

5-1-2003

The Trouble With Interpreting Statistically Nonsignificant Effect Sizes in Single-Study Investigations


Joel R. Levin

University of Arizona, jrlevin@u.arizona.edu

Daniel H. Robinson

University of Texas at Austin

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Levin, Joel R. and Robinson, Daniel H. (2003) "The Trouble With Interpreting Statistically Nonsignificant Effect Sizes in Single-Study Investigations," *Journal of Modern Applied Statistical Methods*: Vol. 2 : Iss. 1 , Article 23.

DOI: 10.22237/jmasm/1051748580

Available at: <http://digitalcommons.wayne.edu/jmasm/vol2/iss1/23>

This Invited Debate is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in *Journal of Modern Applied Statistical Methods* by an authorized editor of DigitalCommons@WayneState.

Invited Debate: Comment
**The Trouble With Interpreting Statistically
Nonsignificant Effect Sizes in Single-Study Investigations**

Joel R. Levin
University of Arizona

Daniel H. Robinson
University of Texas at Austin

In this commentary, we offer a perspective on the problem of authors reporting and interpreting effect sizes in the absence of formal statistical tests of their *chanceness*. The perspective reinforces our previous distinction between single-study investigations and multiple-study syntheses.

Key words: Hypothesis testing, effect sizes, conclusion coherence

Introduction

Yes, everybody has troubles, and not just with trivials (Sawilowsky, 2003). We adopt a different perspective on the Sawilowsky vs. Roberts-Henson debates about appropriate methodologies for, and interpretations of, their respective Monte Carlo investigations (Roberts & Henson, 2002; Sawilowsky & Yoon, 2002).

Although we have decided biases concerning the rights and wrongs of that particular debate, we also have decided not to jump into the fray for two related reasons: (1) Knapp (2003) considers a number of general issues that need to be considered in the context of Monte Carlo simulation studies; and (2) because we regard such issues more as background to certain more fundamental research-related effect-size-reporting foreground issues, we elected to forego additional hammering on the former so that we might nail down the latter.

Single-Study Investigations vs Multiple-Study Syntheses

The major argument promoted here is one that we have presented elsewhere (e.g., Levin, 1998; Levin & Robinson, 2000; see also Onwuegbuzie & Levin, 2003). It can be sum-

marized as follows: Research conductors and consumers need to be more attentive to the different purposes/functions of an educational research article. Is it: (a) to report the results of an *individual* empirical study (a single-study investigation) or is it (b) to summarize a *set* of empirical studies (a meta-analytic multiple-study synthesis)?

If *a*, then we contend that hypothesis testing should be a critical precursor to effect-size estimation in telling the researcher's story; whereas if *b*, then effect-size reporting should play a more prominent role. In that context, a critical point of contention concerns whether the effect sizes associated with a single-study investigation should be interpreted in the absence of statistical significance. We have cast our nay votes on (and justifications for) this issue elsewhere (e.g., Levin, 1993; Levin & Robinson, 1999; Robinson & Levin, 1997; Robinson, Funk, Halbur, & O'Ryan, in press; Wainer & Robinson, in press) and will summarize our stance here.

Almost without exception, introductory statistics textbooks present examples based on single-study investigations. And, of course, a good number of single-study investigations are published in educational-research scholarly journals. Authors are forced to interpret the results of statistical inference tests – and this is where most of the troubles begin. In our previous writings, we have argued that statistical significance should serve a gatekeeper function to screen out effects whose direction has not been determined probabilistically. What may appear to be an interesting or important effect worth talking

Correspondence concerning this article should be addressed to Joel R. Levin, Department of Educational Psychology, University of Arizona, Tucson, AZ 85721. Email: jrlevin@u.arizona.edu.

about can easily be a chance finding, or one that is attributable solely to sampling error. In that case, by screening out spurious effects through a formal statistical test, an author protects the reader from erroneously interpreting the effects as if they were real.

Let us insert an important comment that has rarely been mentioned in relation to the so-called “significance-testing controversy.” It is simply that under the truth of the null hypothesis, testing the hypothesis that, say, two population means are equal or that the correlation between two variables is zero *is equivalent to testing the hypothesis that the effect size is equal to zero*. This may be readily appreciated when inferences about correlation coefficients are desired (because the correlation coefficient itself is an effect-size measure), though not as readily appreciated in the mean-difference situation.

Yet, it becomes apparent when one realizes that if the two population means are equal, then $\mu_1 - \mu_2 = 0$, and the corresponding population Cohen’s d effect-size measure is $0/\sigma = 0$. Thus, if a researcher applies a formal statistical test and then proceeds to report/interpret the sample effect size regardless of the test’s outcome, the question arises: What function did the statistical test serve, and why was it even conducted in the first place? That conclusion coherence issue (Levin & Robinson, 2000) is one that Roberts and Henson (2003) need to reconcile.

