Effects of Sample Size on Effect Size in Systematic Reviews in Education

Robert E. Slavin
Johns Hopkins University and University of York

Dewi Smith
Success for All Foundation

Paper presented at the annual meetings of the Society for Research on Effective Education, Crystal City, Virginia, March 3-4, 2008

This research was carried out under funding from the Institute of Education Sciences, U.S. Department of Education (Grant No. R305A040082). However, any opinions expressed are those of the authors and do not necessarily represent IES positions or policies

It has often been noted by methodologists and authors of systematic reviews of research that studies with small sample sizes tend to have much larger, positive effect sizes than do studies with larger sample sizes. For example, this phenomenon has been described in reviews in medicine by Givens, Smith, & Tweedie (1997), Ioannidis, Cappelleri, & Lau (1998), and Sterne, Gavaghan, & Egger (2000). In reviews of research on approaches to elementary mathematics (Slavin & Lake, in press), secondary mathematics (Slavin, Lake, & Groff, 2007), and secondary reading (Slavin, Cheung, Groff, & Lake, in press), the same phenomenon was documented in education. Median effect sizes for studies with sample sizes less than 250 were two to three times as large as those of larger studies.

The importance of the issue of large effect sizes in small studies has been heightened by controversies over the procedures of the What Works Clearinghouse (WWC). In order to receive the highest rating given by the WWC, "positive effects," programs must have at least one study in which students, classes, or schools were randomly assigned to treatments and in which there were statistically significant positive effects on important outcomes. A second matched or randomized study with significant positive effects is also required, but since matched studies are more common, it is the presence or absence of a randomized study that is usually decisive. As noted by Slavin (2008), many programs have a single randomized evaluation with a very small sample size, and this single study is all that qualifies the program for a top rating. For example, a study by Williams (1986) with only 46 students qualified *Saxon Math* for the only "positive effects" rating in the WWC middle school math review, even though several matched studies had median effect sizes near zero. Similarly, the only program to receive a "positive effects" rating in the English language learners topic report was *Peer Tutoring and Response Groups*, which qualified based solely on a small randomized experiment involving only 46 students.

Much as an emphasis on randomized experiments in program evaluation syntheses is appropriate, there are other methodological factors that may be as important as random assignment, and need to be taken into account in the same way. Sample size is one of these factors.

The present study uses data from the elementary and secondary mathematics reviews by Slavin & Lake (in press) and Slavin et al. (2007) to further explore the effects of sample size on effect size in program evaluations in education. Finally, the findings are used to suggest solutions for the problem of sample size bias in systematic reviews in education.

Why Should Small Samples Produce Large Effect Sizes?

The main reason advanced by methodologists for small sample bias in experimental studies in medicine is that publication bias is more serious in small sample research than in studies with large samples (Givens et al., 1997; Sterne et al., 2000; Rothstein, Sutton, & Borenstein, 2005). In essence, a small study that obtains a negative or nonsignificant effect is unlikely to be published or even to produce a technical report that can later be located by reviewers. Journal editors and the investigators themselves are likely to reason that when an underpowered study fails to find significant differences, little is learned. In contrast, when a large study finds nonsiginificant effects, this is taken as potential evidence that the treatment is in fact ineffective, so the report is more likely to be published or otherwise made available. Further, large studies are more likely to have been funded by government or private funders and are therefore more likely to at least produce a technical report that can be located by reviewers.

It takes a larger effect size to produce statistical significance in a small study than in a large study. Therefore, if studies with significant differences are more likely to be published or otherwise findable, then large studies with both small and large positive effects are likely to be found but small studies will likely be found only if their effect sizes are large. This problem is exacerbated by the fact that effect sizes in small studies are more likely to be highly variable than is the case in large studies, especially if (as is usually the case in school research) students are clustered in a small number of classes or schools, which may add in a consistent direction to program effects. This variability means that any set of small studies is likely to contain a disproportionate number of very positive (and very negative) effect sizes, so if the significant positive effects are more likely to come to light, small study bias is likely to be exacerbated.

To the degree that small sample size bias is due to differential publication bias, then there is a paradox. Each individual small study, especially if random assignment is used, may be entirely free of bias, yet a collectivity of small studies could nevertheless suffer from serious bias.

