Interpreting Effect Sizes of Education Interventions

Matthew A. Kraft[1]

Researchers commonly interpret effect sizes by applying benchmarks proposed by Jacob Cohen over a half century ago. However, effects that are small by Cohen's standards are large relative to the impacts of most field-based interventions. These benchmarks also fail to consider important differences in study features, program costs, and scalability. In this article, I present five broad guidelines for interpreting effect sizes that are applicable across the social sciences. I then propose a more structured schema with new empirical benchmarks for interpreting a specific class of studies: causal research on education interventions with standardized achievement outcomes. Together, these tools provide a practical approach for incorporating study features, costs, and scalability into the process of interpreting the policy importance of effect sizes.

Keywords: education policy; education reform; effect size; evaluation; experimental research; policy analysis; program evaluation

The ability to make empirical analyses accessible and meaningful for broad audiences is a critical skill in academia. Translating empirical analyses correctly is an equally important skill for anyone who consumes and communicates scholarly research. However, interpreting research findings can be a substantial challenge when outcomes are measured in unintuitive units. This is particularly true in fields such as education, where common outcomes like academic achievement are measured using arbitrary scales. Even in fields that typically examine outcomes measured in more intuitive units such as dollars, it remains difficult to compare the relative success of programs evaluated with different metrics. The typical approach for addressing these challenges is to convert unintuitive and disparate measures onto the same scale using a simple statistic: the standardized effect size.

While a common metric helps, it does not resolve the problem that scholars and research consumers face in evaluating the importance of research findings. For example, P. J. Cook et al. (2015) found that integrating intensive individualized tutoring into the school day raised student achievement in math by 0.23 *SD*, whereas Frisvold (2015) found that offering universal free school breakfasts increased achievement in math by 0.09 *SD*. Are these effects substantively meaningful? Is individualized tutoring a better intervention than universal free breakfast? Answering these questions requires appropriate benchmarks and close attention to study design, costs, and scalability.

The default approach to evaluating the magnitude of effect sizes is to apply a set of thresholds proposed by Jacob Cohen over a half century ago (0.2 = small, 0.5 = medium, 0.8 = large; Cohen, 1969).[1] Cohen's conventions continue to be taught and used widely across the social sciences. However, Cohen's standards are based on a handful of small, tightly controlled lab experiments in social psychology from the 1960s performed largely on undergraduates. Recent meta-analyses of well-designed field experiments have found that education interventions often result in no effect or effects characterized as small by Cohen's standards (Cheung & Slavin, 2016; Fryer, 2017; Lortie-Forgues & Inglis, 2019). Cohen (1988) himself advised that his benchmarks were "recommended for use only when no better basis for estimating the [effect size] index is available" (p. 25). We now have ample evidence to form a better basis.

The persistent application of outdated and outsized standards for what constitutes meaningful effect sizes has had a range of negative consequences for scholarship, journalism, policy, and philanthropy. Researchers design studies without sufficient statistical power to detect realistic effect sizes. Journalists mischaracterize the magnitude and importance of research findings for the public. Policymakers dismiss programs with effects that are small by Cohen's standards but are large relative to existing alternatives. Grantmakers eschew investments

[1]Brown University, Providence, RI

Educational Researcher, Vol. 49 No. 4, pp. 241–253
DOI: 10.3102/0013189X20912798

in programs that deliver incremental gains in favor of interventions targeting alluringly large but unrealistic improvements.

In this article, I develop a framework for interpreting effect sizes that attempts to strike a balance between attention to the contextual features of individual studies and practical considerations for interpreting findings quickly and with limited information. The framework consists of two parts: (a) five broad guidelines with simple questions and corresponding interpretations for contextualizing effect sizes and (b) a more structured schema for interpreting effects from a specific class of studies: causal analyses of education interventions with standardized achievement outcomes.

The article contributes to the effect size literature in several ways. First, I update prior reviews (Bloom et al., 2008; Coe, 2002; Lipsey et al., 2012) with insights from a number of new articles (e.g., Baird & Pane, 2019; Cheung & Slavin, 2016; Funder & Ozer, 2019; Lortie-Forgues & Inglis, 2019; Schäfer & Schwarz, 2019; Simpson, 2017; Soland & Thum, 2019). Second, the interpretive guidelines I present synthesize a range of recommendations from the broader literature that have often been considered in isolation.[2] Third, the schema I propose incorporates new, empirically based benchmarks for effect sizes—derived from a sample of almost 750 randomized control trials (RCTs)—and highlights the underrecognized importance of program cost, scalability, and political feasibility for interpreting the policy relevance of research findings.

I begin by providing a brief summary of the evolution of education research, which serves to illuminate the origins of many common misinterpretations of effect sizes. I then briefly review common approaches to translating effects into more intuitive units such as months of learning or percentile changes. Next, I introduce a set of guidelines and a corresponding schema for interpreting effect sizes, apply them to several examples, and conclude by discussing the implications of the proposed framework.

Effect Sizes and the Evolution of Education Research

Until the mid-20th century, researchers often evaluated the importance of quantitative findings based on significance tests and their associated *p* values. Such statistics, however, are a function of sample size and say nothing about the magnitude or practical relevance of a result. As the social sciences slowly moved away from a myopic focus on statistical significance, scholars began reporting on the practical significance of their findings using the standardized effect size statistic (hereafter, "effect size") or Cohen's *d*:

$$Effect\ Size = \frac{\left[Mean_1 - Mean_2\right]}{Standard\ Deviation}. \quad (1)$$

Most basically, effect sizes are a measure of differences in means between two subgroups divided by the standard deviation of the measure of interest (Lipsey et al., 2012). In the context of program evaluations, $Mean_1$ is the mean of the treatment group, and $Mean_2$ captures the mean of the control or comparison group. There are several approaches to estimating the standard deviation, which I discuss in more detail below.

In 1962, Jacob Cohen proposed a set of conventions for interpreting the magnitude of effect sizes, which he later refined in 1969. As Cohen (1969) emphasized in his seminal work on power analysis, researchers needed a framework for judging the magnitude of a relationship to design studies with sufficient statistical power. His conventions provided the foundation for such a framework when little systematic information existed.

Early meta-analyses of education studies appeared to affirm the appropriateness of Cohen's benchmarks for interpreting effect sizes in education research. A review of over 300 meta-analyses by Lipsey and Wilson (1993) found a mean effect size of precisely 0.50 *SD*. However, many of the research studies included in these meta-analyses used small samples, weak research designs, and proximal outcomes highly aligned to the interventions—all of which result in systematically larger effects (Cheung & Slavin, 2016). Influential reviews by Hattie (2009) continued to incorporate these dated studies and ignored the importance of research design and other study features, further propagating outsized expectations for effect sizes in education research.

The "2 sigma" studies conducted by Benjamin Bloom's doctoral students at the University of Chicago provide a well-known example of education research from this period. Bloom's students conducted several small-scale experiments in which fourth, fifth, and eighth graders received instruction in probability or cartography for 3 to 4 weeks. Students randomized to either (a) mastery-based learning classes with frequent formative assessments and individual feedback or (b) one-on-one/small group tutoring also with assessments and feedback outperformed students in traditional lecture classes by 1.0 and 2.0 *SD*, respectively (Bloom, 1984). The Bloom 2 sigma studies and others like them helped to anchor education researchers' expectations for unrealistically large effect sizes, despite early objections (Slavin, 1987).

