5-1-2003

You Think You've Got Trivials?

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

Invited Debate: Target Article
You Think You've Got Trivials?

Shlomo S. Sawilowsky
Educational Evaluation & Research
Wayne State University University

Effect sizes are important for power analysis and meta-analysis. This has led to a debate on reporting effect sizes for studies that are not statistically significant. Contrary and supportive evidence has been offered on the basis of Monte Carlo methods. In this article, clarifications are given regarding what should be simulated to determine the possible effects of piecemeal publishing trivial effect sizes.

Key words: Trivial effect sizes, meta-analysis, Monte Carlo, simulation, Monte Carlo simulation

Introduction

"It would seem that power analysis has arrived" (Cohen, 1988, p. xiii). This was the conclusion of the late Jacob Cohen in reviewing twenty-six years of the literature since he brought the importance of effect size (and sample size) to the attention of behavioral and social science researchers (Cohen, 1962). The explosion of meta-analyses being published, which followed Gene Glass' presidential address to the American Educational Research Association (AERA) in April of 1976, also depends on the proliferation of effect sizes.

Researchers and editors, after neglecting power analyses in the past, or to provide raw materials for future meta-analyses, are now being asked to report effect sizes associated with statistically non-significant results. A recent motivating example of this call was made by Thompson (1996, 1999), who recommended effect sizes "can and should be reported and interpreted in all studies, regardless of whether or not statistical tests are reported" (1996, p. 29), and "even [for] non-statistically significant effects" (1999, p. 67).

Shlomo S. Sawilowsky is Professor of Educational Evaluation and Research (EER), College of Education, Wayne State University, Detroit, MI. He is the program coordinator of (EER), and Wayne State University Distinguished Faculty Fellow. Email: shlomo@wayne.edu. The title of this article is based on Gerrold (1973).

Robinson and Levin (1997; see also Levin & Robinson, 1999) gave a reasoned approach to the reporting of effect sizes. On the basis of a thought experiment, they concluded that it is better to "First convince us that a finding is not due to chance, and only then, assess how impressive it is" (p. 23). Knapp and Sawilowsky (2001) added additional heuristic arguments against the practice.

Sawilowsky and Yoon (2001, 2002) conducted a Monte Carlo simulation to provide rigor for this position. Their results indicated that "effect sizes should not be reported or interpreted in the absence of statistical significance" (Sawilowsky & Yoon, 2002, p. 144). In contrast, Roberts and Henson's (2002) Monte Carlo study came to the opposite conclusion. The purpose of this paper is to bring resolution to these opposing results.

High Quality Monte Carlo Simulation & Sampling With Replacement

It is necessary to preface with a brief discussion of (a) simulation, (b) Monte Carlo, (c) Monte Carlo simulation, (d) sampling *with* vs *without* replacement, and (e) characteristics of a high quality Monte Carlo simulation. This will clarify the study conducted by Sawilowsky and Yoon (2001, 2002), and explicate the flaws in the design and conclusion of the study conducted by Roberts and Henson (2002). It will also serve as a brief review of Monte Carlo simulation methods. (For more complete coverage of the Monte Carlo simulation method, see Sawilowsky & Fahoome, 2003).

Simulation

A simulation "mimics important elements" (Roberts, et. al, 1983, p. xi) of a system or phenomenon. It is "a representation ...in simplified form to study its behavior" (p. 452). Negoita and Ralescu (1987) noted that "In science... 'simulation' is forming an abstract model from a real situation in order to understand the impact of modifications and the effect of introducing various" (p. 29) interventions.

Norlén (1975) stated that simulation can be viewed as a "numerical technique for the carrying out of experiments" (p. 15). As an example, consider simulating the tossing of a fair die. This may be accomplished by accessing an uniform pseudo-random number generator that produces a value on the interval [0,1]. Draw a variate from the generator. Suppose it is .1770 (rounding to four significant digits, or to as many significant digits as desired). Using the assignment in Table 1 below, this process results in the simulation of throwing a fair die and having two spots surface.

Table 1. Simulation of a fair die using uniform variates on the interval [0,1].