Another Troubling, Yet Telling, Hypothetical Example

As a sequel to a perplexing example (Levin & Robinson, 2000, p. 34-35; see also Levin’s, 1993, p. 379), let us consider an instructional intervention study with $n = 2$ participants in each of two conditions, where Condition 1’s scores are both 5 and Condition 2’s scores are both 6. For this example, a nondirectional permutation test would indicate that there is not sufficient evidence to conclude that the two populations are statistically different ($p = 2/6 = .333$, which far exceeds the conventional .05 level of statistical significance).

On the other hand, if an effect-size measure were computed and reported, it would likely be communicated as gigantic or even infinitely large, for in fact, in this particular instance d is equal to ∞ . Alternatively, with effect

size defined as a squared point-biserial correlation coefficient, one would conclude that there is perfect prediction of scores from knowledge of condition, with no score variability left to be explained, for r^2 turns out to be 1.00 here. Never mind that the study included only a couple participants per condition and that a valid statistical test performed on these data indicates a nonsurprising event associated with an outcome this or more extreme (i.e., $p = .333$), assuming that the population-identity hypothesis is true. Moreover, even if each condition were to include a third participant (resulting in $n = 3$) who produced the same scores of 5 and 6 for Conditions 1 and 2, respectively, the associated significance probability would be only $p = 2/20 = .10$, still above the conventional .05 level.

Although this particular example may sound extreme, far fetched, or even ridiculous, consider the myriad experiments in the educational research literature that involve a comparison of two different instructional approaches each based on three teachers, classrooms, or schools. With those teachers/classrooms/schools representing the appropriate data-analysis units (e.g., Levin & O’Donnell, 1999) and with the aggregated data equal to the values just described, the above significance probability of .10 applies.

This example also serves to clarify an oft-made argument that statistically nonsignificant effects are invariably associated with small or trivial effect sizes. Yes, a large-scale study (e.g., $N = 100$) with trivial effects (e.g., $d = .10$) can produce nonsignificant results, but so can a very small-scale study with huge effects (as was just illustrated). Conscientious conclusion-coherent researchers should refrain from interpreting such effects as either real (in both cases) or important (in the second case).

Our example leads to consideration of a converse situation as well, which was earlier discussed by Robinson and Levin (1997). The following question is regularly posed by one of us on Ph.D. qualifying examinations: “What is wrong with a researcher’s claim that ‘although the anticipated outcome did not quite reach statistical significance in this study, it would have if only a few more participants had been included’?” This claim is reminiscent of the substance of Thompson’s (e.g., 1989, 1996) proposed “what if” analyses and something toward which Roberts and

Henson (2003) tread dangerously close. (We are also troubled by the researcher's use of the term "quite" in the qualifying-examination question, as will be reflected in our concluding paragraph.) Thus, in our above amended example based on $n = 3$ participants per condition (for which $p = .10$), can it be claimed that if only one more participant were added to each condition the difference between conditions would have been statistically significant (since with $n = 4$, $p = 2/70 = .029$ according to a two-sample permutation test)? Well, could it?

Only if you are willing also to add that the outcome produced by the two additional scores (resulting in $n = 4$ participants per condition) mimicked exactly what was present in the original data. In the case of a two-sample permutation test, just as all three Condition 2 participants had higher scores than all three Condition 1 participants in the actually conducted study, only if the additional participant in each condition maintained that situation would there be a statistically significant difference at the .05 level. In contrast, if either the additional Condition 1 participant were to score higher than any Condition 2 participant or the additional Condition 2 participant were to score lower than any Condition 1 participant, then $p < .05$ statistical significance would not be attained (see, for example, Fisher, 1960, pp. 11-15).

The key to answering the qualifying-examination question is recognizing that one cannot simply *assume* that the mean difference or pattern will stay exactly the same with the addition of a few more participants. That is precisely the reason why one needs to collect actual data and conduct the analysis, rather than sitting around thinking in hypothetical "what if?" terms. Robinson, Fouladi, Williams, and Bera (2002) provide empirical data bearing on "what if" pondering and Hoenig and Heisey (2001) discuss an equally troubling related issue, post hoc or observed power analyses.

But we have other fish to fry. In Roberts and Henson's (2003) concluding paragraph, it is implied that researchers would be unable either to conduct replication studies or to perform meta-analyses unless authors calculate and report *all* effect sizes – including statistically nonsignificant ones. Let us consider each of the two implied components (replication studies and meta-analyses) of this contention in turn.