At the turn of the 21st century, a growing emphasis on causal inference across the social sciences began to reshape quantitative research in education (Angrist, 2004; T. D. Cook, 2001; Gueron & Rolston, 2013; Murnane & Nelson, 2007). Starting in 2002, the newly established Institute of Education Sciences (IES) began providing substantial federal funding for large-scale randomized field trials, and the U.S. Department of Education increasingly required rigorous evaluations of grant-funded programs. Effect sizes from this new generation of field experiments have been strikingly smaller as new norms about preregistering research designs, hypotheses, and outcomes have emerged. For example, Lortie-Forgues and Inglis (2019) found an average effect size of only 0.06 *SD* among 141 RCTs funded by IES and the UK-based Education Endowment Foundation.

IES also launched The What Works Clearinghouse (WWC), an initiative to review and summarize "gold-standard" evidence on education programs. The WWC (n.d.) has established detailed standards for evaluating causal research designs and produces evidence summaries for specific education interventions by outcome domain. Each summary includes a six-category "effectiveness" rating based on the quality of the research design, the precision and magnitude of findings, and the consistency of

results. This approach shares several similarities with the schema and benchmarks I propose below.

Current Approaches to Translating Effect Sizes

Although Cohen's benchmarks continue to color our interpretation of effect sizes, scholars have increasingly adopted translational approaches to interpreting research findings. These approaches, which convert effect sizes onto more broadly familiar scales, provide helpful intuition when applied with care. Several of these translational approaches are worth highlighting (for detailed descriptions of these techniques, see Baird & Pane, 2019; Hill et al., 2008; Lipsey et al., 2012).

Months of Learning

Scholars often convert effect sizes into months of learning by comparing their estimates to empirical benchmarks for annual learning gains (Bloom et al., 2008). For example, the 0.09 *SD* effect of universal free breakfast on math achievement is approximately 1.6 months of learning.[3] This approach has a strong intuitive appeal but can also be misleading. Year-to-year achievement gains capture learning that occurs both inside and outside of school as well as learning that is the result of naturally occurring cognitive development. An effect might seem trivial compared to the rapid learning rates in early childhood but in fact be quite impressive relative to the effects of early childhood education. Large differences in learning gains across grades make it important to use the appropriate grade-specific benchmark and difficult to translate effects that pool across several grades (Baird & Pane, 2019).[4]

Changes in Percentile Rank

The change in percentile rank approach describes an effect as moving the average student in the sample from some initial percentile to the percentile that corresponds with the effect size of interest. Because the total percentile point change is sensitive to the starting percentile one chooses, it is important to describe both the initial and postintervention percentiles. For example, the effect of individualized tutoring (0.23 *SD*) is equivalent to moving male students in distressed Chicago high schools from the 50th to the 59th percentile of achievement or, alternatively, from the 10th to the 15th percentile.

Achievement Gaps

Benchmarking against achievement gaps helps to frame effects using a widely studied and policy-relevant metric. The effect of universal free breakfast on math achievement (0.09 *SD*) represents 11% of the student-level Black-White achievement gap.[5] This framing is helpful but can also mislead people to believe that an intervention would decrease the Black-White achievement gap by the same magnitude. Whether an intervention raises achievement more for certain groups than others depends on how it is targeted and its relative effects across different student subgroups.

Differences in Teacher or School Effectiveness

Mapping effect sizes onto changes in the distribution of teacher or school effectiveness helps to benchmark effects relative to those that are being achieved within the education system. For example, a 0.09 *SD* effect is equivalent to the difference between the median teacher and a teacher at approximately the 73rd percentile in the distribution of teacher effectiveness or between the median school and a school at roughly the 79th percentile of school effectiveness.[6] This approach, however, is sensitive to the estimate one uses for the magnitude of teacher and school effects.

No single translational approach is uniformly better; their value depends on the context of the study and the audience one is trying to reach. But translations alone are not enough to interpret the policy relevance of an effect size. These unit conversions provide additional intuition but still leave the interpretation to the reader and allow considerable room for disagreement. They are complements, not substitutes, for more direct interpretations that consider study features, program costs, and scalability. It is time researchers update and expand their approach.

Five Guidelines for Interpreting Effect Sizes

1. Results From Correlational Studies Presented as Effect Sizes Are Not Causal Effects

The term *effect size* can be misleading. A logical way to interpret it is as "the size of an effect" or how large the causal effect of X is on Y. This interpretation is accurate when it applies to effect sizes that represent the standardized mean difference between treatment and control groups in RCTs. Random assignment eliminates systematic differences between groups, so any subsequent differences are attributable to the intervention.[7] However, effect sizes often represent simple descriptive relationships between two variables, such as height and achievement. Although the practice of referring to correlation coefficients as effect sizes is largely limited to psychology, education researchers frequently use the term *effect size* to report changes in performance over time and estimates from regression models using observational data. These descriptive effect sizes provide useful information but likely do not reflect underlying causal relationships. Taller students have higher achievement because they are older, on average, not because of their stature.

Knowing whether an effect size represents a causal or correlational relationship matters for interpreting its magnitude. Comparing meta-analytic reviews that incorporate effect size estimates from observational studies (e.g., Hattie, 2009; Lipsey & Wilson, 1993) to those that only include experimental studies (e.g., Hill et al., 2008; Lipsey et al., 2012; Lortie-Forgues & Inglis, 2019) illustrates how correlational relationships are, on average, substantially larger than causal effects. It is incumbent on researchers reporting effect sizes to clarify which type their statistic describes, and it is important that research consumers do not assume effect sizes inherently represent causal relationships.

> ASK: Does the study estimate causal effects by comparing approximately equivalent treatment and control groups, such as an RCT or quasi-experimental study?

INTERPRET: Effect sizes from studies based on correlations or conditional associations do not represent credible causal estimates.

INTERPRET: Expect effect sizes to be larger for correlational studies than causal studies.

2. The Magnitude of Effect Sizes Depends on What, When, and How Outcomes Are Measured

What outcomes are measured. Studies are more likely to find larger effects on outcomes that are easier to change, proximal to the intervention, administered soon after the intervention is completed, and measured with more precision (Ruiz-Primo et al., 2002). Outcomes that reflect short-term decision making and effort, such as passing a class, are easier to influence than outcomes that are the culmination of years of decisions and effort, such as graduating from high school. Similarly, outcomes that are more directly related to the intervention will also be easier to move. For example, teacher coaching has much larger effects on teachers' instructional practice (0.47 *SD*) than on students' achievement (0.18 *SD*; Kraft, Blazar, & Hogan, 2018), and social-emotional learning (SEL) programs have much larger effects on students' SEL skills (0.57 *SD*) compared to their academic performance (0.27 *SD*; Durlak et al., 2011).

