Abstract Title Page
Not included in page count.

Title: How Methodological Features Affect Effect Sizes in Education

Authors and Affiliations: Alan Cheung, The Chinese University of Hong Kong alancheung@cuhk.edu.hk; Robert Slavin, Johns Hopkins University rslavin@jhu.edu

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Throughout the federal government, evidence is playing an increasing role in policy (see Buck & McGee, 2015; Haskins, 2014; Nussle & Orszag, 2014; Slavin, 2013). In particular, certain federal grants are restricted to applicants who can already demonstrate evidence of effectiveness for the programs or practices they are proposing, or at least agree to subject new or untested ideas to rigorous evaluation.

The increasing influence of evidence in education policy contributes urgency to the need to have clear, enforceable, and difficult-to-game standards of evidence indicating that educational programs have acceptable levels of evidence. As of this writing, there are four main sources of definitions for programs with enough evidence to be considered effective. The most influential is the U.S. Department of Education's What Works Clearinghouse (WWC), which has detailed descriptions of standards for inclusion of individual studies as well as procedures for pooling study outcomes (Song & Herman, 2010; WWC, 2014). The WWC categorizes programs as being acceptable "without reservations" or "with reservations," and if studies meet these categories, outcomes are considered positive if they are statistically significant at the proper level of analysis.

In 2012, the Education Department General Administrative Regulations (EDGAR) added definitions of "strong" and "moderate" levels of evidence.

Social Programs That Work (http://evidencebasedprograms.org) uses stringent standards to identify programs that clearly meet "top tier" standards, including successful, replicated randomized evaluations. Programs can be "near top tier" with a single successful evaluation.

Another effort to summarize the findings of educational program evaluations is the Best Evidence Encyclopedia, or BEE (www.bestevidence.org), created at Johns Hopkins University. BEE standards are similar to those of the WWC, but place much more emphasis on issues such as measures aligned with experimental but not control group content. Also, the BEE carries out meta-analyses to determine average effect sizes in evaluating programs and categories of programs, and study authors publish these meta-analyses in peer-reviewed journals.

The Problem: Methodology Correlates With Outcomes

The development of WWC, EDGAR, SPW, and BEE standards, and WWC, SPW, and BEE reviews, are essential underpinnings for evidence-based reform, because they provide policy makers with some assurance that if they encourage use of proven programs, there will in fact be programs that will meet rigorous standards of evaluation and will show positive impacts.

However, in the course of creation of these research syntheses, several nettlesome issues have come up, and these must be resolved or at least understood if evidence-based reform is to have its desired impact on policy and practice. The problem is that certain methodological features are correlated with study effect sizes. All of these correlations may indicate the presence of bias. For example, Slavin & Madden (2011) examined effect sizes of studies that met the standards of the What Works Clearinghouse with or without reservations. Studies were identified that used measures inherent to the experimental treatments, as when experimental students were taught specific content or skills that the control group was not taught, and the measure focused on the content taught to the experimental but not the control group. These same studies also administered tests that were not inherent to the treatment, such as standardized measures, specialized measures made by someone other than the study authors, or measures held to cover the content taught equally in experimental and control groups. The differences in effect sizes

between the inherent and non-inherent measures were striking. Across seven WWC-accepted math studies, the mean effect size was +0.45 for measures with treatment-inherent measures and -0.03 for measures used in the same studies that were not inherent to the treatment. Across 10 WWC-accepted early reading studies, the effect sizes were +0.51 and +0.06, respectively. Similarly, Slavin & Smith (2009) found substantial differences in effect sizes between studies with large and small sample sizes, with an average effect size of +0.44 for studies with fewer than 50 subjects, +0.29 for studies with 51-100 subjects, and +0.09 for studies with sample sizes of more than 2000.

Numerous reviewers have noted substantial differences between published and unpublished articles (e.g., Glass, McGaw, & Smith, 1981; Lipsey & Wilson, 1993). These well-known differences have led most meta-analysts, as well as the WWC, SPW, and BEE to insist on exhaustive searches for all studies on a given topic, including dissertations, technical reports, and other "gray literature."

A recent review of studies of learning strategies interventions by deBoer, Donker, & van der Werf (2014) found that studies using non-standardized tests obtained higher effect sizes than those using standardized tests, as did studies in which the intervention was delivered by the researcher or associates (rather than ordinary teachers). This review did not, however, find significant differences between studies using random (vs. matched) assignment to conditions or between longer and shorter interventions.

The impacts of these differences according to study methodology are no longer academic. If, for example, large, randomized experiments characteristically produce much lower effect sizes than small, matched ones, then it may be unfair to compare effect sizes from these two categories of studies as though they were indicators of substantive differences between the effect sizes of different programs or types of programs. Not only could this mislead educators and policy makers about which programs truly work, but it could encourage publishers or developers to "game the system" by using certain methods and avoiding others to make their programs appear more effective than they are (see Baron, 2003).

For scientific as well as pragmatic reasons, it is important to know how research designs affect effect sizes in program evaluations. Yet research on the relationship between methodology and effect size is sparse, has been focused within reviews of particular subjects or interventions, and has involved relatively few studies. Also, some studies that have evaluated relationships between methodologies and effect sizes have initially included such a broad range of studies that aspects of methodology of no interest to practice cause certain related factors to appear to affect effect sizes, as when reviews include one-hour, tightly controlled lab studies and then conclude that brief interventions with very small samples have extraordinarily large effect sizes, relationships that may or may not be true of experiments involving real classrooms over significant time periods.

