Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1 Article 25

5-1-2003

Trivials: The Birth, Sale, And Final Production Of Meta-Analysis

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

DOI: 10.22237/jmasm/1051748700

Journal of Modern Applied Statistical Methods
May, 2003, Vol. 2, No. 1, 242-246

1538 – 9472/03/$30.00

Invited Debate: Rejoinder
Trivials: The Birth, Sale, And Final Production Of Meta-Analysis

Shlomo S. Sawilowsky
Educational Evaluation & Research
Wayne State University

The structure of the first invited debate in *JMASM* is to present a target article (Sawilowsky, 2003), provide an opportunity for a response (Roberts & Henson, 2003), and to follow with independent comments from noted scholars in the field (Knapp, 2003; Levin & Robinson, 2003). In this rejoinder, I provide a correction and a clarification in an effort to bring some closure to the debate. The intension, however, is not to rehash previously made points, even where I disagree with the response of Roberts & Henson (2003).

Key words: Effect size, meta-analysis, Monte Carlo simulation, trivials

Introduction

Many such techniques were developed throughout the half-century before Gene Glass gave meta-analysis its modern name in 1976. Twenty-four years later, despite considerable developments in the field, Glass (2000) lamented the use of meta-analysis. Nevertheless, there remain powerful lobbyists for meta-analysis, including those who use their editorial position to coerce statistical policy to ensure its survival.

The question arises: Has the advent of meta-analysis in social and behavioral sciences in the past quarter century increased the ability to synthesize and evaluate research, as compared with – for example – traditional scholarly analysis? Or, perhaps has meta-analysis become the favored tool in the hunt for Type I errors? When professional associations and learned societies are lobbied to require their journals report and interpret effect sizes, the coin of the realm of meta-analysis, "in all studies, regardless of whether or not statistical tests are reported" (Thompson, 1996, p. 29) even for "non-statistically significant effects" (Thompson, 1999. p. 67), the answer to the initial question will be negative, and the latter question will be positive.

Email the author at shlomo@wayne.edu. The title of this article is based on Gerrold (1973).

This was the point I made in Knapp & Sawilowsky (2001), and Sawilowsky and Yoon (2001, 2002). A Monte Carlo simulation was conducted to determine what magnitude of effect sizes should be expected if studies, whose results were obtained under the truth of the null hypothesis, were published piecemeal for the sake of meta-analysis. The Monte Carlo simulation indicated that effect sizes near zero should not be expected. Hence, publishing effect sizes for nonstatistically significant study results are ill advised.

Roberts & Henson (2002)

Subsequently, Roberts and Henson (2002) demurred, and the battle was joined. They advanced the following argument: Sawilowsky and Yoon's Monte Carlo simulation (2001) must imply that the bias associated with effect sizes is large under the truth of the null hypothesis. Hence, Sawilowsky and Yoon (2001) cautioned against the publication of effect sizes in the absence of statistical significance. Yet, Roberts and Henson's (2002) Monte Carlo study indicated the bias was near zero. Therefore, the publication of such effect sizes should not be suppressed.

The purpose of the target article (Sawilowsky, 2003) in this debate was to illustrate this is a straw-person argument. The bias associated with effect sizes under population normality is easily determined, and indeed its

average is near zero. This result was known two decades prior to the Roberts and Henson (2002) Monte Carlo study (Cohen, 1988, p. 66). This does not, however, detract from the main pronouncement of Sawilowsky and Yoon (2001, 2002). The expected magnitudes (i. e., absolute value) of the constituent effect sizes are *not* near zero. Publicizing these non-near zero values, for the sake of meta-analysis, will wreak havoc in the literature.

Levin & Robinson (2003)

Levin and Robinson's (2003) comments are very insightful. A premise of Sawilowsky and Yoon (2001, 2002) is that scientific research is by definition comprised of multiple-study investigations, regardless of who actually conducts the experiment.

Knapp (2003)

Knapp's (2003) comments prompt a (1) correction and a (2) clarification.

(1) Material in Knapp's (2003) appendix correctly estimates the non-near zero magnitudes of the effect sizes to be approximately $|\overline{d}|$ = .34, not .17 as indicated in Sawilowsky and Yoon (2001, 2002). I reran the Monte Carlo simulation and got approximately the same value reported by Knapp (2003). I cannot find the errant value in my lab notes, so I must conclude that by some error I halved the result to present the value as a "±" when setting the table for publication. Nevertheless, the correct result *doubles* the warning raised by Sawilowksy and Yoon (2001, 2002), as .34 is situated half-way between what Cohen (1988) loosely defines as a "small" and a "moderate" effect size.

(2) Knapp (2003) estimated the correct value via formulas provided by Kraemer (1983), and thus, he argued that Monte Carlo methods were not necessary. He amplified this with remarks on the general utility of Monte Carlo in the presence of mathematical statistics. As the latter comment goes to the issue of one of the three missions of *JMASM*, it demonstrates to me that the message of the power of Monte Carlo methods requires further demonstration and publicity.