Even among measures of student achievement, effect sizes for researcher-designed and specialized topic tests aligned with the treatment are often 2 to 4 times larger than effects on broad standardized state tests (Cheung & Slavin, 2016; Hill et al., 2008; Lipsey et al., 2012; Lynch et al., 2019). These larger effects on researcher-designed, specialized assessments can be misleading when they reflect narrow, nontransferable knowledge. The Bloom (1984) 2 sigma effects on probability and cartography tests after a month of tutoring are 8 to 20 times larger than the effects on standardized math tests found in several recent studies of even more intensive daily tutoring over an entire school year (P. J. Cook et al., 2015; Fryer & Noveck, 2020; Kraft, 2015).

ASK: Is the outcome the result of short-term decisions and effort or a cumulative set of decisions and sustained effort over time?

INTERPRET: Expect outcomes affected by short-term decisions and effort to be larger than outcomes that are the result of cumulative decisions and sustained effort over time.

ASK: How closely aligned is the intervention with the outcome?

INTERPRET: Expect outcomes more closely aligned with the intervention to have larger effect sizes.

When outcomes are measured. When an outcome is measured also influences the magnitude of effect sizes. Outcomes assessed immediately after an intervention ends are likely to show larger effects than outcomes captured months or years later (Bailey et al., 2017). For example, studies of the effect of attending high-performing charter high schools in Boston using lottery admissions show large effects on contemporaneous achievement outcomes, more moderate effects on college-going outcomes, and limited effects on college completion (Angrist et al., 2016; Setren, 2019). A helpful mental framework for assessing the proximity of an outcome to treatment is to think about the causal chain of events that must occur for an intervention to affect an outcome. The further down this causal chain, the smaller the effect sizes are likely to be.

ASK: How long after the intervention was the outcome assessed?

INTERPRET: Expect outcomes measured immediately after the intervention to have larger effect sizes than outcomes measured later.

How reliably outcomes are measured. Even when comparing similar outcomes measured at the same time, differences in measure reliability can affect the magnitude of effect sizes. This is because the instruments researchers use to measure outcomes are imperfect. The lower the reliability, the greater the error variance, and thus, the greater the measured variance. Dividing by a larger measure of variance in Equation 1 results in a smaller effect size. As Boyd et al. (2008) showed, measurement error can differ substantially across outcomes. They found that measurement error accounted for 17% of the variance in standardized test scores but 84% of the variance in test score gains (changes in students' scores across time).

ASK: How reliably is the outcome measured?

INTERPRET: Expect measures with lower reliability to have smaller effect sizes than comparable measures with higher reliability.

3. Subjective Decisions About Research Design and Analyses Influence Effect Sizes

The study sample. One of the most common findings in social science research is treatment effect heterogeneity—variation in treatment effects across subgroups. For example, growth mindset interventions are consistently more effective among lower-achieving students (Paunesku et al., 2015, Yeager et al., 2019). This heterogeneity makes it important to consider sample characteristics when evaluating the magnitude of an effect size. A variety of factors can influence the composition of the study sample. The intervention design itself may dictate which subjects can be included in the sample. Universal interventions, such as providing universal free breakfasts, allow for population-level samples. More targeted interventions, such as requiring low-achieving students to repeat a grade, can only be studied among more restricted samples (Greenberg & Abenavoli, 2017).

The recruitment process can also affect the composition of the study sample and thus, the resulting effect sizes. Researchers often recruit a limited set of study participants given cost and capacity constraints. Students, teachers, schools, and districts are more likely to participate in a study when they think they will benefit, causing selection bias (Allcott, 2015). When first testing the potential efficacy of an intervention, researchers themselves

often recruit participants that they expect to benefit the most. Targeted interventions and small-scale efficacy trials generally produce larger effect sizes than universal interventions because they target study participants that are most likely to benefit and because there is less variation in outcomes among smaller, nonrepresentative samples (Cheung & Slavin, 2016).

ASK: Are study participants a broad sample or a subgroup most likely to benefit from the intervention?

INTERPRET: Expect studies with more targeted samples to have larger effect sizes than studies with more diverse and representative samples.

The standard deviation. Researchers exercise considerable judgment about what standard deviation they use to calculate an effect size. This involves making two subjective decisions, one about the correct measure to use and another about the appropriate sample for estimating the variance. For example, researchers choose among several different measures to standardize effects on achievement, including variation in student-level test scores, average school-level test scores, or changes in student test scores over time (i.e., gains). Whenever possible, researchers should present effects standardized at the student level, irrespective of the level of treatment or the unit of analysis. This approach directly answers the question policymakers are most often interested in—How much does the intervention benefit kids?—and provides a common point of comparison with the vast majority of effect sizes in education research.

It makes sense to also present effect sizes relative to variation in test-score gains or school-level average achievement when research questions focus explicitly on these quantities. However, scholars and consumers of research should expect these approaches to produce effect sizes that are approximately 1.5 to 3 times larger than effect sizes scaled relative to student-level scores (Boyd et al., 2008; Dee & Dizon-Ross, 2019; Hedges, 2007). This is because there is substantially less variation in both school-level averages and gains compared to student scores.

ASK: Is the effect size standardized relative to the variation in an individual-level measure, an aggregate-level measure, or a change across repeated measures?

INTERPRET: Expect effect sizes that are standardized using variation in aggregate-level measures or changes across repeated measures to be substantially larger than those using individual-level measures.

After selecting the level of standardization, researchers decide what sample to use to calculate the variance. Scholars typically choose between three types: (a) the complete analytic (i.e., pooled) sample, (b) the control group sample, and (c) an estimate from a larger population.[8] For example, the effect of individualized tutoring in P. J. Cook et al. (2015) of 0.23 *SD* uses the control group sample. They also reported effects scaled by the national distribution of test scores, which reduced the estimated effect to 0.19 *SD*. This is because the more homogenous group of students who were offered tutoring had less variable test performance (i.e., smaller standard deviation) than students in an unrestricted national sample. When baseline outcome measures are not available, it is preferable to use the standard deviation of the control group outcome rather than the pooled sample because the intervention may have affected the variation in outcomes among the treatment group.

ASK: What sample produced the standard deviation used to estimate effect sizes?

INTERPRET: Expect effect sizes that are standardized using more homogeneous and less representative samples to have larger effect sizes.

The treatment-control contrast. For RCTs, the contrast between the experiences of the treatment and control groups plays an important role in determining effect sizes. For example, some early evaluations of center-based early childhood education programs, such as the HighScope Perry Preschool Project, compare treatment students to control group students who were almost exclusively cared for by guardians at home (Heckman et al., 2010). In more recent studies, such as the Head Start Impact Study, the difference in child care experiences between the treatment and control groups is far less pronounced because most children in the control group also received center-based care (Puma et al., 2010). This weaker treatment-control contrast is one reason why studies find larger effect sizes for Perry Preschool than for Head Start (Kline & Walters, 2016).

Some education interventions are constrained to have smaller contrasts than others, resulting in potentially systematic differences in effect sizes (Simpson, 2017). Interventions that offer supplemental resources or services such as one-on-one tutoring can be evaluated against a control group that does not receive these supports, providing a large contrast. However, standard educational practices such as student behavior management programs cannot be evaluated relative to a control group where student behavior goes unaddressed. The treatment-control contrast in this case is between a new approach and the current behavioral approach. Interpreting effect sizes from RCTs requires a clear understanding about the nature of the control condition.