However, there may be factors beyond publication bias that contribute to small sample bias. Kjaergard, Villumsen, & Gluud (2001) note that in medical research, small studies tend to be of lower methodological quality. In educational research, the most important source of small study bias, may be what Cronbach et al. (1980) called "superrealization bias." This refers to the fact that in a small experiment, experimenters are able to monitor the quality of implementation, provide additional assistance, or create unrealistic conditions that could never be replicated on a large scale. Many of the small studies that caused various programs to qualify for top ratings in the WWC are clear examples of superrealization bias. For example, in studies of phonemic awareness software called *Daisy Quest*, project assistants sat every day with small groups of children to help them with the software (e.g., Torgersen et al., 1999). In a study of a tutoring model called *SpellRead* (Rashotte, MacPhee, & Torgesen, 2001), project assistants, not teachers, provided the tutoring. Further, knowing that small studies need large effect sizes to be statistically significant, designers of small studies may be more likely than designers of large studies to use, for example, measures closely aligned with experimental (but not control) curricula or to use other procedures that bias the study in the direction of large positive effects. The designer of a large study may not feel the need to use such procedures and may not be able to afford to do them on a large scale in any case.

Methods

In order to examine the effects of sample size on effect size we used data from two systematic reviews of evaluations of mathematics programs, one of elementary programs by Slavin & Lake (in press) and one of secondary programs by Slavin et al. (2007). In both reviews, a common set of inclusion requirements was used, as follows:

1. The studies evaluated mathematics programs. Studies of variables, such as ability grouping, block scheduling, and single-sex classrooms were not reviewed.

2. The studies compared children taught in classes using a given mathematics program to those in control classes using an alternative program or standard methods.

3. Studies could have taken place in any country, but the report had to be available in English. The report had to have been published in 1970 or later.

4. Random assignment or matching with appropriate adjustments for any pretest differences (e.g., analyses of covariance) had to be used. Studies without control groups, such as pre-post comparisons and comparisons to "expected" gains, were excluded.

5. Pretest data had to be provided, unless studies used random assignment of at least 30 units (individuals, classes, or schools) and there had to be no indications of initial inequality. Studies with pretest differences of more than 50% of a standard deviation were excluded.

6. The dependent measures included quantitative measures of mathematics performance, such as standardized mathematics measures. Measures of math objectives inherent to the program (but unlikely to be emphasized in control groups) were excluded.

7. A minimum treatment duration of 12 weeks was required.

A total of 85 elementary and 100 secondary studies going back as far as 1970 met the standards of the reviews. In each case, effect sizes were estimated, using procedures from Lipsey & Wilson (2001), with posttest effect sizes adjusted for pretest effect sizes (see the study reviews and Slavin, 2008, for details of these procedures). No limitations were placed on study size although both reviews required sample sizes of at least 250 students to qualify a given program for the highest rating, "strong evidence of effectiveness."

Sample size information was extracted from the primary studies, except in a few cases in which student sample sizes were estimated (at 20 students per class) in studies that presented sample sizes as classes rather than students.

As is illustrated in Figure 1, sample sizes across studies had a strong negative skew. Sample sizes of qualifying studies ranged from 30 to almost 40,000, but 25% of studies had samples sizes of 100 or less and 52% included 250 students or less. To reduce the influence of extreme scores, sample sizes were recoded into 8 categories. Table 1 shows the categories and the numbers of subjects in each.

