Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1 Article 22

5-1-2003

Not All Effects Are Created Equal: A Rejoinder To Sawilowsky

J. Kyle Roberts
University of North Texas, kroberts@unt.edu

Robin K. Henson
University of North Texas, rhenson@unt.edu

DOI: 10.22237/jmasm/1051748520

Journal of Modern Applied Statistical Methods
May, 2003, Vol. 2, No.1, 226-230

1538 – 9472/03/$30.00

Invited Debate: Response
Not All Effects Are Created Equal: A Rejoinder To Sawilowsky

J. Kyle Roberts
University of North Texas

Robin K. Henson
University of North Texas

In the continuing debate over the use and utility of effect sizes, more discussion often helps to both clarify and syncretize methodological views. Here, further defense is given of Roberts & Henson (2002) in terms of measuring bias in Cohen's *d*, and a rejoinder to Sawilowsky (2003) is presented.

Key words: Effect size, Cohen's *d*, bias, simulation

Introduction

Under a spirit of collegiality and zeal to further the field of research, dialogues like this play an important role in discussing areas where researchers both agree and disagree. Through open-ended dialogue, it is hoped that readers will continue to see the benefit in debate about important topics.

In this brief rejoinder to Sawilowsky (2003), we will provide discussion to the nine minor criticisms and one major criticism point by point. Although the first portion of his paper is lengthy, it does not bear comment on because it was expertly written and we do not disagree with any of the substance laid therein.

As we respond to each of the criticisms, however, we feel it important to note two things. First, the point of our paper was to show whether or not Cohen's *d* contains any amount of bias and is therefore in need of a correction to account for this bias.

Dr. Roberts is an Assistant Professor of educational research whose research interests include hierarchical linear modeling and measurement. Correspondence can be sent to kroberts@unt.edu. Dr. Henson is an Assistant Professor of educational research. His areas of research include applied statistics, measurement, reliability generalization, and self-efficacy theory. Correspondence can be sent to rhenson@unt.edu.

For all practical purposes, our answer to this question was NO. As we stated in our article, "the amount of bias in *d* remained small under most conditions of consideration . . . [and the] incredibly small amount of difference between the population *d* and the average sample *d* leads us to believe that *d* is in fact not biased in terms of practical differences" (p. 247, 251).

Second, we examined Thompson's (2002) proposed correction of *d* for accuracy and to see whether or not the correction was even necessary. In response to this proposed correction, we state, "although this correction of *d* seems to make sense theoretically, it overcorrects for the actual amount of bias" (p. 251).

As we begin our reply, we would like to note that NOWHERE in the rebuttal does Sawilwosky (2003) refute either of these findings. Instead, the arguments fall into two categories: minor criticisms that are mostly methodological, and one major criticism that has to do with the publishing of reported effect sizes. Once again, it bears mentioning that none of these criticisms, once having addressed and clarified the methodological issues, directly calls into suspect the findings of Roberts and Henson (2002).

Responses to Minor Criticisms

Criticism 1: Effect sizes help evaluate

Although we agree with Sawilowsky's statement that effect sizes do not evaluate the effect of a difference or relationship, we want to note that we pointed out in our paper that the purpose of the effect size is to "*help* evaluate the magnitude of a difference" (emphasis ours, p. 241); for judgments are of course made by people.

As Sawilowsky (2003) quoted this very statement, we do not see any point of disagreement here.

Criticism 2: S-PLUS Random Number Generator

As Sawilowsky makes a good point about resetting the random number seed, it should be pointed out that this seed was reset for both populations so that they weren't identical. Concerning the random number generator (RNG) in S-PLUS, however, we feel that the criticisms are unwarranted. The DIEHARD tests for randomness were designed to work on RNGs that assume 32 random bits. The RNG for S-PLUS is 31 bit. As a result it should be assumed that the RNG will fail some of the tests that are 32 bit based. If there is a need for a 32 bit RNG, then S-PLUS users can install a patch that will paste together 16 bits from each of two consecutive numbers and then the S-PLUS RNG will pass all of the DIEHARD tests. Also, the bug which Sawilowsky speaks of only applies to the Chi-Square distribution function when X is large (e.g., 10^13). (Our thanks to Tim Hesterberg from Insightful Corporation for his guidance concerning the RNG).

Criticism 3: Typo!!

The entry of .0611 for the maximum r^2 when d = .00 and $n_1=n_2=10$ in Table 2 should read .611.