Is Effect-Size Information A Necessity For Independent Replication Studies?

First, the replication component. If a researcher chooses to replicate an experiment, knowledge of the specific magnitude of a nonsignificant outcome from that experiment is not a prerequisite. The forefather of experimental design and statistical hypothesis testing, Sir Ronald Fisher, certainly could – and did – replicate his agricultural experiments without betting the farm on a single study's effect sizes. Indeed, Fisher believed that the direction of an effect was only established if he could produce consistent results based on several replications.

As investigators who have collected our share of primary research data, our replication philosophy is similar to Fisher's. And the difference between that philosophy and the one apparently held by Roberts and Henson basically comes down to the difference between the publication of single-shot (one-experiment) studies (their conception of published educational research) and multiple-experiment replication-and-extension studies (our conception). In fact, we contend that much of the fury that characterizes the debates between those who wish to do away with statistical hypothesis testing and those who defend the essence of it (see, for example, Harlow, Mulaik, & Steiger, 1997) would dissipate if researchers refrained from publishing and interpreting single-shot studies.

Results that are statistically significant permit two conclusions. First, they provide evidence that the hypothesis under test (of which the null hypothesis is a special case) is not supported. Second, and less trivially (e.g., Cohen, 1994), they provide evidence of the direction of the difference or relationship. For example, a statistically significant t -test comparing the mean scores of a treatment and control group tells us that it is likely that the treatment group outperformed the control group in the sampled-from populations. Results that are not statistically significant do not permit either of these conclusions.

On the other hand, it is also possible that certain statistically nonsignificant effects are real but too small or fragile to be detected within the parameters of the initial study. In that case, the researcher must decide whether or not the effect is worth pursuing. If so, a replication study is in

order, which may involve changing/tweaking one or more of the initial study's features to make the statistical test of the treatment effect more sensitive – such as by incorporating a larger, more homogeneous, or differently defined sample, strengthening the treatment and/or its implementation, modifying the experimental design and analysis in some way (e.g., through blocking or by including a relevant antecedent variable in the analysis), or improving the psychometric properties of the outcome measure. If the replication study finds the effect to be statistically significant, and if that replication is followed by additional successful replications, then the initially spurned statistically nonsignificant effect will be resurrected.

Is Explicit Effect-Size Reporting A Necessity For Meta-Analytic Literature Syntheses?

Roberts and Henson (2003, p. 226-230) argue (again, at least implicitly) that if multiple-study syntheses are to be conducted, then reporting effect sizes for each experiment allows a meta-analyst to compute an average effect size, as well as to see how the size of the effect may vary as a function of design changes. The argument has been made that single-study investigations should always include effect sizes, even for statistically nonsignificant outcomes, so that meta-analysts will be able to ply their trade using that study's effect-size estimate. What is ignored in this argument is that a meta-analyst does not need the primary researcher to provide explicit effect-size information. As long as the researcher provides sufficient statistics (in the form of either means, variances/covariances, and sample sizes or the associated test statistics) then a competent meta-analyst will be able to calculate the standardized effect-size measures required for multiple-study syntheses (see, for example, Robinson & Levin, 1997).

It is important to note here that we also differ from Roberts and Henson (2003, p. 227-230) in our view of whether research syntheses should consist mostly of meta-analyses or of programmatic replication-and-extension studies. We opt mainly for the latter. We do not disagree that meta-analysis, as conceived by Gene Glass (1976) more than a generation ago, holds great potential for revealing potentially important findings that are shrouded in a literature where

studies are classified only in terms of significant and nonsignificant (see also Hunt, 1997). However, much of what we have witnessed as passing for meta-analyses in the educational and psychological literature since Glass coined the term may be more masking than revealing. For example, certain meta-analytic studies consider all the research on, say, visual aids in learning from text (Robinson, 2002) or phonics/phonemic instruction in beginning reading (Ehri, Nunes, Willows, Schuster, Yaghoub-Zadeh, & Shanahan, 2001) without attending to the type and quality of the materials or the specifics of the instruction. Reporting the average effect sizes in such global meta-analyses may inadvertently misinform the reader.

Finally, we believe that there is another plausible meta-analytic reason to favor single-study authors reporting sufficient summary data rather than the effect-size measures that can be derived from them. It is because (at least in our experience) that it is not unusual for authors to derive effect-size measures incorrectly – in the case of d , often with respect to the particular standard deviation selected for the specific design (e.g., between-subjects, within-subjects, ANCOVA) or question being asked, and in the case of r^2 , by not distinguishing between (or confusing) unconditional and conditional proportions of variance explained (see, for example, Olejnik & Algina, 2000).

