Best Evidence
Encyclopedia (BEE)
Empowering Educators with Evidence on Proven Programs
www.bestevidence.org

How Methodological Features Affect Effect Sizes in Education

Alan C. K. Cheung

The Chinese University of Hong Kong

Robert E. Slavin

Johns Hopkins University

September, 2015

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Research and Reform in Education (CRRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Abstract

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 article is to examine how methodological features such as types of publication, sample sizes, and research designs affect effect sizes in experiments. A total of 645 studies from 12 recent reviews of evaluations of reading, mathematics, and science programs were studied. The findings suggest that effect sizes are roughly twice as large for published articles, small-scale trials, and experimenter-made measures, than for unpublished documents, large-scale studies, and independent measures, respectively. In addition, effect sizes are significantly higher in quasi-experiments than in randomized experiments. Explanations for the effects of methodological features on effect sizes are discussed, as are implications for evidence-based policy.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Introduction

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. In the Departments of Health and Human Services, Labor, and Education, among others, there are now "tiered evidence" competitions, in which applicants with strong evidence of effectiveness can apply for substantial funding to go to scale, those with smaller amounts of evidence can apply for funding for a large-scale evaluation, and those with a promising idea can apply to further develop and formatively evaluate their program.

In the U.S. Department of Education, the prototype for this tiered evidence strategy is Investing in Innovation (i3), which makes "scale-up," "validation," or "development" grants based on the level of evidence. Further, programs with strong or moderate evidence of effectiveness are increasingly being favored in other federal educational programs. In School Improvement Grants (SIG) for very low-achieving schools, Congress added a category of services schools could choose, "evidence-based whole-school reform," which required at least one rigorous evaluation of a whole-school reform model. The Department of Education invited applications from providers, and selected four that met this standard. Similarly, to receive Supporting Effective Educator Development (SEED) grants under Title II for professional development, applicants have to show evidence from at least one rigorous experiment that their program is effective.

A bipartisan coalition of lawmakers is forming to support the move toward evidence-based reform in government, influenced by an organization called Results for America, which

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

has a campaign to get government to "play Moneyball" (i.e., to use evidence to guide decisions [Nussle & Orszag, 2014]). Further, investments by i3, the Institute for Education Sciences (IES), the National Science Foundation (NSF), the Education Endowment Fund (EEF) in the U.K., and other agencies, are producing a steadily growing set of proven programs, greatly facilitating the argument that such programs should be favored when appropriate in federal funding. These and other developments suggest that evidence-based reform may continue to grow in importance in policy and practice.

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. That is, if assignment and treatment were at the school level, analysis must be at the school level, usually using hierarchical linear analysis (HLM; Raudenbush & Bryk, 2002) or comparable methods. Also, programs can be considered effective by the WWC if they do not analyze data at the proper level but produce an effect size of +0.25 or more (WWC, 2014).

In 2012, the Education Department General Administrative Regulations (EDGAR) added definitions of "strong" and "moderate" levels of evidence. EDGAR definitions draw on WWC standards for including specific studies, but focus more on the effectiveness of programs. To be

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

considered "strong" according to EDGAR, a program must have at least one study that meets WWC standards without reservations or two that meet WWC standards with reservations, at least one significant positive effect on a relevant outcome, 350 students or 50 classes or schools, and multiple sites. To meet the "moderate" standard a program must have at least one study that meets WWC standards and at least one significant positive outcome, but there are no requirements for sample sizes or multiple sites.

Social Programs That Work (SPW) (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. BEE standards are described in detail later in this article.

In addition to the general standards applied by the WWC, EDGAR, SPW, and the BEE, federal legislation and requests for proposals that require or provide preference points for proven programs define their own standards, which are similar but not identical to WWC, EDGAR, SPW, or BEE standards.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

The Problem: Methodology Correlates With Outcomes

The development of WWC, EDGAR, SPW, and BEE standards and 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 WWC 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.

Study sample size has also been found to strongly impact effect sizes. Slavin & Smith (2009) found substantial differences in effect sizes between studies with large and small sample

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

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

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

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. For example, many 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.

