they should take advantage of ways to present $P<0.05$ in conjunction with the previously presented options so clarity is fully optimized.

REFERENCES

1. Buhi ER. The insignificance of "significance" tests: Three recommendations for health education researchers. *Am J Health Educ.* 2005; 36(2): 109-112.

2. Watkins DC, Rivers D, Roswell KL, Green BL, Rivers B. A closer look at effect sizes and their relevance to health education. *Am J Health Educ.* 2006; 37(2): 103-108.

3. Zhang J, Hanik BW, Chaney BH. Confidence intervals: Evaluating and facilitating their use in health education research. *The Health Educator.* 2008; 40(1):29-36.

4. Vacha-Haase T, Thompson B. How to estimate and interpret various effect sizes. *J Couns Psych.* 2004;51(4): 473-481.

5. Tyler RW. What is statistical significance? Educ Res Bull. 1931;10:115-118.

6. 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.

7. Cox DR. Statistical significance tests. *Brit J Clin Pharm.* 1982; 14: 325-331.

8. Shrout P. Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychol and Sci.* 1997; 8:1-2.

9. Cohen J. (1994). The earth is round ($p < .05$). *Am Psychol.* 1994; 49:997-7003.

10. Carver R. The case against statistical

significance testing. *Harv Ed Rev*.1978; 48: 378-399.

11. Young MA. Supplementing tests of statistical significance: Variation accounted for. *J Speech Hear Res*. 1993; 36: 644-656.

12. Fleiss J. Significance tests have a role in epidemiologic research: Reactions to A.M. Walker. *Am J Public Health*. 1986; 76:559-561.

13. Poole C. Beyond the confidence interval. Am J Pub Health. 1987; 77:195-199.

14. Thompson W. Statistical criteria in the interpretation of epidemiologic data. *Am J Public Health*. 1987; 77: 191-195.

15. Yates F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *J Am Stat Assoc*. 1951;46: 19-34.

16. Moore DS. *The Basic Practice of Statistics*. New York: W.H. Freeman & Company 1995.

17. Bartz AE. *Basic Statistical Concepts*.(Third Ed.).Upper Saddle River: NJ. 1988:453-459.

18. Cowles M, Davis C. On the origins of the .05 level of statistical significance. *Am Psychol*. 1982; 37: 553-558.

19. Glantz S. (2002). *Primer of Biostatistics*. New York: McGraw-Hill 2002:106-110.

20. McKinlay JB, Marceau LB. (1999). A tale of 3 tails. *Am J Public Health*. 1999; 89: 295 – 298.

21. Walker HM. *Studies in the History of Statistical Method*. Baltimore, Md.: Williams & Wilkins 1929: 88-94.

22. Morrison DE, Henkel RE. Significance tests reconsidered. *Am Scientist*. 1969; 4:131-140.

23. Conn V, Valentine J, Cooper H. Grey literature in meta-analyses. *Nurs Res*. 2003;52:256-261.

24. Rosenthal R. The "file-drawer problem" and tolerance for null results. *Psychol Bull*. 1979;86:638-641.

25. Cook D, Guyatt G, Ryan G. Should unpublished data be included in meta- analyses? *JAMA*. 1993;269:749-753.

26. Savitz D. Interpreting epidemiological evidence - strategies for study design and analysis. Oxford: Oxford University Press 2003:248-251.

27. Huygens C. On reasoning in games. In Bernoulli J, Maseres F, (Eds.) *The Art of Conjecture*. New York: Redex Microprint, 1970. (Originally published 1657).

28. Rosnow R, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. *Am Psychol*. 1989; 44: 1276-1284.

29. Berkson J. Some difficulties of interpretation encountered in the application of the Chi Square Test. *J Am Stat Assoc*.1938; 33: 526-542.

30. Maxwell SE, Delaney HD. *Designing Experiments and Analyzing Data*. Belmont, CA.: Wadsworth Publishing 1990.

31. Goodman S, Royall R. Evidence and scientific research. *Am J Public Health*. 1988; 78:1568 -1574.

32. Glaser D. The controversy of significance testing: Misconceptions and alternatives. *Am J Critical Care*. 1999; 8:13-22.

33. 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.

34. Torabi M. How to estimate practical significance in health education research. *J Sch Health*. 1986; 56:232-234.

35. Rozeboom WW. The fallacy of the null-hypothesis significance test. *Psychol Bull*. 1960; 57: 416 – 428.

36. Fleiss J, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. Hoboken, NJ : John Wiley and Sons 2003.

37. Duryea EJ. Use of cause and effect language in health behavior research literature, *Am J Health Behav*. 2002; 26:221-228.

38. Kocher MS, Zurakowski D. (2004). Clinical epidemiology and biostatistics. A primer for orthopaedic surgeons. *J Bone Joint Surg*. 2004; 86: 607 – 620.

39. Rosenstock L, Lee LJ. Attacks on science: The risks to evidence-based policy. *Am J Public Health*. 2002; 92:14-18.

40. Fagin D, Lavelle M. *Toxic Deception: How the Chemical Industry Manipulates Science, Bends the Law, and Threatens Your Health*. Seacacus, NJ: Birch Lane Press, 1997.

41. Shermer M. Airborne baloney. *Scientific Am*. 2007; 32: 32-33.

42. Price JH, Dake JA, Murnan J, Dimmig J, Akpanudo S. Power analysis in survey research: Importance and use for health educators. *Am J Health Educ*. 2005; 36:202-207.

43. Vaughan RD. The importance of meaning. *Am J Pub Health*. 2007; 97:592-593.