Outcome	Assignment
.0000 - .1666	1 spot
.1667 - .3333	2 spots
.3334 - .5000	3 spots
.5001 - .6666	4 spots
.6667 - .8333	5 spots
.8334 - 1.000	6 spots

Monte Carlo

Monte Carlo, in the sense it is being used in this article, is of rather recent origin (Metropolis & Ulam, 1949). Its usage appeared over a half century ago in reference to the gaming establishments of previous centuries of a famous city in the Monaco principality. It is an explicit reference to the use of *repetition* as a method of discovery of the long run outcome of an event.

More technically, it is the "use of stochastic techniques to solve... a deterministic problem" (Moshman, 1967, p. 250). As such, "one of the simplest and most direct applications of the Monte Carlo methods is to the evaluation of integrals" (Kahn, 1966, p. 249-250), or the area of any geometric figure, but particularly those *irregular* in shape. (The first moment of the uniform distribution over the interval [0,1] can be obtained via the calculus:

$$\int_0^1 x dx = .5 .$$

This result could be estimated via Monte Carlo methods by drawing a large number of variates from a uniform pseudo-random number generator and computing the mean, but *usually* there is little point in doing so.)

As an example, consider the problem of determining the area of an irregular closed figure that is unwieldy to the calculus. Inscribe the figure within a unit square. Draw two variates from the uniform pseudo-random number generator to represent Cartesian coordinates for the ordered pair (x, y), and plot them accordingly. Repeat the previous step many times. The area of the irregular geometric figure is estimated (as accurately as desired) by the ratio of the number of dots that fall within the figure, divided by the total number of repetitions (i. e., pairs of dots created). Note, however, that no system or phenomenon was simulated.

A famous example of the Monte Carlo method was undertaken in 1908 by William Sealy Gosset (Student, 1908a, 1908b), a chemist working for the Guinness brewing company. He bolstered his analytical expression of the distribution of the Pearson product-moment correlation coefficient on small samples via a Monte Carlo conducted by hand. Similarly, he supported the derivation of the t statistic with a Monte Carlo demonstration of the sampling distribution of t.

Monte Carlo Simulation

Statistical historians (e.g., Hald, 1998, p. 196 - 201) noted that multinomial outcomes, such as tossing a fair die with equiprobability of one through six spots surfacing, was determined mathematically by Laplace in 1774. As an

alternative to the mathematical approach, the Monte Carlo simulation approach arose with Buffon in 1777, who tossed a coin 2,048 times and recorded the results. The distribution of outcomes indicated an expectation of heads to occur in 50.693% of the tosses. In 1837, Poisson determined $0.48468 < p < 0.52918$ to be what he called the 99.555% interval of the probability "p" representing the chance of a heads occurring.

A famous Monte Carlo simulation was reported in 1900 by the eugenicist, Karl Pearson. His zoologist colleague and co-founder of *Biometrika*, Walter Frank Raphael Weldon, tossed twelve dice at the same time, recorded the results, and repeated the process 26,306 times. Pearson (1900) procured this data set and applied his newly developed goodness of fit χ^2 test to demonstrate the frequency of obtained outcomes were as expected due to combinatorial analysis.

Norlén noted (1975) "the advent and use of computers... freed the method from manual calculations... and... afford richer possibilities for the creation of complex, dynamic, and multi-variate" (p. 20) problems. Thus, the modern Monte Carlo simulation obviates the physical tossing of a die (or flipping of a coin). The combination of assignment in Table 1 (simulation) with many repetitions (Monte Carlo) via computer software and hardware results in the Monte Carlo simulation of the probability of outcomes in tossing a fair die with far more accuracy than could be achieved with the manual methods used by Buffon or Weldon.

The richness of possibilities for Monte Carlo simulation are truly amazing. Some examples include annealing, electromagnetism, image processing, and genetic linkage (Robert & Casella, 1999); inventory control, queuing systems at a two-minute car wash, expected waiting times, management planning, short-term forecasting, consumer behavior of switching brands, and customer product ordering behavior, (McMillan & Gonzalez, 1968); mass-supply systems, and quality and reliability of products (Sobol, 1974); growth of yeast in a sugar solution, cooling temperature of coffee, development of ability to perform pushups, estimating migration patterns, material or time delays, ecology of the Kaibab Plateau on the rim of the Grand Canyon, urban growth, sale and consumption of commodities, controlling dam water, projection of discovery of natural gas reserves, and heroin addiction's impact on a community (Roberts et. al, 1983); and studying random neutron diffusion in fissile material in the development of the atom bomb during World War II.