ASK: How similar or different was the experience of the treatment group compared to the control or comparison group?

INTERPRET: Expect studies to have smaller effect sizes when control groups have access to resources or services similar to the treatment group.

The type of treatment effect estimated. Researchers who conduct RCTs are often able to answer two important but different questions: What is the effect of *offering* the intervention, and what is the effect of *receiving* the intervention. Assuming not everyone randomized to the treatment group participates in the intervention, we would expect the effect of the offer of the intervention (i.e., intent to treat) to be smaller than the effect of actually receiving it (i.e., treatment on the treated). Returning to the intensive tutoring study (P. J. Cook et al., 2015), the 0.23 *SD* effect on math achievement represents the effect of receiving

tutoring. However, only 41% of all students who were randomly assigned to be offered tutoring took up this offer.[9] Thus, the effect of offering tutoring, which includes all students who received the offer regardless if they took up it, was a smaller 0.13 *SD*. Understanding the degree to which implementation challenges cause eligible individuals not to participate in a program is critical for informing policy and practice.

ASK: Does the effect size represent the effect of offering the intervention or the effect of receiving the intervention?

INTERPRET: Expect studies that report the effect of offering the intervention to have smaller effect sizes than studies that report the effect of receiving the intervention.

4. Costs Matter for Evaluating the Policy Relevance of Effect Sizes

As several authors have argued persuasively, effect sizes should be considered relative to their costs when assessing the importance of an effect (Duncan & Magnuson, 2007; Harris, 2009; Levin & Belfield, 2015). Two things are particularly salient for policymakers examining education programs: the potential returns per dollar invested and the total costs. Spending the marginal dollar on the most cost-effective program makes sense. At the same time, the financial implications of reforms that require large initial investments, such as modernizing school facilities, are very different from programs that are lower cost and flexible with scale, such as free school breakfasts. Policymakers have to consider not only what works but also how well it works relative to costs and the total financial investment required.

Studies increasingly include back-of-the envelope estimates of per-participant costs, which serve to contextualize the return of an education intervention. More comprehensive cost-effectiveness analyses that account for both monetary and nonmonetary costs, such as the opportunity costs of educators' time, would go even further to provide policymakers with valuable information for making difficult decisions with limited resources. At the same time, increased attention to cost-effectiveness should not lead us to uniformly dismiss costlier programs or policies. Many challenges in education such as closing long-standing achievement gaps will likely require a combination of cost-effective and costlier approaches.

ASK: How costly or cost-effective is the intervention?

INTERPRET: Effect sizes from lower-cost interventions are more impressive than similar effects from costlier programs.

5. Scalability Matters for Evaluating the Policy Relevance of Effect Sizes

Similar to program costs, assessing the potential scalability of program effects is central to judging their importance for policy and practice. One of the most consistent findings in the education literature is that effects decrease when smaller targeted programs are taken to scale (Slavin & Smith, 2009). Two related but distinct challenges are behind this stylized fact: (a) Program effects are often heterogeneous, and (b) programs are often difficult to replicate with fidelity at scale. As discussed previously, impressive effects from nonrepresentative samples are unlikely to scale when programs are expanded to more representative populations. Thus, the greater the external validity of a study, the greater its policy importance is.

Even for program effects with broad external validity, it is often difficult to replicate effects at scale due to implementation challenges. In the highly decentralized U.S. education system, the success of most education interventions depends on the will and capacity of local educators to implement them (Honig, 2006). For example, of the 67 education interventions the U.S. Department of Education Investing in Innovation Fund (i3) selected to fund because of prior evidence of success, only 12 produced significant positive effects when taken to scale (Boulay et al., 2018). Efforts to reduce class sizes statewide in California, which were inspired by the large academic gains found in the Tennessee STAR class size experiment, failed to produce similar effects (Jepsen & Rivkin, 2009).

The challenge posed by taking programs to scale is largely proportional to the degree of behavioral change required to implement a program. Top-down interventions that require limited implementation by personnel are often easier to scale. Examples include financial incentives for recruiting teachers, changing school starting times, and installing air conditioning in schools. Interventions that require more coordinated and purposeful implementation among school personnel often face greater challenges. Examples are implementing a new behavioral support system, engaging in professional learning communities, and teaching new curricula.

Political feasibility and unintended consequences also play an important role in determining scalability. Interventions often stall when they face opposition from organized constituencies. Nationwide reforms to teacher evaluation systems did little to remove ineffective teachers or reward highly effective ones given the strong opposition these efforts faced in most districts (Kraft, 2018). As programs scale, their direct effects become even more confounded with any corresponding indirect effects due to how the intervention might cause students, educators, or parents to change their behavior in unexpected ways (Todd & Wolpin, 2003).

To be clear, more technical, top-down interventions are not uniformly better than those that require widespread behavioral change or create political headwinds. At its core, school improvement is about strengthening leadership and instructional practices, both of which require behavioral change that can push educators outside of their comfort zones. What matters is better understanding the behavioral, financial, and political challenges required to expand programs while maintaining their effectiveness.

ASK: How likely is it that the intervention could be replicated at scale under ordinary circumstances?

INTERPRET: Programs are unlikely to maintain their effectiveness at scale if they are only effective with a narrow population, entail substantial behavioral changes, require a skill level greater than that possessed by typical educators, face considerable opposition among the public or practitioners, or depend on the charisma of a single person or a small corps of highly trained and dedicated individuals.

Toward a New Schema for Interpreting Effect Sizes

There exists an inherent tension in providing guidance on interpreting effect sizes. Broad guidelines can be applied widely and flexibly but require a degree of technical expertise and result in subjective interpretations. Fixed benchmarks are easy to use and provide unambiguous answers but cannot fully account for differences in study features or the degree of statistical uncertainty inherent in any estimate. Some scholars argue "there is no wisdom whatsoever" in proposing benchmarks (Glass, McGaw, & Smith, 1981, p. 104) and that "it would be inappropriate to wed effect size to some necessarily arbitrary suggestion of substantive significance" (Kelley & Preacher, 2012, p. 146). At the same time, benchmarks may be a pragmatic necessity given that human cognition relies on comparisons and heuristic shortcuts to make sense of complex information. The persistent application of Cohen's benchmarks despite repeated calls to abandon them suggests that little short of a simple alternative will dislodge them. Nature abhors a vacuum.

One way to ease this tension is for researchers to identify benchmarks for specific classes of studies and outcome types based on the distribution of effects from the relevant literature (e.g., Tanner-Smith et al., 2018). Benchmarking based on existing interventions applies a practical counterfactual to answer a specific question: "How large is the effect relative to other studies with broadly comparable features?" Benchmarks based on comparable studies would provide a more informed starting place for interpretations that we can then adjust based on the characteristics of the relevant study.

