5-1-2003

A Different Future For Social And Behavioral Science Research

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Journal of Modern Applied Statistical Methods
May, 2003, Vol. 2, No. 1, 128-132

1538 – 9472/03/$30.00

A Different Future For Social And Behavioral Science Research

Shlomo S. Sawilowsky
Educational Evaluation and Research
College of Education
Wayne State University

The dissemination of intervention and treatment outcomes as effect sizes bounded by confidence intervals in order to think meta-analytically was promoted in a recent article in *Educational Researcher*. I raise concerns with unfettered reporting of effect sizes, point out the con in *con*fidence interval, and caution against thinking meta-analytically. Instead, cataloging effect sizes is recommended for sample size estimation and power analysis to improve social and behavioral science research.

Key words: Effect size encyclopedia, bracketed interval, confidence interval, sample size, power

Introduction

Recently, an article appeared in *Educational Researcher* describing a possible future of social science research. It was one in which research results were reported in terms of effect sizes bounded by so-called confidence intervals. The notion of thinking meta-analytically was touted, and to that end, the publication of effect sizes was promoted (Thompson, 2002).

Bracketed Intervals (BI)

I prefer the phrase "bracketed interval" (BI) instead of confidence interval, for reasons discussed below. The Frequentist perspective of the BI was described by Thompson (2002) as a 95% degree of confidence that the interval contains the parameter in question. According to this view it would be inappropriate to say there is a 95% probability that μ, the population mean, is within the interval, but it would not be inappropriate to say there is a 95% level of confidence that μ is in the interval.

Shlomo S. Sawilowsky is Professor of Educational Evaluation and Research, and Wayne State University Distinguished Faculty Fellow. He is the editor of *Journal of Modern Applied Statistical Methods*. Email him at shlomo@wayne.edu.

The first intervals of a statistical nature were developed by de Moivre between 1733 - 1742, but they were not positioned for interval estimation. That feat was first accomplished by Lagrange in 1776.

De Moivre stated that the interval refers to "the probability that the value of [a parameter] is enclosed between the [upper and lower] limits" (cited by Hald, 1998, p. 23). Thus, in modern classification schemes, the original expression of bracketed intervals was from a Frequentist perspective.

Now, return to the term *confidence*. The general idea originated with Pytkowski (1932), but the first use of the phrase *confidence interval* and its theoretical development was by Neyman (1934, 1937, 1939). He referred to

> determining certain intervals, which I propose to call the confidence intervals (see Note 1), in which we may assume are contained the values of the estimated characters of the population, the probability of an error in a statement of this sort being equal to or less than $1 - \epsilon$, where ϵ is any number $0 < \epsilon < 1$, chosen in advance. The number ϵ I call the confidence coefficient. (1934, p. 562)

He opined that "the solution of the problem which I described as the problem of confidence intervals has been sought by the greatest minds since the work of Bayes 150 years ago" (Neyman, 1934, p. 563). However, because Jerzy Neyman, along with Egon Sharpe Pearson, originated the Frequentist version of modern statistics (Neyman & Pearson, 1928a, 1928b), his definition was purposefully not "Bayesian", and instead followed the Frequentist paradigm.

The student of Bayes would demur, claiming it doesn't make sense to ascribe the 95% moniker to μ being found within the interval. The 1-α% probability only pertains prior to the collection of data, whereas afterwards either the parameter falls within the interval or it doesn't.

Instead, the Bayesian perspective is that the judicious usage of specific prior information regarding the estimate is the only meaningful way to obtain such a probability. Thompson (2002) characterized this as "a better definition" (p. 26).

The weakness of the Bayesian approach, (which Fisher, Neyman, Wald, and others rejected) is the reliance on subjective prior information. I cannot resolve the philosophical debate between the Frequentist and the Bayesian, but it is inappropriate to call either perspective "better", as did (Thompson, 2002, p. 26).

Furthermore, the philosophical controversy Thompson (2002) alluded to is not relevant in practical application. What is of importance is the role of interval estimation vs hypothesis tests. There has been a flurry of activity since the early 1990s where the usage of hypothesis tests was taken to task, particularly within the American Educational Research Association (AERA) and other professional organizations. For example, Carver (1992) presented a paper to the AERA attempting to make a case against statistical significance testing, and recommended banning its usage altogether.