Sampling *With* vs *Without* Replacement

Sampling via Monte Carlo simulations can be conducted *with* or *without* replacement. In the examples using dice or coins, the correct sampling technique is *with* replacement. Once the result for the experiment has been recorded, the value obtained from the uniform pseudo-random number generator is returned to the repository of values that may again be drawn. This is because the spots don't leave the dice after being tossed and the heads don't leave the coin after being flipped.

Conversely, sampling *without* replacement would be appropriate in simulating the turning of cards. Once the Queen of Hearts has been turned, it is no longer in the deck, and cannot reappear. The Queen of Hearts must be prevented from further assignment. The choice of which technique to use in a Monte Carlo simulation is determined by what is being simulated.

The matter of sampling *with* vs *without* replacement is practically irrelevant when drawing variates from the continuous uniform distribution, which is represented by an infinite number of real numbers, each in turn with an infinite string of digits. Furthermore, this consideration is often moot with asymptotically large data sets. However, Monte Carlo simulation based on discrete and bounded distributions, and even more so with small sample data sets, may lead to different results based on which sampling technique is used.

Characteristics Of A High Quality Monte Carlo Simulation

There are a variety of factors that must be attended to in order to assure a Monte Carlo simulation is correct and useful. Some of these factors are as follows:

- the pseudo-random number generator has certain characteristics (e. g. a long "period" before repeating values)
- the pseudo-random number generator produces values that pass tests for randomness

- the number of repetitions of the experiment is sufficiently large to ensure accuracy of results
- the proper sampling technique is used
- the algorithm used is valid for what is being modeled
- the study simulates the phenomenon in question

Sawilowsky and Yoon (2001, 2002) vs Roberts and Henson (2002)

The Monte Carlo *simulation* by Sawilowsky and Yoon (2001, 2002) was conducted with

> A Fortran 95 program "written to randomly draw variates from a de Moivreian (i. e., normal) distribution and then randomly assigned to two groups ($n_1 = n_2 = 10$), with the first group designated the treatment group and the second the control. A two-sided two independent samples t test was conducted with nominal $\alpha = 0.05$. 10,000 repetitions were conducted. (p. 143).

Under the truth of the null hypothesis, the results indicated that the average of the absolute values of the effect size, Cohen's d, was not near zero, but rather, was approximately what Cohen (1988) categorized as a small treatment effect. Thus, the conclusion of their brief report was the publishing of the constituent effects sizes would be misleading.

The Monte Carlo *study* by Roberts and Henson (2002) was designed to examine the "amount of bias in the effect size" (p. 241). They used an S-Plus macro to

> generate two normally distributed populations of 1 million cases... the factors in this simulation study included the size of Cohen's d in the population, the standard deviation of the two populations, and the sample sizes of the two groups... A total of 5,000 pairs of

> samples were drawn from the populations within each condition of the simulation study. (p. 245)

The results of their study found "the amount of bias in d remained small under most conditions of consideration" (p. 247). Because the "average across samples tended to more closely approximate zero" under the truth of the null hypothesis, meaning "Cohen's d does not appear to be biased in practical terms" (p. 252), they concluded the opposite of Sawilowsky and Yoon (2001). Therefore, they supported the reporting of effect sizes for results that are not statistically significant.

Criticism of Roberts and Henson's (2002) Study
Nine Minor Criticisms

(1) Roberts and Henson (2002) claimed that "effect sizes can serve a valuable function to help evaluate the magnitude of a difference or relationship" (p. 241). Although effect sizes do quantify the magnitude of a difference or relationship, they do not evaluate it. Content knowledge of the research question is required to decide if the difference or relationship is of theoretical, clinical, or practical importance.