Methods

In order to investigate the relationships between study methodological features and effect sizes, we analyzed all 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. Table 1 shows the reviews and information about numbers of studies and breakdowns of studies in key methodological categories. At the bottom of the table is information on the full set of studies. Note that because of overlaps (e.g., studies in the technology reviews often overlapped those in the comprehensive reviews), the study N's from each review add up to a larger number than the grand total.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

==========

TABLE 1 HERE

==========

Study Inclusion Criteria

A consistent set of study inclusion criteria was used across all studies, 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 prekindergarten-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

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

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. Brief studies may not allow programs to show their full effect. On the other hand, brief studies often advantage experimental groups that focus on a particular set of objectives during a limited time period while control groups spread that topic over a longer period. Studies with brief treatment durations that measured outcomes over periods of more than 12 weeks were included, however, on the basis that if a brief treatment has lasting effects, it should be of interest to educators. The one exception to the 12-week requirement was elementary science, where there were numerous studies of science units (e.g., electricity) that lasted less than 12 weeks.
9. Studies had to have at least two teachers in each treatment group to avoid compounding of treatment effects with teacher effect.
10. 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.

A total of 645 studies from these 12 reviews were included in our final analysis (studies included in multiple reviews were only used once). In each of these reviews, effect sizes were computed as the difference between experimental and control individual student posttests after adjustment for pretests and other covariates, divided by the unadjusted posttest pooled standard deviation (SD). Procedures described by Lipsey & Wilson (2001) and Sedlmeier & Gigerenzor

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

(1989) were used to estimate effect sizes when unadjusted standard deviations were not available, as when the only standard deviation presented was already adjusted for covariates or when only gain score SD's were available. If pretest and posttest means and SD's were presented but adjusted means were not, effect sizes for pretests were subtracted from effect sizes for posttests. F ratios and t ratios were converted to effect sizes when means and standard deviations were not reported.

The following methodological features were extracted from each of the 12 reviews: 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). 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.

==========

TABLE 2 HERE

==========

Results

Publication Bias

Across the 12 reviews, there were a total of 262 published articles and 383 unpublished dissertations and technical reports (often collectively referred to as "gray literature). As indicated in Table 2, the overall effect sizes for published articles and unpublished reports were +0.30 and +0.16, respectively. The Q-value (Q_B =58.47, df=1, and p<0.00) clearly indicates publication bias in this set of studies. The average effect size of published studies was about

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

twice as large as that of studies found in technical reports or dissertations. The findings are consistent with previous studies in social science, psychological science, and medicine (Ferguson & Heene, 2012; Glass, McGraw, & Smith; 1981; Hopewell, McDonald, Clarke, & Egger, 2007; Lipsey & Wilson, 1993; McAuley, Pham, Tugwell, Moher, 2000). For example, when examining 11 meta-analyses published between 1976 and 1980 in the areas of psychotherapy and counseling, Glass, McGraw, and Smith (1981) discovered that the average effect sizes found in published journals were larger than those found in unpublished reports such as theses and dissertations. Similarly, in their study, Lipsey and Wilson (1993) found that there was a significant difference between the mean treatment effect sizes derived from published studies and those derived from unpublished studies in a large body of meta-analyses of psychological, educational, and behavioral treatment research. The mean effect sizes for published studies and unpublished studies from the 92 meta-analyses they examined were +0.53 and +0.39, respectively. When examining publication bias in the health care field, Hopewell, McDonald, Clarke, and Egger (2007) found that on average the effect sizes for published trials were about 9% larger than those found in the gray literature.

The findings that published studies yield larger effect sizes than unpublished studies should come as no surprise. First, positive and significant results are perceived more favorably by reviewers and editors alike, making studies that present them more likely to be published (Cook et al., 1993; Hopewell, Clarke, & Mallett, 2005). Atkinson, Furlong, & Wampold (1982) carried out a study to examine this issue. Consulting APA editors were asked to review manuscripts that were identical in all aspects except that the findings were either statistically significant or not significant. What they found was that manuscripts with significant results were more than twice as likely to be recommended for publication compared to those with non-significant findings. In

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

addition, the "reviewers" reported that studies with significant results had better research designs than those with non-significant results, though the methods used were in fact exactly the same in the two sets of studies.