Figure 1
Sample Sizes in Mathematics Studies

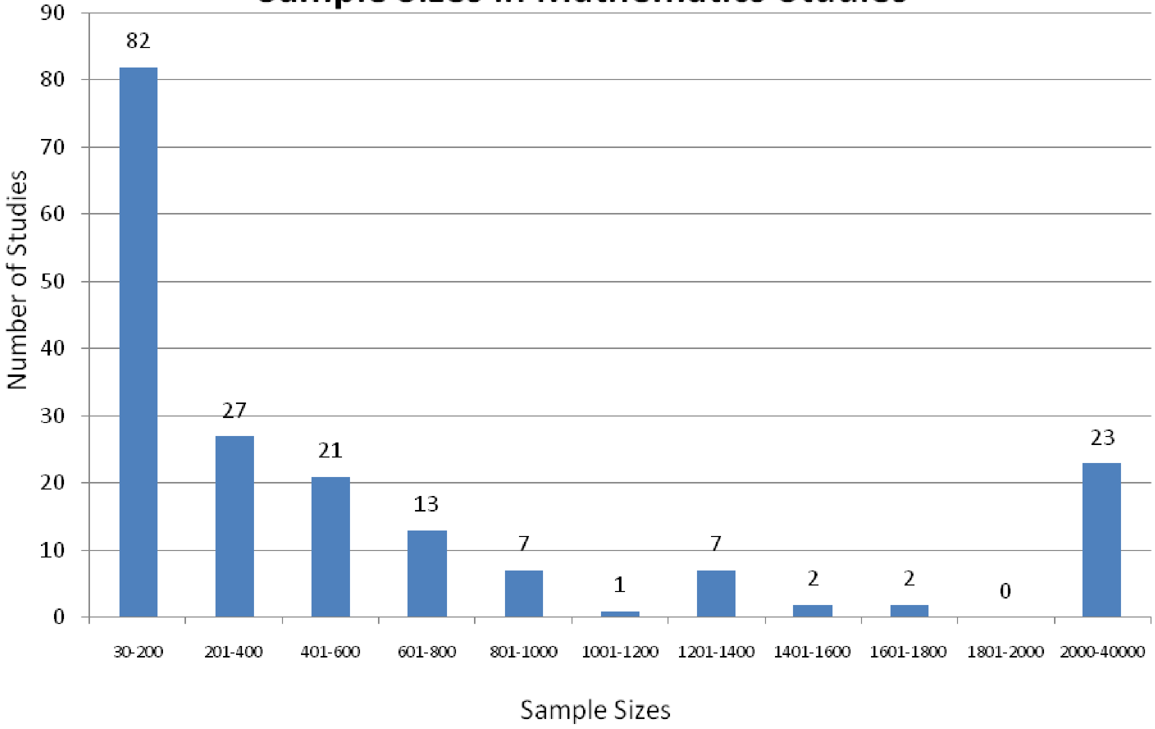


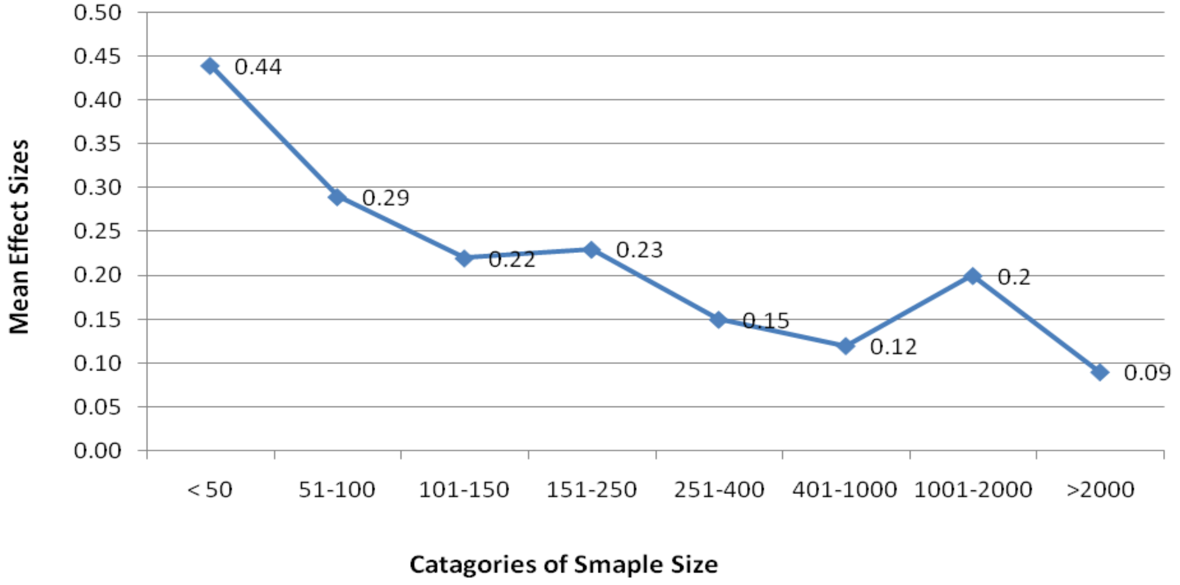
Table 1
Total Sample Size

Recode	Range	Number of Studies
1	Up to 50	10
2	51-100	36
3	101-150	18
4	151-250	31
5	251-400	14
6	401-1000	41
7	1001-2000	12
8	2001 or more	23
TOTAL		185

A Pearson correlation was computed between recoded sample size and pretest-adjusted effect sizes. The overall correlation was -.28, p<.001. Figure 2 shows that effect sizes for very small studies, those with sample sizes below 50, averaged +0.44, while those with sample sizes of more than 2000 had effect sizes of only +0.09. Studies with sample sizes below the median of about 250 had a mean effect size of +0.27 while those with larger sample sizes had an average effect size of +0.13, less than half as much.