(2) Their Monte Carlo study was written in a recent albeit dated version of S Plus, which is a superb statistical package. There are advantages of using statistical packages over programming languages, such as ease of use. There have been bugs, however, in this software's pseudo-random number generator (e.g., see the discussion at www. insightful.com/support/faqdetail.asp?FAQID=137 &IsArchive=0).

On the positive side, if a glitch due to this bug occurred it should have produced an observable error message. The built generator has an excellent period length (i. e., $2^{64} - 2^{32}$) compared with most other statistical packages, but the algorithm it is based on fails at least four DIEHARD tests of randomness (available at http://stat.fsu.edu/~geo/. The default option requires the programmer to reset the seed, which was not mentioned by Henson and Roberts (2002). Otherwise, the two "populations" of 1 million values would be identical. The current version of S-Plus eliminated these potential concerns.

(3) The entry of .0611 for the maximum r^2 when d = .00 and n_l = n_2 = 10 in Table 2 is obviously a typographical error.

(4) They presented "descriptive statistics" (p. 247), including the minimum and maximum *d*, in Tables 1 - 3. Roberts and Henson (2002) mistakenly labeled and considered the strongest negative effect size as a "minimum". Although mathematically it is a "minimum", in the context of effect sizes, the minimum *d* is, of course, defined as zero.

(5) Whereas Sawilowsky and Yoon (2001, 2002) used 10,000 replications and reported results to three significant digits, Roberts and Henson (2002) used 5,000 repetitions, but reported results to four significant digits. The number of repetitions was likely due to the limitations of using an S-Plus macro instead of Fortran, as the latter is far more flexible to program and faster in terms of execution. (It is not uncommon to use millions of repetitions to gain precision.)

(6) Roberts and Henson (2002) conducted their study on "5,000 pairs of samples" that "were drawn from the populations" (p. 245). Thus, they used sampling *without* replacement. This is incorrect if the intent was to simulate the occurrence of test scores, group means, p values, or effect sizes. For example, the appearance of an IQ score of 107.5 as one sample mean should not preclude another sample from having the same mean. Each sample mean of a pair must be returned to the population, with the chance of being drawn again being equal to every other possible sample mean. This is accomplished by sampling *with* replacement.

(7) Because the study was conducted on Cohen's *d* (and r^2), which is a standardized value, there was no need for Roberts and Henson (2002) to include three different population standard deviations, and hence, two-thirds of their study (i. e., Tables 2 - 3) is redundant.

(8) There is little justification for publishing Monte Carlo work when results can be computed easily and directly. The bias in *d* can be computed analytically under population normality, which is the only distribution Roberts and Henson (2002) examined. Cohen (1988) noted:

> It has been shown by Hedges (1981) and Kraemer (1983), in the context of the use of d_s in meta-

analysis that the absolute value of d_s is positively biased by a factor of approximately (4df - 1)/(4df - 4), which is of little consequence except for small samples. (p. 66)

Their Monte Carlo results for the bias of Cohen's *d* = .2, .5, and .8 in Table 2 for n_l = n_2 = 10 differ from (4df - 1)/(4df - 4) by only .005, -.014, and -.013, respectively. The results should converge as the number of repetitions in their Monte Carlo study increase.

(9) Roberts and Henson (2002) cited literature reviews indicating authors inadequately documented effect sizes. They cited editors who promoted citing effect sizes. They cited the same list of journals previously given by Thompson (2001, p. 83), whose editors require reporting of effect sizes. Their point is well taken, despite the apparent recanting of this form of persuasion by Thompson (2002), who cautioned "headcounts of views are not perfect indicators of truth" (p. 85). Nevertheless, Roberts and Henson's (2002) Monte Carlo study did not present any compelling reason to report effect sizes *when the null hypothesis remains tenable.*

Major Criticism

Sawilowsky and Yoon (2001, 2002) never "argued that small effects can in some cases be due solely to sampling error" (Roberts & Henson, 2002, p. 245), as claimed by Roberts and Henson and which was the premise of their counter-study. Instead, Sawilowsky and Yoon (2002) demonstrated the trouble with reporting effect sizes for studies that were not statistically significant by simulating the process and examining the false impression that would subsequently be created in the literature. The following fabricated data sets (Data Set A and Data Set B) represent two possible patterns of results in terms of effect sizes when the null hypothesis is tenable.