The schema I propose provides new, baseline benchmarks for one class of studies: causal research that evaluates the effect of education interventions on standardized student achievement. The motivation for this focus is threefold. First, it serves to narrow the contextual differences that make benchmarks impractical when considering a more diverse body of research. Second, standardized achievement tests are taken annually by tens of millions of public school students and are strong predictors of a range of positive outcomes in adulthood (Goldhaber & Özek, 2019). Third, we now have a large literature of causal research evaluating programs using standardized achievement outcomes on which to base new benchmarks.

New Empirical Benchmarks

I propose the following baseline benchmarks for effect sizes from causal studies of preK–12 education interventions evaluating effects on student achievement: less than 0.05 is *small*, 0.05 to less than 0.20 is *medium*, and 0.20 or greater is *large*. These proposed benchmarks are based on the distribution of 1,942 effect sizes from 747 RCTs evaluating education interventions with standardized test outcomes (for source data and coding details, see Appendix A available on the journal website). As shown in Table 1, these values divide the overall distribution, with a median of 0.10 *SD*, into approximate thirds (37th and 69th percentiles).

If calling an effect size of 0.20 *SD* large seems overly enthusiastic, consider this: By fifth grade, student achievement improves about 0.40 *SD* or less over the course of an academic year (Bloom et al., 2008), and schools only account for around 40% of these achievement gains (Chingos et al. 2015; Konstantopolus & Hedges, 2008; Luyten et al., 2017). Formal schooling, our society's defining education intervention, is delivered over more than 1,000 hours a year, costs over $10,000 per student, and barely qualifies as producing large effects in middle and high school. Alternatively, raising student achievement by 0.20 *SD* results in a 2% increase in annual lifetime earnings on average (Chetty et al., 2014).

Others might object to characterizing a 0.05 *SD* as a medium-sized effect, but raising academic achievement is difficult. One in four effect sizes from RCTs of education interventions with standardized test outcomes described in Table 1 are zero or negative, with many more small, positive effects that cannot be distinguished from zero. Even this likely understates the failure rate among interventions given publication bias against null findings.

These baseline benchmarks provide a simple, general heuristic but in doing so, average across heterogeneity in effects related to study characteristics such as sample size, subject, grade level, and test type. Thus, they provide an informed starting point that should be adapted based on specific study characteristics, not the definitive interpretation of an effect size from causal studies of education interventions with preK–12 achievement outcomes.

Adapting the Benchmarks

In Table 1, I explore how we might adapt these baseline benchmarks to account for effect size heterogeneity. Overall, effect sizes in reading are slightly larger than those found in math. However, disaggregating by grade level reveals that the larger average effects in reading are driven exclusively by the considerably large effects on standardized tests of early-literacy skills in prekindergarten through third grade. This heterogeneity is evident in Figure 1, which depicts the median, interquartile range, and 10th to 90th percentiles of effect sizes in math (Panel A) and reading (Panel B) across grade levels (for specific statistics, see Appendix Table B1 available on the journal website).

In math, the distribution of effect sizes is relatively stable across grade levels despite students making much larger learning gains in early childhood than during adolescence (Bloom et al., 2008; Lee, Fin, & Liu, 2019). Median effects in math cluster tightly between 0.04 and 0.09 *SD* across all grades above prekindergarten and are similar in magnitude to effect sizes in reading across 4th to 12th grades (median between 0.04 and 0.08 *SD*). These results suggest that the proposed benchmarks are broadly applicable, if not even slightly high thresholds, for most grade and subject combinations with the exception of prekindergarten and lower elementary grades in reading. One might adjust benchmarks for evaluating effect sizes based on assessments of early literacy upward to 0.10 and 0.30 *SD*.

Similar to prior studies, I find further evidence that larger studies with broad achievement measures have systematically smaller effect sizes. Effect sizes from studies with samples greater than 2,000 students are several times smaller than studies with 100 students or fewer (medians of 0.03 vs. 0.24 *SD*). Some of this difference is likely driven by publication bias and budget constraints that make resource-intensive interventions more

Table 1
Empirical Distributions of Effect Sizes From Randomized Control Trials of Education Interventions With Standardized Achievement Outcomes

		Subject		Sample Size					Scope of Test		DoE
	Overall	Math	Reading	≤100	101–250	251–500	501–2,000	>2,000	Broad	Narrow	Studies
Mean	0.16	0.11	0.17	0.30	0.16	0.16	0.10	0.05	0.14	0.25	0.03
Standard deviation	0.28	0.22	0.29	0.41	0.29	0.22	0.15	0.11	0.24	0.44	0.16
Mean (weighted)	0.04	0.03	0.05	0.29	0.15	0.16	0.10	0.02	0.04	0.08	0.02
P1	−0.38	−0.34	−0.38	−0.56	−0.42	−0.29	−0.23	−0.22	−0.38	−0.78	−0.38
P10	−0.08	−0.08	−0.08	−0.10	−0.14	−0.07	−0.05	−0.06	−0.08	−0.12	−0.14
P20	−0.01	−0.03	−0.01	0.02	−0.04	0.00	−0.01	−0.03	−0.03	0.00	−0.07
P30	0.02	0.01	0.03	0.10	0.02	0.06	0.03	0.00	0.02	0.05	−0.04
P40	0.06	0.04	0.08	0.16	0.07	0.10	0.06	0.01	0.06	0.11	−0.01
P50	0.10	0.07	0.12	0.24	0.12	0.15	0.09	0.03	0.10	0.17	0.03
P60	0.15	0.11	0.17	0.32	0.17	0.18	0.12	0.05	0.14	0.22	0.05
P70	0.21	0.16	0.23	0.43	0.25	0.22	0.15	0.08	0.20	0.34	0.09
P80	0.30	0.22	0.33	0.55	0.35	0.29	0.19	0.11	0.29	0.47	0.14
P90	0.47	0.37	0.50	0.77	0.49	0.40	0.27	0.17	0.43	0.70	0.23
P99	1.08	0.91	1.14	1.58	0.93	0.91	0.61	0.48	0.93	2.12	0.50
k (number of effect sizes)	1,942	588	1,260	408	452	328	395	327	1,352	243	139
n (number of studies)	747	314	495	202	169	173	181	124	527	91	49

Note. A majority of the standardized achievement outcomes (95%) are based on math and English language art test scores, with the remaining based on science, social studies, or general achievement. Weights are based on sample size for weighted mean estimates. For details about data sources, see Appendix A, available on the journal website. DoE = U.S. Department of Education.

likely to be evaluated in smaller samples. Studies that use broad achievement measures produce effect sizes that are noticeably smaller than those on narrow measures (medians of 0.10 vs. 0.17 *SD*). RCTs funded by the U.S. Department of Education, which requires scholars to preregister their research design and report their findings, have a median effect size of 0.03 *SD* across 139 effect sizes from 49 RCTs. These patterns suggest that effects of 0.15 or even 0.10 *SD* should be considered large and impressive when they arise from large-scale field experiments that are preregistered and examine broad achievement measures.

