The Trouble With Trivials (p > .05)

Name of 1st Author: Shlomo S. Sawilowsky
Evaluation & Research, College of Education, Wayne State University
351 EDUC, Detroit, MI USA 48202
shlomo@edstat.coe.wayne.edu

Name of 2nd Author: Jina S. Yoon
Educational Psychology, College of Education, Wayne State University
347 EDUC, Detroit, MI USA 48202
j.yoon@wayne.edu

1. Introduction

Among various reforms suggested to the American Educational Research Association=s editorial policies directed at Aeditors, program chairs, and reviewers@ (p. 28), Thompson (1996) recommended the reporting of effect sizes Aregardless of whether statistical tests are or are not reported@ (p. 29), Aeven [for] non-statistically significant effects@ (1999, p. 67). Similar advice was given by Carver (1993), Hulburt (1994), Rosnow and Rosenthal (1989), and Wilkinson (1999).

Heuristic support in the form of a thought experiment designed to illustrate concern with this suggested reform was given by Robinson and Levin (1997). They concluded that a better editorial practice is to AFirst convince us that a finding is not due to chance, and only then, assess how impressive it is@ (p. 23).

2. Purpose of This Study

This study presents Monte Carlo evidence, which is more convincing than a thought experiment, to demonstrate the perils of reporting and interpreting effect sizes arising from nonstatistically significant research studies.

3. Methodology

A Fortran 95 program was written to randomly draw variates from a Gaussian distribution and randomly assigned to two groups ($n_1 = n_2 = 10$), with the first group designated the treatment and the second the control. A two-sided two independent samples t test was conducted with nominal á = 0.05. 10,000 repetitions were conducted.

The effect sizes were considered (a) under the truth of the null hypothesis, and (b) for shift in location parameter, which was simulated by adding a constant Ac@, representing 0.52ó (a moderate effect size according to Cohen, 1988). This shift was selected to produce a power of about .2 for the t test for the given sample size and á level.

Small sample size and power level were chosen to mimic applied research. A balanced layout and a theoretically normally distributed data set were chosen to demonstrate what happens under the best of circumstances with regard to layout and data distribution assumptions. Nominal á was selected at 0.05 due to Cohen (1994).

4. Results

The results are compiled in Table 1. The upper panel represents the various outcomes due to random numbers, where the effect size is modeled as zero. The entries were obtained by averaging the absolute value of d, given by the formula $d = (0_t - 0_c)/s_{pooled}$, where s_{pooled} refers to the pooled ó. (The absolute value was taken because the order of 0_t and 0_c is arbitrary). The upper panel demonstrates the trouble with reporting and interpreting effect sizes when the results of the experiment are statistically trivial. A fail to reject decision was reached in 95% of the repetitions of the experiment. Reporting an average effect size of 0.17, which is approximately what Cohen (1988) judged to be a small effect size, is misleading because these effect sizes are specious. There can be no effect size because none was modeled

in the data generation.

(The remaining results aren=t relevant to the main pronouncement of this paper, but are presented to complete the illustration. The adverse effects of making a Type I error is demonstrated, because an average effect size of 0.51 was obtained, a medium effect size (Cohen, 1988), when in fact the true effect size is zero.

In the second case, depicted by the lower panel, one-tailed power is represented by averaging the effect sizes. As predicted by Cohen=s (1988) power tables, when the false null hypothesis is rejected, the average effect size reported and interpreted is a moderate 0.54. This is a meaningful effect size to report and interpret.

However, when the t test failed to reject the false null hypothesis, the resulting calculations indicate the effect size under consideration was only 0.18. Similar results were obtained for the t test when data were drawn from nonnormally distributed data, indicating that the t test is (a) robust with respect to Type II errors, but more importantly, (b) is less powerful than competitors such as the Wilcoxon Rank-Sum test, which would have rejected many more of these false null hypotheses.)

Table 1. Effect Sizes for $n_1 = n_2 = 10$, Gaussian distribution, nominal á = 0.05.

	H_o	
	True	False
	Random Numbers	
Decision		
Fail To Reject	0.169 ∀ .003	n/a
	(Type I Errors)	
Reject	0.508 ∀ .007	n/a
	Shift = 0.52ó, Power = 0.20	
		(Type II Errors)
Fail To Reject	n/a	0.180 ∀ .006
Reject	n/a	0.540 ∀ .005

5. Conclusion. It was shown that effect sizes should not be reported or interpreted in the absence of statistical significance. As Shaver (1993) noted, even A an effect size of 1 or larger may reflect a *trivial* result@ (p. 303, emphasis added). This is the trouble with trivials.

6. REFERENCES

Carver, R. P. (1978). The case against statical significance testing, revisited. *Journal of Experimental Education,* 61, 287-291.

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

Cohen, J. J. (1994). The Earth is round (p < .05). American Psychologist, 49, 997-1003.

Hulburt, R. T. (1994). *Comprehending behavioral statistics*. Pacific Grove, CA: Brooks/Cole.

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

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.

Shaver, J. P. (1993). *What statistical significance testing is, and what it is not. Journal of Experimental Education*, *61*, 293-316.

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.

Wilkinson, L., & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.