As noted in the target article (Sawilowsky, 2003), there usually is no need to invoke Monte Carlo methods when results may be obtained easily, conveniently, and accurately via mathematical statistics. For example, the statistical properties of the t test, under asymptotic conditions, can easily be determined through an expansion of moments. The question in applied statistics, however, pertains to the small samples properties of this test, and, its properties under departures from underlying assumptions, especially for real data sets. Here, asymptotic mathematical statistics have utterly failed, and have misled the discipline. Monte Carlo methods, however, have been used successfully and convincingly to set the record straight regarding the properties of the t and other statistics.

Methodology

Sawilowsky and Yoon (2001, 2002) was remiss in not explaining that in Monte Carlo work, (1) should desirable results be obtained when underlying assumptions are met, it is still necessary to proceed to when underlying assumptions are not met, but, (2) should undesirable results be obtained when underlying assumptions are met, there is little point in proceeding to when underlying assumptions are not met. Thus, when non-near zero results were obtained under normality, the remainder of the Monte Carlo simulation results obtained became irrelevant and were not presented in Sawilowsky and Yoon (2001, 2002). However, to respond to Knapp's criticism against appealing to the use of Monte Carlo methods, these results are provided below.

Results

Table 1 contains the Type I error rates of the two independent samples t test under the De Moivre distribution for the purpose of demonstrating the viability of the algorithms used. The $|\overline{d}|$ for fail to reject H_o is shown to be about .34 for α=.05, and about .38 for α=.01, when the sample size is 10. The 95% bracketed interval for $|\overline{d}|$ is [.2841489 - .4107949] for α=.05, and [.2968488 - .4601668] for α=.01.

Because Knapp was concerned about this sample size, new results are presented below for samples of size 20 and 30. To address concerns regarding the number of repetitions, it was increased from 10,000 to ten million. Additional precision was obtained by using critical values to six decimals. The warning of Sawilowsky and Yoon (2001, 2002) remains fully supported by these new results.

Table 1. Two Independent Samples t Test Type I Error Rates, $\overline{|d|}$ (Fail To Reject H_o), $\overline{|d|}$ (Reject H_o); For De Moivre (Normal) Distribution, And Various Sample Sizes And α Levels.

Statistic	α=.050000	α=.010000
	$n_1=n_2=10$	
Type I Error Rate	.0499992	.0099861
Fail to Reject H_o	.3474719	.3785078
Reject H_o	1.217658	1.571810
	$n_1=n_2=20$	
Type I Error Rate	.0499181	.0099800
Fail to Reject H_o	.2348740	.2547229
Reject H_o	.7940045	1.001228
	$n_1=n_2=30$	
Type I Error Rate	.0500528	.0099930
Fail to Reject H_o	.1891833	.2053082
Reject H_o	.6326703	.7928227

Notes: Critical t Taken To Six Decimals. Each Cell Entry Is Based On 10,000,000 Repetitions.

Knapp (2003) obtained approximately $\overline{|d|}$ = .34 without appealing to a Monte Carlo procedure. (Indeed, in e-mail correspondence, he delivered yet another method to obtain these results. It was a less satisfying solution 3, as it depended on the simulation of values with unknown characteristics by hand, instead of values with known characteristics by machine.) However, Sawilowsky and Yoon (2001, 2002) was not a Monte Carlo *study* to determine this value; it was a Monte Carlo *simulation* designed to determine the magnitude of effect sizes expected under the truth of the null hypothesis. In retrospect, perhaps the use of $\overline{|d|}$ to communicate the study results obscured the objective.

Indeed, it takes a Monte Carlo simulation to determine the values in Table 2, which are the first 20 of ten million from the first run of the Fortran program that produced the value of .3474719 in Table 1. The simulation results are understood as follows. The first study to appear in the literature regarding a certain outcome, that is not statistically significant, will publicize a large effect size of .9. The second study to appear in the literature will be about .24, followed by a study that obtained an effect size of about -.18. The subsequent study will follow with an effect size of .31, and so forth.

Table 2. First Twenty Of 10,000,000 Simulated Values of $\overline{|d|}$ For (Fail To Reject H_o) For De Moivre (Normal) Distribution, $n_1=n_2=10$, α=.05.

#	ES	#	ES
1	.902532	11	-.214086
2	.239664	12	-.386423
3	-.184106	13	.100410
4	.311091	14	-.682867
5	.291022	15	.305013
6	-.204143	16	-.537210
7	-.105137	17	-.410020
8	.662463	18	-.330778
9	.111973	19	.168260
10	-.366065	20	.202596

The objective of Sawilowsky and Yoon (2001, 2002) was to have proponents of publishing these effect sizes imagine the incorrect message this will promote in the literature. After all, these are effect sizes obtained for an intervention modeled as random numbers! Clearly, the magnitudes of these values are non-near zero. (It should be recognized that the interpretation of the simulation results can begin at any arbitrary point within the 10 million effect sizes.)