Further, studies with small or negative insignificant results are often shelved by researchers or program developers themselves before submitting them for publication. This is called the "file drawer effect" (Glass et al., 1981). Ferguson and Heene (2012) believed that "publication bias may be more pernicious at the level of the individual scholar than it is at the journal level" (p. 556).

There is some debate as to whether gray literature should be included in research reviews and meta-analyses. Some meta-analysts only included published articles, arguing that since published materials have to go through a rigorous peer-reviewed process, their quality is generally higher than that of unpublished works. However, Lipsey and Wilson (2001) disagreed, stating "this rationale is generally not very convincing. In many research areas, unpublished material may be as good as the published and in any event, the decision is better made on the basis of explicit methodological criteria than by using publication status as a proxy" (p.19). In a similar vein, Jefferson, Alderson, Davidoff, & Wager (2003) argued that although a peer review process can help ensure the scientific quality of studies, it still can be open to bias since it still relies on the opinions of expert reviewers. Our current findings show that excluding studies in the gray literature could bias results of a meta-analysis. Clearly, it is critical to screen and assess both published and unpublished works using rigorous inclusion criteria to guarantee the quality of included studies.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Researcher-Made vs. Independent Measures

As noted earlier, Slavin & Madden (2011) found substantial differences in effect sizes within WWC-accepted studies between measures inherent to (or over-aligned with) the treatment as opposed to those that were independent of the treatment. This issue has often been posed in previous research as researcher-made vs. standardized tests, and it has been frequently found that researcher-made tests are associated with much higher effect sizes than are standardized tests. For example, Scammacca et al. (2007) carried out a meta-analysis to examine effective reading interventions for adolescent struggling readers. A total of 33 studies were included in this meta-analysis, and the overall effect size was +0.95. However, studies that used standardized measures had smaller effect sizes than those that used researcher-developed measures, which were often closely aligned with the specific intervention tested. In the 11 studies that used standardized, norm-referenced measures, the average effect size was only +0.42. Similar results were found in other research reviews that compared studies that used researcher-developed measures and standardized tests, which often measure more generic skills (Edmonds et al, 2009; Wanzek, Wexler, Vaughn, & Ciullo, 2010). In their meta-analysis, Edmonds and her colleagues (2009) examined the effects of reading interventions for older struggling readers. In addition to the overall effect size, they also examined whether there was a difference in terms of effect sizes of studies that used researcher-developed or curriculum-based measures and studies that used standardized measures. The results suggest that average effect size of studies that used researcher-developed measures (ES=+1.19) were two and a half times larger than that of studies that used independent outcome measures (ES=+0.47). In a review of research on the effects of computer technology on students' mathematics learning, Li and Ma (2010) found that out of the 46 included studies, half of the outcome measures used were teacher-made or researcher-

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

developed. The effect size of studies that used non-standardized measures (ES=+0.86) was much larger than that of studies that used standardized instruments (ES=+0.57).

==========

TABLE 3 HERE

==========

In the current review, far fewer included studies used experimenter-made measures, because BEE standards reject 'treatment-inherent" measures. Yet even after this adjustment, effect sizes for studies using experimenter-made measures were twice the size of effect sizes from assessments not made by the researchers. As indicated in Table 3, the effect sizes for experimenter-made measures and standardized tests were +0.40 and +0.20, respectively.

==========

TABLE 4 HERE

==========

Sample Size

As noted earlier, Slavin & Smith (2009) examined the relationship between sample size (numbers of students) and effect sizes, and found a substantial negative relationship. The current review did a similar comparison but with a much larger set of studies. We first divided studies into those with sample sizes less than 250 subjects (n=335 studies) and those with sample sizes greater than 250 subjects (n=310 studies).