Amazingly and inexplicably, proponents of the case against hypothesis testing are also proponents of the usage of interval estimation. The root of their misconception is the misnomer *confidence*, as if bracketed intervals have a certain amount of confidence to them that hypothesis tests do not. There is no more confidence associated with an interval based on $(1-\alpha)100\%$ than in a point null hypothesis based on α.

Thompson (2002) incorrectly construed my position in *Educational Researcher*, claiming I "erroneously equate CIs and statistical significance tests" (p. 29). In an article with Thomas Knapp, I pointed out that the statistical criteria regarding the probabilities associated with bracketed intervals are the same as those for point null hypothesis tests, but certainly the two procedures cannot be equated. Regarding the equivalency of probabilities: (1) Is zero really not in the interval? (Type I error), (2) is zero really in the interval? (Type II error), and (3) is the width of the interval at a minimum (comparative statistical power)? The probabilities associated with these criteria are exactly the same (Knapp & Sawilowsky, 2001).

These three points are congruent with a careful examination of Neyman (1934). He equated the boundaries of the interval with the probabilities of classical Fisherian "fiducial" limits of $\theta_1(x)$ and $\theta_2(x)$, which represent the lower and upper bound of the bracketed interval. With a passing reference to the famous debate in the literature on what Sir Ronald Fisher meant by fiducial, Neyman (1934) did not dissociate the so-called confidence of the bracketed interval from the probabilities used in its construction:

> Since the word "fiducial" has... caused misunderstandings I have already referred to, and which in reality cannot be distinguished from the ordinary concept of probability, I prefer to avoid the term and call the intervals $[\theta_1(x), \theta_2(x)]$ the confidence intervals. (p. 590)

Although Wald (1950) subsumed both hypothesis tests and interval estimation in a single model, and expressed them as specific cases of the general theory of statistical decision functions, that does not mean the two procedures are equivalent in every respect. After pointing out the probabilities associated with BIs and hypothesis tests are the same, I noted there is an advantage of BIs over point null hypothesis tests. It results in a range of possible values wherein the parameter might fall, whereas hypothesis tests do not.

This doesn't appear to be the tremendous advantage that many proponents claim it to be. What added benefit is there in knowing, for

example, that the BI for a student's *Wechsler* IQ was 97-103 from a educator's perspective? Furthermore, in Knapp and Sawilowsky (2001), we mentioned specific data analysis situations where the BI would be preferred over the hypothesis test, as well as the reverse.

I also pointed out there are areas of concern in unbridled promotion of BIs (Knapp & Sawilowsky, 2001): (1) Some statistics are not amenable to the determination of standard errors, relying instead on theoretically interesting but practically questionable asymptotic variances (which are mathematical inventions pertaining to the world of infinite sample sizes). This may make the BI yield poorer statistical properties than point hypothesis testing. (2) There is the question of whether or not the interval should be symmetric about the sample statistic (Low, 1997).

(3) There is the problem of the effects of measurement error in constructing the interval (Nunnally, 1978). (4) Here, I add yet another concern: Bienaymé's complaint in 1852 against using BIs based on a single parameter expressed as a continuum on a line. Instead, he proposed the concept of Bracketed Ellipsoids, where simultaneous regions are constructed taking into account multiple parameters. For example, two parameters result in an ellipsoid continuum on a Cartesian plane.

Meta-Analysis

These issues regarding BIs apply to all statistics, including effect sizes. Thompson (2002) focused on effect sizes to provide fodder for meta-analyses. This became necessary following Gene Glass' presidential address on meta-analysis to the AERA in April of 1976, because modern meta-analysis depends on the proliferation of effect sizes.

Thompson (2002) viewed effect sizes as the enabler in thinking meta-analytically. His exuberance with meta-analysis led him to recommend that effect sizes "can and should be reported and interpreted in all studies, regardless of whether or not statistical tests are reported" (Thompson, 1996, p. 29), and "even [for] non-statistically significant effects" (Thompson, 1999, p. 67). The same argument had previously been made by Carver (1979, 1993).