This could easily lead an incautious, or unchecking, meta-analyst down the wrong estimation path. Meta-analysts are generally more skilled in the nuances of effect-size types and variations and are less prone to calculating effect sizes incorrectly. Therefore, might it not even be a more judicious research practice/recommendation that meta-analysts routinely calculate effect sizes themselves based on a researcher's provided summary statistics?

Conclusion

In summary, and in contrast to Roberts and Henson's (2003) research philosophy, we argue that in the context of single-study investigations statistically nonsignificant effect sizes should not be reported or interpreted. That is because such reporting/interpreting may lead readers to believe – unwarrantedly – that evidence has been provided

concerning the direction of the effect. Reporting and interpreting effect sizes (with corresponding confidence intervals) in multiple-experiment studies where the effect of interest is replicated (i.e., its direction is confirmed) may provide readers with more useful information concerning the believability and magnitude of the effect, along with the consistency with which it can be produced. Additionally, when a multiple-experiment study is programmatic in nature (i.e., where the design is cumulatively extended to estimate the effect under differing contextual and procedural variations), then reporting effect sizes may be helpful in pinpointing the conditions under which the effect is strongest.

We hope that editors of educational research journals will encourage authors to report work consisting of multiple-experiment studies that replicate and extend initial findings. This is routine procedure in many behavioral science disciplines; and as a clear illustration of editorially practicing what we are preaching, see Levin (1991, p. 5-6). For each experiment conducted, *a priori* α levels, *a posteriori* *p*-values, sample-size and power information, and sufficient statistics should be reported.

In terms of summarizing the multiple experiments, an author may wish to quantify replicated effects, if that serves to inform practitioners who are considering adopting the intervention. At the same time, we are not so naive as to believe that a journal-policy change of this kind will happen overnight. Thus, until the practice of publishing single-shot, non-replicated findings changes, at least we hope that statistically nonsignificant results will be regarded as evidence that the direction of an effect of interest remains undetermined and further research is needed before a more definitive conclusion can be made. Single-study investigators should not routinely provide effect-size estimates for statistically nonsignificant outcomes.

Multiple-study synthesizers can capture those effect sizes from the sufficient statistics reported. Finally, single-study authors should not persist in interpreting or promoting a statistically nonsignificant effect (which includes use of the terms “not quite significant,” “almost significant,” or “approaching significance”), due to the risk of consumers regarding the effect as having been formally screened as believable – when, in fact, no

formal evidence to that effect has been provided. With editorial changes such as these, we strongly suspect that many of educational research’s analysis-and-reporting troubles would simply burst like bubbles!

References

- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.
- Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, A., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel’s meta-analysis. *Reading Research Quarterly*, *36*, 250-283.
- Fisher, R. A. (1960). *The design of experiments*. New York: Hafner Publishing Company.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*, 19-24.
- Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage.
- Knapp, T. R. (2003). Was Monte Carlo necessary? *Journal of Modern Applied Statistical Methods*, *2*(1), 237-241.
- Levin, J. R. (1991). Editorial. *Journal of Educational Psychology*, *83*, 5-7.
- Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.
- Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools*, *5*(2), 43-53.
- Levin, J. R., & O’Donnell, A. M. (1999). What to do about educational research’s credibility gaps? *Issues in Education: Contributions from Educational Psychology*, *5*, 177-229.
- 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.

- Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher*, 29(1), 34-36.
- Olejnik, S., & Algina, J. (2000). Measures of effects size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241-286.
- Onwuegbuzie, A. J., & Levin, J. R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods*, 2(1), 133-151.
- Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, 62, 241-253.
- Roberts, J. K., & Henson, R. K. (2003). Not all effects are created equal: A rejoinder to Sawilowsky. *Journal of Modern Applied Statistical Methods*, 2(1), 226-230.
- Robinson, D. H. (Ed.). (2002). Spatial text adjuncts and learning. Special issue of *Educational Psychology Review*, 14(1).
- Robinson, D. H., Fouladi, R. T., Williams, N. J., & Bera, S. J. (2002). Some effects of including effect size and "what if" information. *Journal of Experimental Education*, 70, 365-382.
- Robinson, D. H., Funk, D. C., Halbur, D., & O'Ryan, L. (in press). The .05 level of significance in educational research: Traditional, arbitrary, sacred, magical, or simply psychological? *Research in the Schools*.
- 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. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, 2(1), 218-225.
- Sawilowsky, S. S., & Yoon, J. S. (2002). The trouble with trivials ($p > .05$). *Journal of Modern Applied Statistical Methods*, 1, 143-144.
- Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development*, 22, 2-6.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25, 26-30.
- Wainer, H., & Robinson, D. H. (in press). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*.