Incorporating Costs and Scalability

Simply reclassifying the magnitude of effect sizes is not sufficient from a policy perspective because effect sizes do not reflect the cost of a program or how likely it is to scale with fidelity. The schema shown in Table 2 combines the baseline effect size benchmarks with a corresponding set of empirically based per-pupil cost benchmarks in 2016 constant dollars: less than $500 is *low*, $500 to under $4,000 is *moderate*, and $4,000 or greater is *high* (for more details see, Appendix Table C1 available on the journal website).[10] Given that these cost benchmarks are derived from a sample of only 68 education interventions, they should be viewed as a rough guide for classifying effect sizes into the simple cost-effectiveness ratios shown in this 3 × 3 matrix.

The matrix in Table 2 helps to clarify two key insights about interpreting effect sizes: Large effects are not uniformly more important than smaller effects, and low-cost interventions are not uniformly more favorable than costlier interventions. One can see this in the different combinations of effect sizes and costs that have similar cost-effectiveness ratios on a given downward-sloping diagonal, with green shading representing higher and red shading representing lower cost-effectiveness ratios. At the same time, interventions with similar cost-effectiveness ratios are not interchangeable because policy decisions depend on local priorities, resources, and politics as well.

The last interpretive step is assessing whether an intervention is easy, reasonable, or hard to scale. Because there are no clear benchmarks to apply here, this step requires the subjective judgement of the interpreter. Ask, would the effects be similar if the intervention were offered to a large, diverse population of students? Is it likely the intervention would be implemented with fidelity by others? Is it politically feasible to scale the intervention? Reasonable people will disagree about the answers to these questions. The larger point is to introduce scalability into the process of interpreting effect sizes and to consider whether an intervention falls closer to the easy-to-scale or hard-to-scale end of the spectrum. Assessing scalability helps to provide a measure of the challenges associated with expanding a program so that these challenges are considered and addressed.

Some Examples

Applying the proposed framework to several examples helps to illustrate the importance of interpreting effect sizes across multiple dimensions. Consider, for example, the previously cited

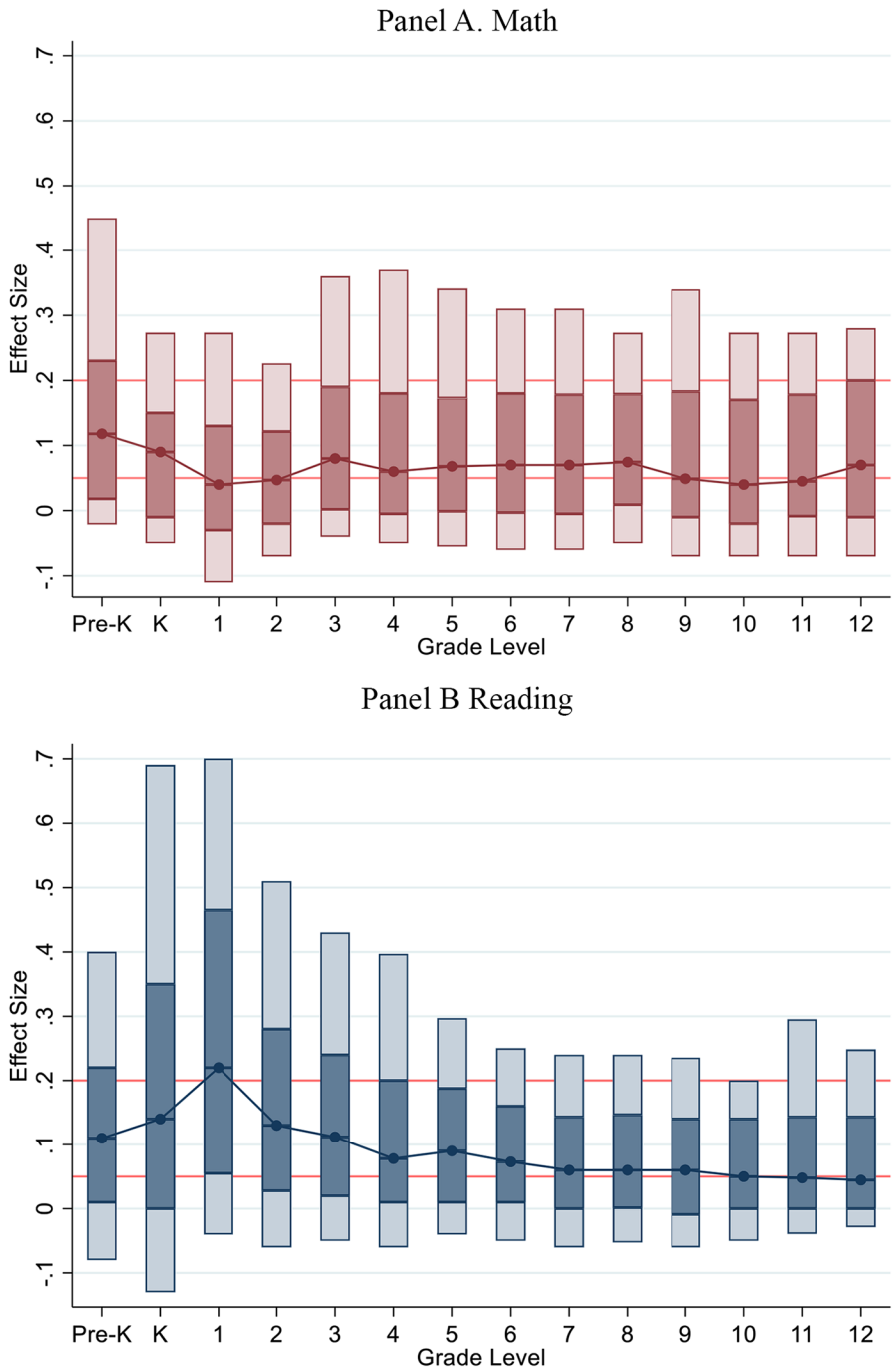

FIGURE 1. *The distribution of effect sizes from randomized control trials of education interventions with student achievement outcomes by subject and grade level.*

Note. Vertical bars represent 10th to 90th percentile ranges with darker shaded interquartile ranges (25th–75th percentiles). Connected line dots illustrate changes in median effect sizes across grade distributions. Red horizontal lines indicate proposed effect size benchmarks.

studies evaluating the effects of universal free breakfast (0.09 *SD*; Frisvold, 2015) and individualized tutoring (0.23 *SD*; P. J. Cook et al., 2015). In many ways, these studies share similar core features. Both studies employed causal methods and examined effects on broad, reliable state achievement tests in math, standardized at the student level, and assessed at the end of the school year in which the interventions were implemented. Both studies analyzed sizable samples of over 2,000 students in grades with few systematic differences in the average effect size of education interventions (4th/5th vs. 9th/10th).