Table 2. Hypothetical Effect Sizes (e. g., Cohen's d) For Data Sets A & B Over Six Replications.

A	.001	-.004	.003	.008	-.003	-.005
B	.23	.12	-.07	.17	-.27	-.17

To appreciate the impact of the information (hypothetical results) in the table above, consider the following vignette. First, consider Data Set A. Readers of the literature will see an effect size of .001 published in a study of interest, -.004 in the subsequent study, and so forth. If the reader has a good memory, it would be remembered that the typical positive effect size averaged .004, and the typical negative effect size averaged -.004. The sign of the effect size, to be discussed further below, depends on the context of the study. Prior to the sought-after and highly prized meta-analysis, what message will have formed in the mind of the reader of the literature? Most likely, there isn't much here.

Now consider Data Set B. The effect size for the first study was .23. Although marginally respectable, the study was published to publicize a subtle, yet detectable treatment effect in education or psychology research. A year later, a replication study appeared in the literature. The magnitude of the effect size was only .12. Explanations were given for the reduction (e. g., the reliability estimate was lower, the sampling plan was inadequate, the period of treatment was reduced). After another year passed and the next replication appeared in the literature, serious questions regarding the veracity of the intervention arose. This was because the effect size for the third non-statistically significant study was only -.07.

This impression dissipated somewhat with the appearance of the fourth study and its effect size of .17. After the fifth and six studies, however, readers of the literature were thoroughly confused on the effectiveness of the intervention. What message might be formed in their minds? A reader with a good memory may recall the magnitude of the effect sizes averaged approximately .2, indicating there was a small but important treatment effect. Readers who (a) recalled the oldest studies maintained the direction was positive, or (b) recalled the newest studies maintained the direction was negative.

When the readers are presented with the published meta-analysis on the series of non-statistically significant studies, they will realize they have been misled. In the absence of a Type I error, the meta-analytic synthesis will determine the studies conducted over the past half-decade are not statistically significant. The meta-analysis, and the misconceptions it clarified, would have been obviated initially had effect sizes for non-statistically significant studies not been published in the first place.

The Sawilowsky and Yoon (2002) Monte Carlo was a simulation designed to determine which type of data set should readers of the literature expect to see under the truth of the null hypothesis. Are the magnitudes clustered about 0.0? The absolute value was taken, and it was determined that the typical magnitude expected is not near zero, but rather, what Cohen (1988) labels a small treatment outcome. Their simulation showed readers should expect to see results such as that depicted by Data Set B, not Data Set A. In contrast, Roberts and Henson's (2002) work was a Monte Carlo study of the bias of d, which does not relate to the process being simulated.

(Without remarking on it, Roberts and Henson, 2002, with slightly different study parameters, found the strongest effect sizes to be -2.31 and 2.06 for negatively and positively signed d's, respectively. You think you've got trivials? These huge results occurred with a treatment modeled by random numbers! Publishing specious effect sizes of such astronomically high magnitude (i.e., ±2.19) could wreak havoc in the literature. Sawilowsky and Yoon, 2001, 2002, considered reporting results in this fashion. It was decided, however, that to be realistic, the simulation should depict the typical magnitude expected, not extrema.)

Conclusion

Consider the chaotic fashion in which meta-analyses are currently being conducted. One researcher is not the holder of results from many tightly integrated experiments, publishing only the final meta-analysis. If that were the case, the presence of effect sizes for non-statistically significant results, duly noted and preserved as they occurred, would never become a misleading menace to the public.

Therefore, Sawilowsky and Yoon's (2001, 2002) brief report was based on taking the absolute value of Cohen's d to determine the typical magnitude expected when an intervention was random numbers. Roberts and Henson's (2002) argument against taking the absolute value was "in real experiments, it is known which group received the intervention" (p. 244). Is their

position correct as far as readers of the literature are concerned? In some treatment vs control studies, the effect of the treatment is demonstrated when the mean of the treatment group is *higher* than that of the control group; in other contexts when the mean of the treatment group is *lower* than that of the control group. For example, the *same* intervention might be used to increase self-esteem scores (treatment group mean is greater than control group mean), and reduce the number of times per week the bed was wet (treatment group mean is lower than control group mean).