As indicated in Table 4, a statistically significant difference was found between large studies and small studies (Q_B =55.28, df=1, and p<0.00). Studies with small sample sizes produced twice the effect sizes of those with large sample sizes (ES=+0.30 vs ES=+0.16, respectively). Table 4 also breaks down the study-sizw effect into seven categories, from fewer

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

than 100 to more than 3000 students. This shows a linear effect at all levels of sample size, with a ratio of 3.5 to 1 comparing the smallest to the largest categories. The results support the findings of other reviews that made similar comparisons (Liao, 1999; Pearson, Ferdig, Blomeyer, & Moran, 2005). For instance, when Pearson et al. carried out a meta-analysis on the use of digital tools and learning environments to enhance literacy acquisition, they found that studies with smaller sample sizes (N<30) were much more likely to achieve higher treatment effects than those with larger sample sizes. Slavin and Smith (2009) found that "studies with sample sizes below the median of about 250 had a mean effect size of +0.27, whereas those with large sample sizes had a mean effect size of +0.13" (p. 503).

There are several possible explanations for these findings. First, small-scale studies are often more tightly controlled than large-scale studies, and therefore are more likely to produce positive results. The positive results of small-scale studies could be due to what Cronbach et al. (1980) called the "super-realization" effect. That is, in small-scale experiments, researchers or program developers are more likely able to maintain high implementation fidelity or provide additional support that could never be replicated on a large scale.

Second, researcher-developed measures are more likely to be used in small-scale studies while standardized tests, which may be less sensitive to treatments, are often used in large-scale studies (Edmonds et al., 2009; Li & Ma, 2010; Scammacca et al., 2007; Wanzek, Wexler, Vaughn, & Ciullo, 2010).

Further, small studies may appear to have high effect sizes because their limited statistical power requires high effect sizes to reach statistical significance. Studies with small samples that do not produce significant differences may be shelved by researchers as "pilots" or declined by journal editors, as noted previously. If a "pilot" just happens to produce a large effect size, it is

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

likely to be submitted and accepted somewhere, and may therefore be easier for reviewers to find.

==========

TABLE 5 HERE

==========

Randomized vs. Quasi-Experiments

In addition to publication bias and sample sizes, research designs may also affect effect sizes in systematic reviews in education. As indicated in Table 5, we categorized 196 studies as randomized and 449 as well-matched quasi-experiments. The average effect size for randomized experimental studies and quasi-experimental studies were +0.16 and +0.23, respectively. Though significant, this difference in effect sizes is, in proportional terms, less than that found for large vs. small samples, researcher-made vs. independent measures, and published vs. unpublished studies, and previous reviews produced mixed results on this comparison (de Boer, Donker, van der Werf, 2014; Heinsman & Shadish, 1996; Kulik & Kulik, 1991; Li & Ma, 2010; Rake, Valentine, McGatha, & Ronau, 2010; Slavin, Lake, & Groff, 2007; Torgerson, 2007). In a meta-analysis to examine the effect of computer technology on mathematics learning in K-12 classrooms, Li and Ma (2010) found that effect sizes in studies using randomized experiments and those in quasi experiments were essential the same. When reviewing four meta-analyses on educational interventions, Heinsman and Shadish (1996) found that results were generally similar in randomized and quasi-experimental studies. On the other hand, Melby-Lervag & Hulme (2003) and Chiu (1998) found that though there was no significant difference in effect sizes between randomized studies and non-randomized studies, effect sizes for matched studies were generally higher.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Our findings indicate that effect sizes were significantly higher in quasi-experiments than in randomized experiments. In addition, almost two-thirds of the qualifying studies (66%) included in these 12 reviews were quasi-experiments, including matched control, and randomized quasi-experiments (where clusters are randomly assigned to experimental or control conditions but there are too few of them to analyze at the cluster level). Out of the 645 qualifying studies, only 196 (34%) were randomized experiments.