However, Sawilowsky and Yoon (2001, 2002) reported a brief Monte Carlo simulation demonstrating the trouble with reporting research findings via effect size in the absence of statistical significance. The practice will wreak havoc in the literature, as the Monte Carlo simulation demonstrated that an intervention of random numbers will produce typical effect sizes that are not near zero, but rather, are at a magnitude Cohen (1988) calls a small treatment effect.

Roberts and Henson (2002) purported to rebut these results. However, their study was not a Monte Carlo simulation of typical effect sizes produced under the truth of the null hypothesis. Instead, it was a Monte Carlo study of the bias in d, a topic irrelevant to the point being made. See the ensuing *Invited Debate* in this issue of the *Journal of Modern Applied Statistical Methods*.

There have been many articles published here and there by a variety of authors, including myself, that addressed specific methodological and substantive issues with meta-analyses. In addition, I have raised questions about thinking meta-analytically (e.g., Knapp & Sawilowsky, 2001). Rather than reviewing that literature here, I find it more instructive to recite an excerpt from Glass' (2000) most recent vision of research synthesis:

> In the twenty-five years between the first appearance of the word "meta-analysis" in print and today, there have been several attempts to modify the approach, or advance alternatives to it, or extend the method to reach auxiliary issues. If I may be so cruel, few of efforts have added much... If our efforts to research and improve education are to prosper, meta-analysis will have to be replaced by more useful and more accurate ways of synthesizing research findings.

Sample Size Estimation and Power Analysis

The role of effect sizes in sample size determination and power analysis is an entirely different matter from that of meta-analysis. The first part of my professorial career could be summarized by the many consultations I had with students, teachers, faculty, and researchers outside of academe on the "how large should my sample be?" question. The bottleneck was obtaining an

estimate of the effect size, which is necessary to enter Cohen's (1988) sample size and power tables. I was not alone; every colleague I discussed this matter with in the past twenty years has reported the same difficulty.

I wrestled with this problem for a decade. During that time I had a series of written and telephone conversations with, and initiated by, Jacob Cohen. He recognized the weaknesses in educated guessing (Cohen, 1988, p. 12) or using his rules of thumb for small, medium, and large effect sizes (p. 532). I suggested cataloging and cross-referencing effect size information for sample size estimation and power analysis as a more deliberate alternative.

Cohen expressed keen interest in this project. His support led to me to delivering a paper at the annual meeting of the AERA on the topic of a possible encyclopedia of effect sizes for education and psychology (Sawilowsky, 1996). The idea was to create something like the "physician's desk reference", but instead of medicines, the publication would be based on effect sizes. (I presented papers every year at AERA from 1985 - 2000, but this session had a higher attendance than most of them put together.) I doubt any of those listening to the presentation envisioned a future for quantitative social and behavioral science research with sample size estimation and power analysis forever relegated to prestidigitation.

Encouraged by colleagues, in 1999 and again in 2000, I submitted proposals to the U. S. Department of Education to fund a print and electronic encyclopedia project. Thirty-five experts on effect sizes and meta-analysis wrote supportive letters (Table 1). A summit would be held with these experts, the most recent ten years of ninety journals in education and psychology would be culled for effect sizes and cataloged, and an internet-based data-base would be created in which authors/journal editors could submit additions or updates. Alas, the proposals were not judged to be a funding priority. Subsequently, I had a series of e-mail and telephone conversations with Herbert Walberg on creating the encyclopedia sans funding, but the enormity of the project was prohibitive.

Table 1. Supporters of the Encyclopedia of Effect Sizes Project:

William Asher, Purdue University
Betsy Becker, Michigan State University
John Behrens, Arizona State University
Patricia Busk, University of San Francisco
C. Mitchel Dayton, University of Maryland
Robert Donmoyer, Ohio State University
Susan Embretson, University of Kansas
Gene Glass, Arizona State University
Robert Grissom, San Francisco State University
John Hunter*, Michigan State University
Carl Huberty, University of Georgia
Harvey Keselman, University of Manitoba
John Kim, San Francisco State University
Roger Kirk, Baylor University
Thomas Knapp, Ohio State University
Dennis Leitner, Southern Illinois University
Joel Levin, University of Wisconsin-Madison
Lisa Lix, Private Scholar
Jorge Mendoza, University of Oklahoma
Theodore Micceri, University of South Florida
Isadore Newman, University of Akron
Steve Olejnik, University of Georgia
Liora Pedhazur-Schmelkin, Hofstra University
Bob Rosenthal, University of California-Riverside
Donald Rubin, Harvard University
Frank Schmidt, University of Iowa
Michael Seaman, University of South Carolina
Ronald Serlin, University of Wisconsin-Madison
Juliet Shaffer, University of California-Berkeley
Bruce Thompson, Texas A&M University
Howard Wainer, ETS
Herbert Walberg, University of Illinois-Chicago
Rand Wilcox, University of Southern California
Joe Wisenbaker, University of Georgia
Bruno Zumbo, University of N. British Columbia

Notes: *Deceased. Affiliations were accurate in 1999-2000.

Conclusion

Sample size estimation and power analysis in every grant funded by the U. S. Department of Education and every article published in AERA journals are based on guessing or Cohen's (1988) rules of thumb. Those practices could be discontinued in a different future of social and behavioral science research. Along with a re-commitment to true experimental design

(Sawilowsky, 1999), a compendium of effect sizes could improve research design in education and psychology, and propel disciplined inquiry forward in a scientific fashion.

The encyclopedia could be a globally cooperative effort among professional organizations and learned societies, their journal editors, and authors. It could be internet-based and updated in real-time, cross-referenced by discipline/sub-discipline and independent variable, have effect size entries categorized by statistically significant studies at various α levels, and classified according to whether the journal was peer reviewed. Finally, entries should be categorized based on whether the effect size arose from a true experimental design vs. quasi-experimental, post hoc, survey, and other non-experimental designs.

References

Carver, R. P. (1979). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Carver, R. P. (April, 1992). *The case against statistical significance hypothesis testing, revisited*. Annual Meeting of the American Educational Research Association, San Francisco, CA.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Hillsdale, NJ: Erlbaum.

Glass, G. (2000). *Meta-analysis at 25*. http://glass.ed.asu.edu/gene/papers/meta25.html.

Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. NY: Wiley & Sons.

Knapp, T., & Sawilowsky, S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education*, *70*, 65-79.

Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics*, *25*, 2547-2554.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method, *Journal of the Royal Statistical Society*, *97*, 558-625.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London*, A(*236*), 333-380.

Neyman, J. (1939). L'estimation statistique traitée comme un problème classique de probabilité. *Actualités Scientifiques et Industrielles*, *739*, 25-57.

Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part 1. *Biometrika*, *20*, 175-240.

Neyman, J., & Pearson, E. S. (1928b). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, A(*231*), 289-337.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed). NY: McGraw-Hill.

Pytkowski, W. (1932). Outline of the income in small farms upon their area, the outlay and the capital invested in cows. Monograph #31, *Biblioteka Pulawska*, Poland: *Agricultural Research Institute of Pulaway*.

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, *62*, 241-253.

Sawilowsky, S. S. (April, 1996). *Encyclopedia of educational and psychological effect sizes*. Annual Meeting of the American Educational Research Association, Division D, Measurement and Research Methodology, NY, NY.

Sawilowsky, S. S. (1999). Quasi-experimental design: The legacy of Campbell and Stanley. In (Bruno D. Zumbo, Ed.) *Social indicators/ quality of life research methods: Methodological developments and issues, Yearbook 1999*. Norwell, MA: Kluwer Academic Publishers.

Sawilowsky, S. S., & Yoon, J. (2001). *The trouble with trivials (p > .05)*. 53rd Session of the International Statistical Institute, Seoul, South Korea.

Sawilowsky, S. S., & Yoon, J. (2002). The trouble with trivials (p > .05). *Journal of Modern Applied Statistical Methods*, *1*, 143-144.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26-30.

Thompson, B. (1999). Five methodology errors in educational research: A pantheon of statistical significance and other faux pas. In B. Thompson (Ed.), *Advances in social science methodology*, *5*, 23-86.

Thompson, B. (April, 2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, p. 25-32.

Wald, A. (1950). *Statistical decision functions*. NY: Wiley.