Roberts and Henson (2002) indicated the maximum effect sizes obtained in their simulation. It was so huge that it prompted the title of Sawilowsky (2003). The maximum effect sizes obtained here for $n_1=n_2=10$, when there was a fail to reject decision under the truth of the null hypothesis, was max $\overline{|d|}_{\alpha=.05}$ = .9942942 and max

$\overline{|d|}_{\alpha=.01}$ = 1.56907 for the De Moivre distribution. This means that an intervention modeled by random numbers can produce an effect size as large as d = ±.99 or d = ±1.6, for α = .05 and .01, respectively! Why would the members of any committee on statistical practices and reporting empowered by their professional association or learned society give credence to the position of the lobbyist who promotes the piecemeal publication of apparently huge albeit trivial effect sizes?

It is likely possible, although difficult, to obtain mathematical solutions for $\overline{|d|}$ for small samples under population nonnormality for certain theoretical distributions. It is easy, however, to obtain results via the Monte Carlo method, as indicated in Table 3. It is impossible, however, to obtain solutions for $\overline{|d|}$ using mathematical statistics for the populations represented by real data sets. The results are easily obtained, however, via Monte Carlo methods, as indicated in Table 4.

Table 3. $\overline{|d|}$ (Fail to Reject H_o) For Various Theoretical Distributions, Sample Sizes, And α Levels.

Distribution	α=.050000	α=.010000
	$n_1=n_2=10$	
Uniform	.3439692	.3748572
Mixed Normal	.4028708	.4149501
Cauchy	.4047977	.4177936
	$n_1=n_2=20$	
Uniform	.2336624	.2535020
Mixed Normal	.2713618	.2781797
Cauchy	.2766581	.2851480
	$n_1=n_2=30$	
Uniform	.1885313	.2046196
Mixed Normal	.2133092	.2209231
Cauchy	.2228022	.2299003

Notes: Critical t Taken To Six Decimals. Each Cell Entry Is Based On 10,000,000 Repetitions. The Mixed Normal distribution is comprised of two distributions: (1) Z(0,1) with frequency of 95%, (2) Z(22,10) with frequency of 5%.

Table 4. $\overline{|d|}$ (Fail to Reject H_o) For Various Psychology/Education Data Sets, Sample Sizes, And α Levels.

Data Set	α=.050000	α=.010000
	$n_1=n_2=10$	
Bimodal (P)	.3408427	.3716145
Asymmetry (P)	.3594031	.3877410
Mass At Zero (E)	.3646502	.3864528
	$n_1=n_2=20$	
Bimodal (P)	.2314171	.2512609
Asymmetry (P)	.2372115	.2572745
Mass At Zero (E)	.2355214	.2562985
	$n_1=n_2=30$	
Bimodal (P)	.1877642	.2036923
Asymmetry (P)	.1902705	.2064020
Mass At Zero (E)	.1909938	.2073510

Notes: Critical t Taken To Six Decimals. Each Cell Entry Is Based On 10,000,000 Repetitions. P = psychometric instrument, A = education test.

Conclusion

As Knapp (2003) pointed out, "Kraemer (1983) showed that d follows the t sampling distribution with $n_1 + n_2 - 2$ degrees of freedom" (p. 242). From this statement alone it should be obvious that the publishing of effect sizes should be handled the same as p values associated with the t statistic in hypothesis testing (as opposed to so-called significance testing, which in my view is outside the boundary of the scientific method).

A nonsignificant obtained t is interpreted, based on the samples, as the difference in means between the two groups are not statistically significantly different from zero. More formally, there is no evidence that the two samples were drawn from populations with different values of μ. For this reason, it is the policy at many journals that p values for nonsignificant t statistics are suppressed from publication. (Typically, the author supplies an "*" in tabled statistical material to indicate the result was not significant at the à priori specified α level.)

The same should hold true for d. When the t is not statistically significant, the effect size (regardless of its magnitude) is not statistically

significantly different from zero. Unfortunately, this type of argument has not been compelling to the meta-analysis lobby.

The purpose, therefore, for the Monte Carlo simulation by Sawilowsky and Yoon (2001, 2002), was to provide another type of demonstration that the publicizing of effect sizes associated with nonstatistically significant results are an invitation to disaster in the literature. One has but to consider the effects of the proliferation of trivials (e.g., such as those in Table 2) to reject the position of lobbyists seeking to promote the piecemeal publishing of effect sizes for meta-analysis in a fashion never envisioned by its developers.

References

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.

Knapp, T. R. (2003). Was Monte Carlo necessary? *Journal of Modern Applied Statistical Methods, 2*(1), 238-242.

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

Kraemer, H.C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics, 8,* 93-101.

Levin, J. R., & Robinson, D. H. (2003). The trouble with interpreting statistically nonsignificant effect sizes in single-study investigations. *Journal of Modern Applied Statistical Methods, 2*(1), 232-237.

Roberts, K. J., & Henson, R. K. (2003). Not all effects are created equal: a rejoinder to Sawilowsky. *Journal of Modern Applied Statistical Methods, 2*(1), 227-231.

Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods, 2*(1), 218-226.

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.

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.