Matched quasi-experiments may produce higher effect sizes than randomized experiments because in matched studies, selective factors may work in favor of the treatment groups. For example, if 20 schools using a particular program are compared to 20 that are using other methods, it is likely that the 20 schools using the program may have chosen to do so because they are more oriented toward innovation, feel more confident in their skills, or are otherwise a stronger staff or have stronger leadership. Even if all quantitative factors are matched in the two sets of schools (e.g., pretests, ethnicities, percent free lunch, teacher experience), there is no way to control for the teachers' motivation or capacity to use the program. When a given program is difficult to use, and especially if some schools have dropped the program, the surviving schools are particularly likely to have an advantage. However, it is important to note that despite these potential biases, reviews have not found strong or consistent differences between matched and randomized experiments whose students are well matched at pretest.

==========

TABLE 6 HERE

==========

Table 6 categorizes the 645 studies by both sample size (more or less than 250) and random assignment/matched design. The table shows that small, matched studies (n=229) have

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

substantially higher effect sizes (ES=+0.33) than large, randomized studies (n=90, ES=+0.12). This table provides a sobering perspective to researchers, developers, and policy makers who wonder why many large, randomized experiments currently being funded by i3 and other funders so often fail to find significant and substantial impacts of educational treatments, even though smaller matched studies did find positive effect sizes.

The current finding carries special importance for researchers and policy makers. First, the small number of randomized studies in this set of studies suggests that there is an urgent need for more randomized experiments in the field. Everything being equal, randomized experiments should be preferred because they eliminate selection bias. In their review, Niemiec, Samson, Weinstein, and Walberg (1987) found that "methodologically weaker studies produced different results than strong studies … [and] the results of quasi-experimental studies have larger variances." Unequal variances may produce results that could be potentially unreliable and misleading (Hedges, 1984). Slavin and Smith (2009) also argued, "randomization provides an important safeguard against selection bias. Selection bias may balance out in the long run, over many studies, but in an area in which small numbers of studies determine conclusions about program effects, such balancing cannot be counted on. Random assignment is essential in building confidence that program outcomes are what they appear to be" (pp. 8-9). The current findings do not suggest that non-randomized experiments should be excluded from meta-analyses, but they do suggest that when carrying out matched studies, researchers or program developers should use every possible means to avoid selection bias and ensure that the treatment and control conditions are comparable. Since matched studies are often less expensive and more feasible to carry out in educational settings, matched control studies, if well-designed, can be a pragmatic alternative to randomized experiments, if interpreted with caution.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Discussion

The findings of this review have major implications for evidence-based reform in education, and for educational research more broadly. In every category examined, methodological factors were associated with substantial differences in effect sizes. In the case of published vs. unpublished papers, the difference was a ratio of almost 2 to 1. Smaller studies (n<250) had twice the effect sizes of larger ones (n>250), and differences were even greater for studies with N's less than 100 (mean ES=+0.38) compared to those with N's greater than 3000 (mean ES=+0.11), a ratio of 3.5 to 1. Even after excluding measures that were deemed to be inherent to (or over-aligned with) treatments, the effect size ratio was 2 to 1 between researcher-made and independent measures. The smallest difference among factors we examined was between randomized and quasi-experimental studies. Quasi-experimental studies were associated with significantly higher effect sizes than were randomized experiments (p<.001), but the ratio was 1.44 to 1, relatively moderate in comparison to the other categories but still reason for substantial concern. Putting together two categories, small quasi-experiments were associated with average effect sizes that exceeded those characteristic of large randomized studies by a ratio of 2.75 to 1.

There is a legitimate question in each of these categories about which is the "true" effect size. For example, unpublished studies are often dissertations done, by definition, by students with less experience and fewer resources than experienced researchers, perhaps with major grants, who are more likely to publish their work. In the case of researcher-made measures, there is good reason to believe that such measures are more sensitive to treatment than are independent tests, which are usually standardized. In the case of sample sizes, large studies may show smaller effect sizes because quality of implementation diminishes in large experiments, and large studies are more likely to use (relatively insensitive) standardized tests as outcome measures. So it might

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

be argued, at a minimum, that the "true" effect size may lie somewhere between the extremes reported here.