Criticism 4: Negative values for d

Although Sawilowsky (2003) disagrees, there are instances when a minimum *d* is actually less than zero. Consider the directional hypothesis t-test where we are comparing the effects of a diet pill on 100 people. We randomly assign people to one of two groups; experimental and control. The point of the study is to show the effect of the diet pill on the experimental group. Let's suppose that when we compare the mean weights of the people at the beginning of the study and note that both group means are 200, and then again at the end of the study and note $\overline{X}_{\exp} = 225$ and the $\overline{X}_{control} = 200$. If we were to consider that the σ = 35, then we could compute the *d* for this study as:

$$d = \frac{200 - 225}{35} = -0.714. \qquad (1)$$

Consider that it would be incorrect to interpret the absolute value of this formula (Cohen, 1988, formula 2.2.2) because we are witnessing an actual negative effect of the diet pill (e.g., people who took the diet pill actually gained weight). If we were to follow the logic of Sawilowsky, we would either interpret this as a positive effect or simply assume the effect is zero. In this case, interpreting a negative effect is *important*. It means that the diet pill worked worse than if we had done nothing at all! Sawilowsky also mistakenly states that the minimum effect (or *d*) should be defined as zero when in fact this is not true (c.f., Cohen, 1988, formula 2.2.1, p. 20).

As this formula applies to our study, we explicitly stated in our manuscript (p. 247) that the design of the study was to test this specific effect with a directional hypothesis where the expected effect was that the experimental group would have a larger mean in the population than did the control group (except for the case where *d* = .00).

Criticism 5: Repetitions

Although Sawilowsky and Yoon (2001) used 10,000 replication, we felt that 5,000 was plenty to obtain generalizability. This was not a limitation due to using a macro in S-PLUS as S-PLUS is a programming language and changing the number of replications is as simple as typing a new number into the script file. However, since Sawilowsky posited this as a criticism of the study, we re-ran all analysis with 10,000 replications and noticed that even under extreme conditions, estimates typically did not differ until the 1000^{th} decimal place!

Criticism 6: Sampling without replacement

We feel that we may have been misleading with our statement, "5,000 pairs of sample data were randomly drawn without replacement at the specified sample sizes" (Roberts & Henson, 2002, p. 246). What would have been better stated is that we sampled without replacement *within* each given replication. After people were drawn from the population for the replication, they were then re-inserted into the population at the completion of that replication.

We chose this method because it seemed counterintuitive to allow for the inclusion of the same person twice within each study (although the probability for being chosen twice is less than 1% for n = 100). We should have been clearer in pointing out that we sampled *with* replacement across the replications, just not inside each replication.

Criticism 7: Redundancy is reinforcement!!

Although Sawilowsky points out that there was no need for 2/3 of our study since there was no change in the standardized values, we felt it important to further reinforce the point that the spread of the data make simply a marginal difference in effecting the bias (or lack thereof) in both *d* and r^2. We would argue that if the results really were redundant then we would see exactly the same values in each of the tables, which we in fact did not. Therefore the inclusion of all three tables serves to reinforce the point that under multiple conditions, *d* shows practically no bias.

Criticism 8: Results that shouldn't be published?

This criticism probably should have been labeled under the "major criticisms" because it states "there is little justification for publishing Monte Carlo work when results can be computed easily and directly." As per our manuscript, we would again point out that the purpose of it was two-fold: to see if *d* contained bias and to see if Thompson's (2002) correction formula should be applied. If nothing else than to show that Thompson's formula "overcorrects for the actual amount of bias" (Roberts & Henson, 2002, p. 251), then the manuscript has merit. Furthermore our study shows that even though the correction cited by Sawilowsky may apply to meta-analysis, it seems of little concern to attempt to correct *d* in directional hypothesis settings.

Criticism 9: Compelling reasons to report effect sizes

We might restate that it was not the purpose of our study to present a "compelling reason to report effect sizes *when the null hypothesis remains tenable*." Our purpose was to investigate the bias in *d*. However, having said that we would like to add that in any given study, we *may* obtain a result in which the null hypothesis is tenable, *but that doesn't mean that the effect is not real*! We will deal more thoroughly with this in the next section.

Response to Major Criticism

Is the Effect Trivial or Not?

Sawilowsky (2003) suggests that he and Yoon (2001) never "argued that small effects can in some cases be due solely to sampling error" as we summized (Roberts & Henson, 2002, p. 245). Nevertheless, in their paper Sawilowsky and Yoon (2001) noted that reporting their simulated average Cohen's *d* effect of .17 would be "misleading because these effect sizes are specious" (p. 2). In their conclusion, the authors claimed: "It was shown that effect sizes should not be reported or interpreted in the absence of statistical significance" (Sawilowsky & Yoon, 2001, p. 4). (It should be noted as well that only the Sawilowsky & Yoon [2001] paper was referenced in our original article. Sawilowsky and Yoon's 2002 article resulting from this paper was not in print during our manuscript development, and therefore was not considered in our article.)