In this sample of 185 studies, bias due to small sample size was far greater than bias due to lack of random assignment. True randomized experiments (n=34) had a mean effect size of +0.24, and randomized quasi-experiments (RQEs) (n=28), in which random assignment took place at the school or classroom level but analysis was at the individual level, had a mean effect size of +0.29. Matched studies (n=124) had a mean effect size of +0.17. None of the differences between these means are statistically significant. The non-significantly larger effect sizes in the randomized studies may be accounted for in part by the fact that randomized studies had smaller sample sizes. True randomized experiments had an unrecoded mean sample size of 212 in comparison to 558 for RQEs and 2570 for matched studies. Effect sizes for large studies, with sample sizes of 250 or more, were similar for large randomized studies (n=8, ES=+0.11), large RQEs (n=15, ES=+0.16) and large matched studies (n=67, ES=+0.13). These effect sizes are not significantly different from each other.

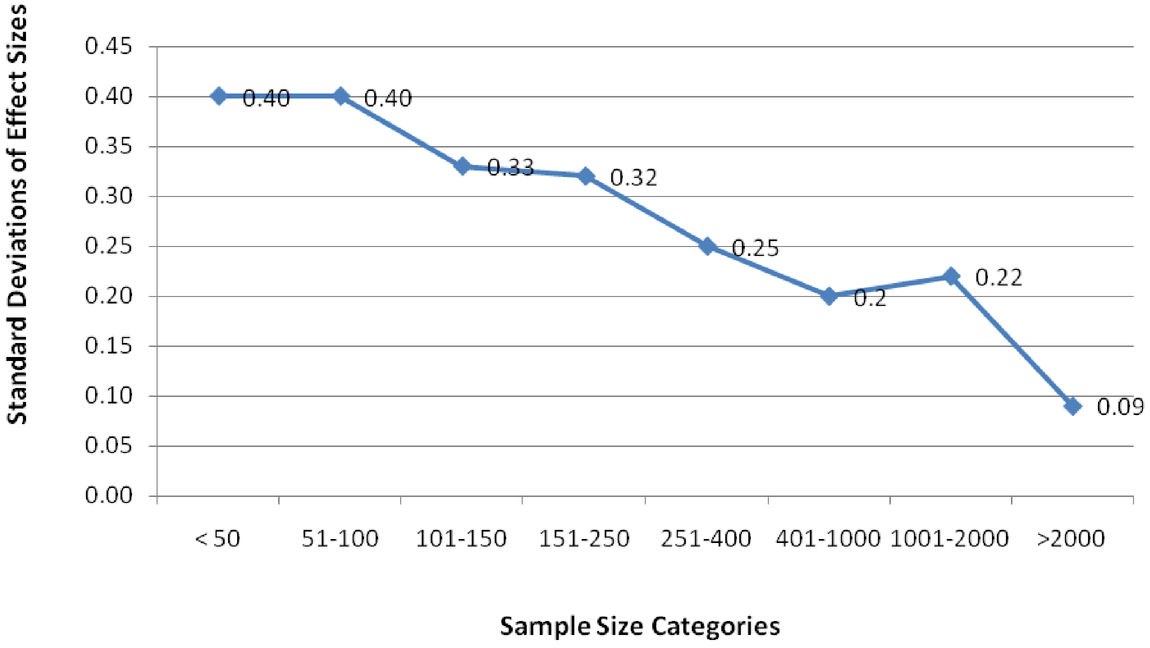
Figure 2
Relationship Between Sample Size and Effect Size



Within the randomized studies (n=34) the relationship between sample size and effect size was particularly strong, with a correlation of -.40, p<.02.

The Slavin & Lake (in press) and Slavin et al. (2007) syntheses review a wide variety of curricula, computer-assisted instruction programs, and instructional process approaches (such as cooperative learning). To eliminate confounding between research designs and types of interventions, we looked at the studies of computer-assisted instruction, the largest set of studies (n=75) on a set of fairly similar interventions. Within this group, the correlation between recoded sample size and effect size was -.25, p<.03.

Figure 3
Relationship Between Sample Size and Standard Deviations of Effect Sizes



As predicted, the variability of effect sizes also diminished with increasing sample size. As shown in Figure 3, the standard deviations of effect sizes averaged 0.40 in both of the smallest categories of sample size, ≤ .50 (n=10) and 51-100 (n=36). In the largest studies, with sample sizes of more than 2000 (n=23), the standard deviation of effect size estimates was only .09. This reduction in standard deviations as sample size increases tracks closely on reductions in the mean effect sizes themselves. It also suggests that as sample sizes increase, effect sizes become more reliable and less likely to be artifacts of unequally distributed school, teacher, or class effects.

How Should Sample Size be Treated in Systematic Reviews?

The research presented in this paper justifies close attention to the issue of sample size in systematic reviews in education. Unfortunately, studies with small and large samples are not randomly distributed in the literature; some programs are usually or always evaluated in small studies, while others are usually or always evaluated in large ones. This means that programs typically evaluated in small studies may greatly overstate mean program impacts.