The problem with this line of reasoning is that in an applied field like education, the ultimate goal of any program evaluation is to estimate what would happen if the program were implemented at large scale under ordinary circumstances. When program outcome data are used for policy purposes, in particular, it may be of little importance what effect sizes were obtained in small experiments with researchers closely involved in ensuring quality implementation. In pragmatic implementations outside of research, it is reasonable to assume that samples will be large, quality of implementation will be variable, and outcome measures will be standardized. So the lower effect size estimates, for large-scale studies with independent measures, are probably a closer approximation to reality.

The importance of the differences found in this review and others is that in comparing the impact of various interventions on student outcomes, it may matter a great deal which methods tended to be used in their evaluations. For example, imagine that Program A and Program B each have mean effect sizes of +0.20. However, the studies evaluating Program A all used independent, standardized tests, while those evaluating Program B all used researcher-made measures. Are their outcomes truly equal? Program A's effect size is right at the average of the BEE studies for independent tests (ES=+0.20), while Program B's effect size is half of the average for studies using research-made measures (ES=+0.40). Or imagine that Program X has an average effect size of +0.20, all from large, randomized experiments, while Program Y has a mean effect size of +0.30, all from small matched experiments. Program Y appears much more effective, but its effect size is below that for all BEE studies using small quasi-experiments,

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

while the effect size for Program X is almost twice that typical of BEE studies using large randomized designs (ES=+0.12).

The situation becomes more complicated when you consider the role of statistical significance. Assuming a given effect size and typical covariates (such as pretests), the main factor that determines statistical significance is sample size. Most educational experiments today, except those involving approaches directed to individuals or small groups (such as tutoring), apply to entire classes or schools, and therefore should be analyzed at the class or school level, usually using hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002). Typically, HLM requires 40-50 classes or schools for adequate statistical power to detect an effect size of 0.20. The importance of this in the present context is that in order to find statistically significant effects in a cluster randomized trial, the sample size must usually be very large, in terms of numbers of students, and therefore (according to our review) the effect size is likely to be small. In fact, large cluster randomized trials, which are the design of choice in federally funded research, very frequently fail to find statistically significant differences for this reason.

Recommendations For Research and Policy

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 the studies are published or not.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

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. For example, if an evaluation of Program X only has 20 schools (10 experimental, 10 control), and achieves an effect size of +0.20, this is unlikely to be statistically significant. However, if two such studies find effect sizes of +0.20, this could be seen as strong evidence of positive effects across the two underpowered experiments. Building up small experiments in this way would allow less well-funded researchers to do high-quality evaluations over time and to learn from them as they go.
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 covers experimental and control objectives.
4. Policy makers and educators should insist on large, high-quality evaluations 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.

Evidence-based reform has great potential to improve the quality of programs students receive and to fuel much interest and investment in development, research, and dissemination of

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

effective approaches. However, evidence-based policies will prevail only if the evidence itself is rigorous and meaningful. The findings of the analyses in this article and those of many previous analyses tell us the consequences of compromising on the quality of the evidence. These findings should be taken into account in crafting evidence-based policies at all levels of government.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

*****References marked with an asterisk indicate reviews included in this study

References

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, *29*, 189-194.

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

Best-Evidence Encyclopedia (2015). www.bestevidence.org.

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.

*Chambers, B, Cheung, A., & Slavin, R. E. (2015). *Literacy and language outcomes of balanced and developmental approaches to early childhood education: A systematic review*. Manuscript submitted for publication.

*Cheung, A., & Slavin, R. E. (2013a). Effects of instructional technology applications on reading outcomes for struggling readers: A best-evidence synthesis. *Reading Research Quarterly*, *48*(3), 277-299.

*Cheung, A., & Slavin, R. E. (2013b). The effectiveness of educational technology applications on mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9*(1), 88-11.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

*Cheung, A., & Slavin, R. E. (2012a). Effective reading program for Spanish-dominant English language learners (ELLs) in the elementary grades: A best-evidence synthesis. *Review of Educational Research, 84*(4), 351-395.

*Cheung, A. & Slavin, R. E. (2012b). How features of educational technology programs affect student reading outcomes: A meta-analysis. *Educational Research Review. 7*(3), 198-215.