However, differences in sample characteristics, analytical approaches, costs, and scalability all indicate these effect sizes might be more similar in practical importance than their magnitudes suggest. P. J. Cook et al. (2015) targeted their tutoring study to male youth of middling achievement in distressed Chicago high schools, a narrow population for which the

Table 2
A Schema for Interpreting Effect Sizes From Causal Studies of Education Interventions With Standardized Achievement Outcomes

	Cost-Effectiveness Ratio				
	Cost Per Pupil				
Effect Size	Low (<$500)	Moderate ($500 to <$4,000)	High (≥$4,000)		Scalability
Small (<.05)	Small ES / low cost	Small ES / moderate cost	Small ES / high cost		Easy to scale
Medium (.05 to <.20)	Medium ES / low cost	Medium ES / moderate cost	Medium ES / high cost	&	Reasonable to scale
Large (≥.20)	Large ES / low cost	Large ES / moderate cost	Large ES / high cost		Hard to scale

Note. Green and red shading represent higher and lower cost-effectiveness ratios, respectively. Effect size and cost benchmarks provide empirically informed starting places that should be adapted based on the characteristics of individual studies. ES = effect size.

intervention is specifically designed and in which there is less variance in outcomes. They also focused on the effect of *receiving* tutoring, whereas Frisvold (2015) reported on the effect of *offering* a universal intervention—free breakfast—to all elementary school students. Both of these differences in study features likely contribute to the larger effect size for tutoring.

Considering costs further illustrates how the smaller effect of universal free breakfast is, from a policy standpoint, equally if not more impressive than the large effect of individualized tutorials. Studies suggest a conservative estimate for the annual cost of universal free breakfast is $50 to $200 per student, depending on state and federal reimbursement rates (Schwartz & Rothbart, in press). P. J. Cook et al. (2015) reported that the annual cost of individualized tutoring is more than $2,500 per student. Universal free breakfast produces a medium effect size at a low cost compared to individualized tutoring with a large effect size at a moderate cost.

Incorporating scalability demonstrates again how smaller effect sizes can be more meaningful than larger ones. Implementing individualized tutorials requires schools to reorganize their schedule to incorporate tutoring throughout the school day. Much of the effect of tutoring depends on the ability to recruit, select, train, and support a corps of effective tutors. I would characterize these implementation challenges as nontrivial but reasonable given they do not require major behavioral changes on the part of core school staff. In contrast, a universal free breakfast program requires little skill or training on the part of cafeteria workers and can be provided using existing cafeteria equipment. I would characterize universal free breakfast as easy to scale. The greater likelihood of scaling universal free breakfast programs with fidelity compared to individualized tutoring makes it that much more of a policy-relevant effect.

Two additional examples further highlight the complexity of interpreting effect sizes when their policy relevance varies across multiple dimensions. Fryer and colleagues (2012) studied a pay-for-performance incentive scheme that leveraged loss aversion where teachers were paid a bonus in advance but had to return it if students did not make achievement gains. They found that third- through eighth-grade teachers randomized to the loss aversion policy raised student achievement on state standardized tests in math by 0.22 *SD*. The expected value of the bonus was $4,000, placing the per-pupil cost between $40 and $200 depending on student-teacher ratios. This loss aversion approach to performance pay produced a large effect size at a low cost but may have more limited policy relevance given potential obstacles to replicating these effects at scale. The intervention itself could be implemented at scale with fidelity but may not replicate at scale if other teachers are less motivated by or effective at responding to the loss incentives. Such a program would also appear politically infeasible given likely opposition from teacher unions to an incentive scheme where their members would have to pay back bonuses.

Finally, consider investments in school facilities. Cellini, Ferreira, and Rothstein (2010) used a regression discontinuity design to show that passing bonds to fund school facility investments increased third-grade students' achievement on state standardized tests in math by 0.08 *SD* 6 years after bonds were passed. They also estimated that narrowly passing a bond increased per-pupil spending across these 6 years by a total of approximately $5,000. Thus, the investments in school infrastructure examined in this study produced a medium effect size at a high cost. Despite these high costs, investments in school infrastructure remain policy relevant because of their scalability. Modernizing school buildings is a technical intervention that can be implemented with fidelity at scale. It is also likely to produce similar effects when scaled because the intervention requires few behavioral changes on the part of educators. Finally, investments in school infrastructure are one of the few spending categories for which there is political support across the aisle, although it remains to be seen if communities, states, or the federal government are willing to make these investments.

Conclusion

Rigorous evaluations of education interventions are necessary for evidence-based policy and practice, but they are not sufficient. To effectively inform policy, scholars and policymakers must be able to interpret findings correctly and judge their substantive significance. This is challenging because what, when, and how outcomes are measured as well as subjective decisions researchers make about study design and analyses all shape the magnitude of program effects. This article provides broad guidelines for incorporating these study features along with program

costs and potential for scalability into the interpretation process. I propose interpreting effect size magnitudes relative to the empirical distribution of effects from specific classes of studies and outcome domains. In practice, the vast majority of education interventions fail to produce effects that would even be judged as small by Cohen's standards. We need to update our expectations and consider the multidimensional nature of policy relevance when interpreting program effects. Effect sizes that are equal in magnitude are rarely equal in meaning.

NOTES

Alex Bolves, Halle Bryant, Alvin Christian, Sarah Conlisk, Lucy Duda, Hannah Sexton, and Emily Skahill provided excellent research assistance. I am grateful to Matt Barnum, Drew Bailey, Howard Bloom, Brooks Bowden, Christina Claiborne, Carrie Conaway, Thomas Dee, Angela Duckworth, Avi Feller, Dan Goldhaber, Michael Goldstein, Jonathan Guryan, Doug Harris, Heather Hill, Jing Liu, Susanna Loeb, Katie Lynch, Richard Murnane, Lindsay Page, James Pustejovsky, Todd Rogers, Nathan Schwartz, James Soland, John Tyler, Dylan Wiliams, Jim Wykoff, and David Yeager for their helpful feedback and advice.

[1]These benchmarks are specifically for effect sizes derived from standardized differences in means, which are the focus of this article.

[2]For example, prior studies have focused on defining effect sizes (Kelley & Preacher, 2012), calculating effect sizes (Hedges, 2008; Rosenthal, Rosnow, & Rubin, 2000; Soland & Thum, 2019), illustrating how research designs influence effect sizes (Cheung & Slavin, 2016; Simpson, 2017), developing empirical benchmarks for interpreting effect sizes (Bloom et al., 2008; Hill et al., 2008), translating effect sizes into more intuitive terms (Baird & Pane, 2019; Lipsey et al., 2012), considering cost-effectiveness (Duncan & Magnuson, 2007; Harris, 2009; Levin & Belfield, 2015), and interpreting effect sizes in the fields of child development (McCartney & Rosenthal, 2000) and psychology (Funder & Ozer, 2019).

[3]Calculation based on an average annual gain in effect size from third to fourth grade of 0.52 *SD* in math from Bloom et al. (2008).

[4]For example, second graders typically make average gains of 1.00 *SD* in math over the course of the school year, whereas ninth graders gain only 0.25 *SD* in math, on average. Dividing each of these annual gains by 9 months to arrive at an approximate magnitude of average gains per month of school illustrates that an effect size of 0.20 *SD* in math is less than 2 months of learning for a second grader (0.2 *SD* × [9 months / 1.00 *SD* annual gain]) but over 7 months for a ninth grader (0.2 *SD* × [9 months / 0.25 *SD* gain]).

[5]Calculation based on author's calculation of a 0.85 *SD* Black-White achievement gap in eighth grade math using data from the 2017 National Assessment of Educational Progress.