How can the impact of small sample bias be minimized in program evaluation reviews in education? The evidence does not justify ignoring the results of small studies, but it does argue that all other things being equal, the findings of large studies should be considered as more conclusive evidence of the effects of a given program than the findings of small studies.

One possible solution is to weight effect sizes by their sample sizes in arriving at average effect sizes for a given program. This was done by Borman, Hewes, Overman, and Brown (2003) in a review of research on comprehensive school reform models and by Slavin, Cheung, Groff, & Lake (in press) in a review of research on middle and high school reading models. One danger of weighting is that it can give too much weight to a few enormous studies, and for this reason, Slavin et al. (in press) use a cap weight of 2500. Weighting by log transformations of sample sizes may serve the same function.

Weighting by sample size is not sufficient, however, because a mean effect size for a given program could still be based on very small studies. For this reason, Slavin & Lake (in press), Slavin et al. (2007), and Slavin et al. (in press) set a sample size criterion for categorization of programs in the top two ratings: “Strong evidence of effectiveness” and “moderate evidence of effectiveness.” Both require at least two studies with sample sizes of at least 250, or a larger number of smaller studies with collective sample sizes of 500.

In addition to strategies for weighting, it may be useful to set a minimum sample size for study inclusion. The What Works Clearinghouse excludes studies with only one school or classroom per treatment, on the basis that such studies confound school or class effects with treatment effects. In Best Evidence Synthesis reviews, studies with one teacher or classroom per treatment are excluded for the same reason. These policies remove the very smallest of studies.

There may be many other ways of emphasizing larger studies over smaller ones, and future research might model consequences of various methods. What is clear, however, is that this issue must be attended to. Small sample bias has significant potential to undermine the scientific validity and the practical utility of program effectiveness reviews in education.

Although this paper focuses on the statistical impact of small sample bias, it is important to note that large sample studies are desirable for external validity as well. Small studies within a small set of schools and classrooms risk being context-specific. The larger the number of students, teachers, classes, and schools involved in a given study or set of studies, the more confidence we can have that the results are generally applicable beyond a specific site or situation. As noted earlier, “superrealization” due to extra assistance or attention to treatment classes becomes less likely in large studies, as it is practically very difficult to provide non-replicable help to large sets of schools, especially if large numbers of students imply many schools and classrooms across diverse contexts, large studies can test programs as they would be disseminated or scaled up in real life.

Large studies do not guarantee internal or external validity, but if other design features are also optimal, large studies should be emphasized. When educators have a broad range of programs to choose from, all of which have been successfully evaluated in large, randomized experiments with valid measures, then evidence-based reform can become a practical reality for all schools.

References

Borman, G.D., Hewes, G.M., Overman, L.T., & Brown, S. (2003) Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73* (2), 125-230.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R.O., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco: Jossey-Bass.

Givens, G. H., Smith, D. D., and Tweedie, R.L. (1997) Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science,* 12, 221—250.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Rashotte, C. A., MacPhee, K., & Torgesen, J. K. (2001). The effectiveness of a group reading instruction program with poor readers in multiple grades. *Learning Disabilities Quarterly, 24*(2), 119–134

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments.* Chichester, UK: John Wiley.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher.*

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2007). *Effective reading programs for middle and high schools: A best-evidence synthesis*. Baltimore, MD: Johns Hopkins University, Center for Data-Driven Reform in Education.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (in press). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*.

Slavin, R.E., & Lake, C. (in press). *Effective programs in elementary math: A best evidence synthesis*. Manuscript submitted for publication.

Slavin, R.E., & Madden, N.A. (2008, March). *Understanding bias due to measures inherent to treatments in systematic reviews in education.* Paper presented at the annual meeting of the Society for Research on Effective Education, Crystal City, VA

Slavin, R.E., Lake, C., & Groff, C. (2007). *Effective programs in middle and high school math: A best evidence synthesis*. Manuscript submitted for publication.

Sterne J., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in literature. *Journal of Clinical Epidemiology, 53,* 1119–1129.

Taylor, S., & Tweedie, R. (1998). *A non-parametric "trim and fill" method of assessing publication bias in meta-analysis*. Denver: University of Colorado Health Sciences Center.

Torgesen, J., Wagner, R., Rashotte, C., Rose, E., Lindamood, P., Conway, T., et al. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91,* 579–593.

Williams, D. D. (1986). *The incremental method of teaching Algebra I*. Unpublished research report, University of Missouri, Kansas City.