*Cheung, A., Slavin, R. E., Lake, C., & Kim, E. (2015). *Effective science programs: A best evidence synthesis*. Manuscript submitted for publication.

Chiu, C. W. T. (1998). *Synthesizing metacognitive interventions: What training characteristics can improve reading performance*? Paper presented at the annual meeting of the American Educational Research Association, San Diego, California.

Cook, D. J., Guyatt, G. H., Ryan, G., Clifton, J., Willian, A., Mcllroy, W., & Oxman, A. D. (1993). Should unpublished data be included in meta-analyses? Current convictions and controversies. *JAMA, 269*(21), 2749-2753.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. O., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation: aims, methods, and institutional arrangements.* San Francisco, CA: Jossey-Bass.

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.

Dickersin, K., & Rennie, D. (2003). Registering clinical trials. *Journal of American Medicine Association, 290*(5), 516-523.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

comprehension outcomes for older struggling readers. *Review of Educational Research,79*(1), 262-300.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives On Psychological Science, 7*(6), 555-561.

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

Hedges, L. V. (1984). *Meta-analysis: Statistics issues*. Paper presented at the American Educational Research Association, New Orleans, USA.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: when do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*(2), 154-169.

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Of Systematic Reviews*, (2), (DOI: 10.1002/14651858.MR000010.pub3)

Hopewell, S., Clarke, M., & Mallett, S. (2005). Grey literature and systematic reviews. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: prevention, assessment and adjustments* (pp. 49-72). West Sussex, England: John Wiley & Sons.

Kulik, C. L. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior, 7*(1-2), 75-94.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Jefferson, T. O., Alderson, P., Davidoff, F., & Wager, E. (2003). Editorial peer-review for improving the quality of reports of biomedical studies. *Cochrane Methodological Review, Issue 4*, 2003

Khalili, A., & Shashaani, L. (1994). The effectiveness of computer applications: A meta-analysis. *Journal of Research on Computing in Education, 27*(1), 48-62.

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review. 22*, 215-243.

Liao, Y. K. C. (1999). Effects of hypermedia on students' achievement: a meta-analysis. *Journal of Educational Multimedia and Hypermedia*. *8*(3), 255-277.

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

McAuley, L., Pham, B., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectivenes reported in meta-analyses? *The Lancet, 357*, 1228-1231.

Melby-Lervag, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*(2), 270-291.

Niemiec, R. P., Samson, G., Weinstein, T., & Walberg, H. J. (1987). The effects of computer based instruction in elementary schools: A quantitative synthesis. *Journal of Research on Computing in Education, 20*(2), 85-103.

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

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Pearson, P. D., Ferdig, R. E., Blomeyer, R. L., & Moran, J. (2005). *The effects of technology on reading performance in the middle-school grades: A meta-analysis with recommendation for policy*. Naperville, IL: Learning Point Associates.

Rake, C. R., Valentine,, J. C., McGatha, M. B., & Ronau, R. N. (2010). Methods of instructional improvement in Algebra: A systematic review and meta-analysis. *Review of Educational Research, 80*(3), 372-400.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models (2nd ed.).* Thousand Oaks, CA: Sage.

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., Torgensen, J. K. (2007). *Interventions for adolescent struggling readers: A meta-analysis with implications for practice*. Portsmouth, NH: RMC, Research Corporation, Center on Instruction.

Sedlmeier, P., Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin, 105*, 309-316.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Slavin, R. E. (2008). What works? Issues in synthesizing education programs. *Educational Researcher*, 37(1), 5-14.

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

*Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly, 43*(3), 290-322.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

*Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: a best evidence synthesis. *Review of Educational Research, 78*(3), 427-515.

*Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*(4), 1391-1466.

*Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review, 6*, 1-26.

*Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high mathematics: a best evidence synthesis. *Review of Educational Research, 79*(2), 839-911.

*Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching, 51*(7), 870-901. doi: 10.1002/tea.21139

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.

Smith, M. L. (1980). Publication bias and meta-analysis. *Evaluation in Education*, *4*, 22-24.