[6]Calculation based on an estimates of 0.15 *SD* for the standard deviation of teacher effects in math from Hanushek and Rivkin (2010) and 0.11 *SD* for the standard deviation of school effects in math from Grissom, Kalogrides, and Loeb (2015).

[7]This assumes no major threats to the validity of the randomization process or substantially differential attrition.

[8]This first approach is equivalent to Cohen's *d* when the sample size for the treatment and control groups are the same, and the second approach is known as Glass's Δ.

[9]This lower take-up rate is due to some treatment students not taking up the offer of tutoring and others never receiving the offer because they did not return to the school they were enrolled in the previous year.

[10]Per-pupil costs can be converted into per-teacher or per-school costs by making a simple assumption about average class and school sizes.

REFERENCES

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, *130*(3), 1117–1165.

Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy*, *20*(2), 198–212.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, *34*(2), 275–318.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7–39.

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher*, *48*(4), 217–228.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328.

Boulay, B., Goodson, B., Olsen, R., McCormick, R., Darrow, C., Frye, M., Gan, K., Harvill, E., & Sarna, M. (2018). *The Investing in Innovation Fund: Summary of 67 evaluations. Final report. NCEE 2018-4013*. National Center for Education Evaluation and Regional Assistance.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Overview of measuring effect sizes: The effect of measurement error. Brief 2*. National Center for Analysis of Longitudinal Data in Education Research.

Cellini, S. R., Ferreira, F., & Rothstein, J. (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *The Quarterly Journal of Economics*, *125*(1), 215–261.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Chingos, M. M., Whitehurst, G. J., & Gallaher, M. R. (2015). School districts and student achievement. *Education Finance and Policy*, *10*(3), 378–398.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Coe, R. (2002, September). *It's the effect size, stupid: What effect size is and why it is important*. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England.

Cook, T. D. (2001). Sciencephobia. *Education Next*, *1*(3). https://www.educationnext.org/sciencephobia/

Cook, P. J., Dodge, K., Farkas, G., Fryer, R. G., Guryan, J., Ludwig, J., & Mayer, S. (2015). *Not too late: Improving academic outcomes for disadvantaged youth*. Institute for Policy Research Northwestern University Working Paper WP-15-01.

Dee, T. S., & Dizon-Ross, E. (2019). School performance, accountability, and waiver reforms: Evidence from Louisiana. *Educational Evaluation and Policy Analysis*, *41*(3), 316–349.

Duncan, G. J., & Magnuson, K. (2007). Penny wise and effect size foolish. *Child Development Perspectives*, *1*(1), 46–51.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–432.

Frisvold, D. E. (2015). Nutrition and cognitive achievement: An evaluation of the School Breakfast Program. *Journal of Public Economics*, *124*, 91–104.

Fryer, Jr., R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments* (Vol. 2, pp. 95–322). North-Holland.

Fryer, Jr., R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). *Enhancing the efficacy of teacher incentives through loss aversion: A field experiment* (No. w18237). National Bureau of Economic Research.

Fryer, Jr., R. G., & Noveck, M. H. (2020). High-dosage tutoring and reading achievement: Evidence from New York City. *Journal of Labor Economics*, *38*(2), 421–452.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

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

Greenberg, M. T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness*, *10*(1), 40–67.

Goldhaber, D., & Özek, U. (2019). How much should we rely on student test achievement as a measure of success? *Educational Researcher*, *48*(7), 479–483.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis*, *37*(1), 3–28.

Gueron, J. M., & Rolston, H. (2013). *Fighting for reliable evidence*. Russell Sage Foundation.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271.

Harris, D. N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis*, *31*(1), 3–29.

Hattie, J. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. Routledge.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*(1–2), 114–128.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370.

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Honig, M. I. (2006). *New directions in education policy implementation*. SUNY Press.

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement the potential tradeoff between teacher quality and class size. *Journal of Human Resources*, *44*(1), 223–250.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152.

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics*, *131*(4), 1795–1848.

Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reforms? *Teachers College Record*, *110*(8), 1611–1638.

Kraft, M. A. (2015). How to make additional time matter: Integrating individualized tutorials into an extended day. *Education Finance and Policy*, *10*(1), 81–116.

Kraft, M. A. (2018). Federal efforts to improve teacher quality. In R. Hess & M. McShane (Eds.), *Bush-Obama school reform: Lessons learned* (pp. 69–84). Harvard Education Press.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588.

Lee, J., Finn, J., & Liu, X. (2019). Time-indexed effect size for educational research and evaluation: Reinterpreting program effects and achievement gaps in K–12 reading and math. *The Journal of Experimental Education*, *87*(2), 193–213.

Levin, H. M., & Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. *Journal of Research on Educational Effectiveness*, *8*(3), 400–418.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181–1209.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166.

Luyten, H., Merrell, C., & Tymms, P. (2017). The contribution of schooling to learning gains of pupils in Years 1 to 6. *School Effectiveness and School Improvement*, *28*(3), 374–405.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, *71*(1), 173–180.

Murnane, R. J., & Nelson, R. R. (2007). Improving the performance of the education sector: The valuable, challenging, and limited role of random assignment evaluations. *Economics of Innovation and New Technology*, *16*(5), 307–322.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784–793.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start impact study. Final report*. Administration for Children & Families.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *39*(5), 369–393.

Schäfer, T., & Schwarz, M. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, Article 813. https://doi.org/10.3389/fpsyg.2019.00813

Schwartz, A. E., & Rothbart, M. W. (in press). Let them eat lunch: The impact of universal free meals on student performance. *Journal of Policy Analysis and Management*.

Setren, Elizabeth. (2019). *The impact of targeted vs. general education investments: Evidence from special education and English language learners in Boston Charter Schools* (EdWorkingPaper: 19-100). http://www.edworkingpapers.com/ai19-100

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardized effect sizes. *Journal of Education Policy*, *32*(4), 450–466.

Slavin, R. E. (1987). Mastery learning reconsidered. *Review of Educational Research*, *57*(2), 175–213.

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

Soland, J., & Thum, Y.M. (2019). *Effect sizes for measuring student and school growth in achievement: In search of practical significance* (EdWorkingPaper No.19-60). http://edworkingpapers.com/ai19-60

Tanner-Smith, E. E., Durlak, J. A., & Marx, R. A. (2018). Empirically based mean effect size distributions for universal prevention programs targeting school-aged youth: A review of meta-analyses. *Prevention Science*, *19*(8), 1091–1101.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33.

What Works Clearinghouse. (n.d.). *What Works Clearninghouse procedures handbook, Version 4.0*. U.S. Department of Education, Institute of Education Science.

Yeager, D. S., Hanselman, P., Walton, G. M., Crosnoe, R., Muller, C. L., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*, 364–369.

AUTHOR

MATTHEW A. KRAFT, EdD, is an associate professor of education and economics at Brown University, PO Box 1938, Providence, RI 02192; *mkraft@brown.edu*. His research focuses on efforts to improve educator and organizational effectiveness in K–12 public schools.

Manuscript received December 17, 2018
Revisions received August 21, 2019,
and November 4, 2019
Accepted December 17, 2019