Purpose / Objective / Research Question / Focus of Study:

As evidence-based reform becomes increasingly important in educational policy, it is becoming essential to understand how research design might contribute to reported effect sizes in experiments evaluating educational programs. The purpose of this study was to examine how methodological features such as types of publication, sample sizes, and research designs affect effect sizes in experiments.

Setting: Not applicable

Population / Participants / Subjects:
Students in PK-12 in reading, math, and science classes using an innovative program and control classes using an alternative program or standard methods.

Intervention / Program / Practice:
The following methodological features were extracted from each of 645 studies in 12 reviews carried out for the Best Evidence Encyclopedia (BEE): type of publication (published vs unpublished), size of the sample (small, N ≤250 vs large, N>250), research design (randomized vs matched), and outcome measures (experimenter-made vs. independent).

Research Design:
In the BEE reviews from which the 645 studies were derived, a consistent set of study inclusion criteria was used, with just a few variations. These criteria were as follows:

1. The studies evaluated reading, mathematics, or science programs designed to improve student achievement.
2. The studies involved students in grades PK-12.
3. The studies compared students taught in classes using an innovative program to those in control classes using an alternative program or standard methods.
4. Studies could have taken place in any country, but the report had to be available in English.
5. 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" scores, were excluded.
6. Pretest data had to be provided, unless studies used random assignment of at least 30 units (individuals, classes, or schools) and there were no indications of initial inequality. Studies with pretest differences of more than 50% of a standard deviation were excluded.
7. The dependent measures included quantitative measures of student performance, such as standardized outcome measures. Experimenter-made measures were accepted if they were comprehensive measures of reading, mathematics, or science, which would be fair to the control groups, but measures of objectives inherent to the program (but unlikely to be emphasized in control groups) were excluded (see Slavin & Madden, 2011).
8. A minimum study duration of 12 weeks was required. This requirement was intended to focus the review on practical programs intended for use for the whole year, rather than brief investigations. Studies had to have at least two teachers in each treatment group to avoid compounding of treatment effects with teacher effect.
9. Studied programs had to be replicable in realistic school settings. Studies providing experimental classes with extraordinary amounts of assistance (e.g., additional staff in each classroom to ensure proper implementation) that could not be provided in ordinary applications were excluded.

Data Collection and Analysis:
In order to investigate the relationships between study methodological features and effect sizes, we analyzed 645 studies that met the standards of inclusion for any of 12 reviews written for the Best Evidence Encyclopedia and (in most cases) published in review journals. The reviews cover programs in elementary and secondary math, elementary and secondary science, and elementary

and secondary reading, as well as a review of elementary reading programs for struggling readers and a review of early childhood education. Studies included in reviews focusing on technology applications in reading and math were also included. Comprehensive Meta-Analysis software Version 2 (Borenstein, Hedges, Higgins, & Rothstein, 2005) was used to carry out all statistical analyses such as Q statistics and overall effect sizes.

Findings / Results:
The findings suggest that effect sizes are roughly twice as large for published articles, small-scale trials, and experimenter-made measures, than for unpublished reports, large-scale studies, and independent measures, respectively. In addition, effect sizes are significantly higher in quasi-experiments than in randomized experiments.

Conclusions:
Based on the findings of our analyses, it is clear that researchers as well as policy makers need to take into account research design, sample size, measures, and type of publication before comparing effect sizes from program evaluations. Some specific recommendations are as follows.

1. In meta-analyses and other quantitative syntheses, reviewers should search for all studies that meet well-justified standards, regardless of whether or not the studies are published.
2. Researchers should use cluster randomized trials whenever possible. When they are not possible or when it is clear that effect sizes are potentially meaningful but the sample size (of clusters) is too small to reach statistical significance, researchers should be encouraged to pool similar studies to build up sample size over time.
3. In reviews of program evaluations intended to inform policy and practice, reviewers should eliminate researcher-developed measures. These greatly overstate effect sizes. However, this is not to say that only standardized tests should be used. Evaluators might choose valid non-standardized tests made by various organizations, tests developed by researchers other than themselves, or tests from other states or other countries, as long as the tests equally cover experimental and control objectives.
4. Policy makers and educators should insist on high-quality research to validate promising programs, even if this means reducing the number of programs available in a given area. It is apparent that small and low-quality studies can greatly overstate program impacts, or at a minimum allow great variations in outcomes. If important decisions are to be made based on evidence, that evidence should be as convincing as possible.

Appendix A. References.

Baron, J. (2003). *How to assess whether an educational intervention has been "proven effective" in rigorous research.* Washington, DC: Coalition for Evidence-Based Policy.

Borenstein, N., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). *Comprehensive meta-analysis (Version 2)*. Englewood, NJ: Biostat.

Buck, S. & McGee, J. (2015). *Why government needs more randomized control trials: Refuting the myths.* Houston: Laura and John Arnold Foundation.

de Boer, H., Donker, A. S., & van der Werf, M. P. C. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research, 84*(4), 509-545.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hill, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation form meta-analysis. *American Psychologist, 48*, 1181-1209.

Nussle, J., & Orszag, P. (Eds.). (2014). *Moneyball for government.* Washington, DC: Disruption Books.

Slavin, R. E. (2013). Overcoming the four barriers to evidence-based education. *Education Week 32* (29), 24.

Slavin, R. E., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*: 370-380.

Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic review in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500-506.

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse. *Educational Evaluation and Policy Analysis, 32* (3), 351-371.

What Works Clearinghouse (2013). *Procedures and standards handbook (version 3.0).* Washington, DC: Author.