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.

Swanson, H. L, Hoskyn, M., & Lee, C. (1999). *Interventions for students with learning disabilities.* New York: Guildford.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Torgerson, C. J. (2007). The quality of systematic reviews of effectiveness in literacy learning in English: a "tertiary review. *Journal of Research in Reading, 32*(3), 287-315.

Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. *Reading and Writing*, *23*, 889-912.

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

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Table 1
Review Characteristics

Reference	Topic	Number of studies[1]	Publication types P=Published U=Unpublished	Research design R=Randomized M=Matched	Size L=large S=Small
Chambers, Cheung, Slavin (2015)	Early Childhood	N=32	P=9 U=23	R=27 M=5	L=17 S=15
Slavin & Lake (2008)	Elementary Math	N=50	P=26 U=24	R=11 M=39	L=27 S=23
Slavin, Lake, & Groff (2009)	Secondary Math	N=62	P=25 U=37	R=10 M=52	L=34 S=28
Cheung & Slavin (2013b)	Technology In Math	N=74	P=27 U=47	R=22 M=52	L=36 S=38
Slavin, Lake, Hanley, & Thurston (2014)	Elementary Science	N=23	P=14 U=9	R=8 M=15	L=14 S=9
Cheung, Slavin, Lake, & Kim (2015)	Secondary Science	N=21	P=5 U=16	R=11 M=10	L=14 S=7
Cheung & Slavin (2012a)	Reading for English learners	N=32	P=16 U=16	R=10 M=22	L=5 S=27
Slavin, Lake, Chambers, Cheung, & Davis (2009)	Elementary Reading	N=142	P=54 U=88	R=19 M=123	L=84 S=58
Slavin, Cheung, Groff, & Lake (2008)	Secondary Reading	N=23	P=7 U=16	R=3 M=20	L=17 S=6
Slavin, Lake, Davis, & Madden (2011)	Struggling readers	N=82	P=47 U=35	R=36 M=46	L=8 S=74
Cheung & Slavin (2013a)	Elementary struggling readers	N=20	P=11 U=9	R=13 M=7	L=8 S=12
Cheung & Slavin (2012b)	Technology in Reading	N=84	P=21 U=63	R=26 M=58	L=46 S=38
		Total=645	P=262 U=383	R=196 M=449	L=310 S=335

[1] Duplicate studies were taken out

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Table 2
Published Articles vs. Unpublished Reports

	Number of Studies	Point estimate	Standard error	Q-value	df (Q)	P-value
Published	262	0.30	0.01			
Unpublished	383	0.16	0.01			
Total between	645			58.47	1.00	0.00

Table 3
Experimenter-Made vs. Independent Measures

	Number of Studies	Point estimate	Standard error	Q-value	df (Q)	P-value
Independent Measures	611	0.20	0.01			
Researcher-Made Measures	34	0.40	0.02			
Total between	645			24.06	1.00	0.00

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Table 4
Sample Size and Effect Size

	Number of Studies	Point estimate	Standard error	Q-value	df (Q)	P-value
Small	335	0.30	0.02			
Large	310	0.16	0.01			
Total between	645			55.28	1.00	0.00

Detailed Sample Size Analysis

Up to:	Number of Studies	Point estimate
100	154	+0.38
200	42	+0.26
300	91	+0.21
500	72	+0.19
1000	88	+0.17
2000	46	+0.13
3000	52	+0.11

Table 5
Quasi-Experiments vs Randomized Experiments

	Number of Studies	Point estimate	Standard error	Q-value	df (Q)	P-value
Quasi	449	0.23	0.01			
Randomized	196	0.16	0.01			
Total between	645			19.72	1.00	0.00

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Table 6
Research Design and Sample Size

	Number of Studies	Point estimate	Standard error	Q-value	df (Q)	P-value
Small Matched	229	0.33	0.02			
Large Matched	220	0.17	0.01			
Small Randomized	106	0.23	0.02			
Large Randomized	90	0.12	0.01			
Total between	645			68.54	3.00	0.00

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.