If Sawilowsky is not arguing that these effect sizes could be solely due to sampling error, then why not report and interpret them? Indeed, the average *d* of .17 was presented as a case when a non-zero effect was obtained from purely random numbers. Surely the logic of this conclusion suggests that small effects can be obtained even when the null hypothesis remains tenable under a statistical significance test. If the significance test is to be trusted over the small effect size, then from whence must the researcher conclude the effect originated? Under this logic, the effect must have been a function of sampling error.

Confused vs Informed Methodology and Readership

Sawilowsky (2003) proceeds in his major criticism by presenting two literatures of effect sizes (A and B). He supposes that after reading one of these literatures, a reader may be "thoroughly confused on the effectiveness of the intervention" (p. 223) because of the presence of non-statistically significant results mixed with other, presumably, statistically significant results. We agree that interpretation of such a literature may present certain challenges. Nevertheless, we would be hopeful that a *more informed use of*

statistics would be the solution to this difficulty rather than avoidance of potential confusion by replacing it with another source of misleading information.

(As a caveat, we would also be hopeful that even a modestly informed consumer of research would be able to determine the expected directionality of an effect, and whether the experimental group is expected to outperform or underperform the control on relevant outcomes. This assumes, perhaps, at least a modestly effective job at communication from the authors.)

It is at this point that we fundamentally disagree with Sawilowsky (2003). It is perhaps very appealing to some to employ statistical significance as a gatekeeper for reporting and interpreting meaningful outcomes. As we cited previously, Robinson and Levin (1997) and Levin and Robinson (2000) propose a reasoned argument for just such a two-stage process, where a finding must be deemed statistically significant before evaluation of the effect size is permitted. Of course, this would work only to the extent that the gatekeeper is effective in performing its duties.

This process also will only work when (a) the readership of the article understands fully the factors impacting statistical significance tests and the elements of power that underlie them and (b) the author understands and communicates these issues directly. Unfortunately, empirical studies have demonstrated that there are a great number of misconceptions about statistical significance testing (cf., Nelson, Rosenthal, Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; M. Zuckerman, Hodgins, A. Zuckerman, & Rosenthal, 1993), and so neither of these outcomes is likely on a widespread basis. Is this the method's fault or our own? We would suggest, of course, primarily the latter. Unfortunately, statistical significance testing has come to be treated among many researchers as a truly dichotomous outcome that relates directly to result importance. This interpretation is a result of many factors, none of which make the misinterpretation any more correct. As Sawilowsky (2003) correctly indicated, the context of the study is critical when interpreting both statistical significance and effect size outcomes.

It is of course very true that a small effect size may be due to sampling error. It is also just as true that the same small effect size may be a *real* effect in spite of it not being statistically significant due to a lack of power. The arguments presented by Sawilowsky (2003) simply do not discount the possibility (and yes, historical truth) that some very real effects may exist but be at risk of not being discovered due to a lack of statistical significance. Meta-analytically speaking, however, when these small but non-statistically significant effects are examined across studies, a more meaningful outcome may be discovered. While it is very easy for methodologists to say that these studies should have had more power, it is much more difficult to attain sufficient power for every study in all applied situations. Should we pay more attention to power? Yes, of course. Should we also recognize that some small effects may indeed be reasonable outcomes not due entirely to sampling error? Absolutely!

A better approach to this issue, in our view, would not just result in discussion of whether statistical significance should be the gatekeeper, or even whether small effects should necessarily be reported and/or interpreted, but rather how methodologists and applied researchers can seek a more informed understanding and use of both of these statistics for what they are.

Conclusion

Effect sizes are not final determinants regarding whether a result is meaningful any more than statistical significance tests are, and if we interpret effect sizes with the same rigidity that we have historically interpreted statistical significance testing, we are guilty of committing the same error yet again. Instead, researchers ought to view their studies in context with prior literature, make comparisons between their outcomes and those from prior studies, attend to power issues, and interpret the findings to the readership for what they are.

Is a small yet non-statistically significant effect important? Maybe, maybe not. We certainly would not know for sure without replication and some form of meta-analysis. We certainly could not do either of these, at least in a world where Type II error exists as much as its Type I counterpart, unless these same small effects were reported.

References

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

Levin, J. R., & Robinson, D. H. (2000). Rejoinder: Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher, 29*, 34-36.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41*, 1299-1301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement, 62*(2), 241-253.

Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. *Journal of Psychology, 55*, 33-38.

Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods, 2*(1), 218-225.

Sawilowsky, S. S., & Yoon, J. (2001, August). *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. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49-53.