The direction (sign of + or -) of that same intervention is entirely arbitrary. The sign depends on the context of the use of the intervention. If one researcher held all of the interim results, then the interpretation could safely rest on the meta-analysis, as the context would be known. However, the reader of the literature, who is getting these results piecemeal, will have the nigh impossible task of making sense of the contexts of a series of independently conducted studies published sporadically over time.

In addition to the above vignette, consider using a compound designed to block the serotonin uptake pump in a treatment one vs treatment two study on patients at risk for suicide. Suppose 30 mg, a common dosage for depression, was being compared with 70 mg, a common dosage for trichotillomania and other obsessive-compulsive disorders. Which dosage is the intervention? Clearly, the resulting direction (sign of + or -) is arbitrary. Thus, both the *magnitude* and the *sign* of published effect sizes for non-statistically significant studies mislead the public.

Cohen (1988) noted the researcher "hardly needs convincing of the centrality of the concept of effect size (ES) to the determination of power or necessary sample size in research design" (p. 531). "It is, after all, what science is all about" (p. 532).Yet, Cohen (1988, p. 10) opined that of all the factors in research design, behavioral scientists understand effect size the least. "Whatever the manner of representation of a phenomenon ... the null hypothesis always means the effect size is zero...[but] when the null hypothesis is false, it is false to some specific degree, i.e., *the effect size* (ES*) is some specific nonzero value in the population*" (Cohen, 1988, p. 10). Thompson (1996, 1999), supported by Roberts and Henson (2002), called for publishing specific nonzero

values under the truth of the null hypothesis. According to Cohen (1988), however, "the ES serves as an index of degree of departure *from* the null hypothesis" (p. 10, italics added for emphasis).

References

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153.

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

Gerrold, D. (1973). *The story behind a* Star Trek *show! "The trouble with tribbles": The birth, sale, and final production of one episode.* NY: Ballantine.

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

Kahn, H. (1966). Multiple quadrature by Monte Carlo methods. In A. Ralston & H. S. Wilf, (Eds.) *Mathematical methods for digital computers*. *Vol. 1*, 249 - 257.

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

Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*, *11*, 143-155.

McMillan, C., & Gonzalez, R. F. (1968). *Systems analysis: A computer approach to decision models.* (Revised ed.). Homewood, IL: Irwin.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*, 335-351.

Moshman, J. (1967.) Random number generation. In A. Ralston & H. S. Wilf, (Eds.) *Mathematical methods for digital computers*. *Vol. 2*, 249 - 263.

Negoita, C. V., & Ralescu, D. (1987). *Simulation: Knowledge-based computing, and fuzzy statistics*. NY: Van Nostrand Reinhold Co.

Norlén, U. (1975). *Simulation model building: A statistical approach to modelling in the social sciences with the simulation method.* NY: John Wiley & Sons.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157-175.

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

Roberts, C. P., & Casella, G. (1999). *Monte Carlo statistical methods.* NY: Springer-Verlag.

Roberts, N., Andersen, D., Deal, R., Garet, M., & Shaffer, W. (1983). *Introduction to computer simulation: A system dynamics modeling approach*. Portland, OR: Productivity Press NY: John Wiley & Sons, Inc.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*, 21-26.

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., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation via Fortran*. Rochester Hills, MI: SS, Inc.

Sawilowsky, S. S., & Yoon, J. (2001). *The trouble with trivials (p > .05).* Paper presented at the 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.

Sobol, I. M. (1974). The Monte Carlo Method. (Translated and adapted from the second Russian edition by R. Messer, J. Slone, and P. Fortini. ERIC No. ED 184845.

Student. (1908a). On the error of counting a haemacytometer. *Biometrika*, *5*, 351-360.

Student. (1908b). The probable error of a mean. *Biometrika*, *6*, 1-25.

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. (2001). Significance, effects sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80-93.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, *80*, 64-71.