ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-10

Interpreting Effect Sizes of Education Interventions

Matthew A. Kraft
Brown University

Researchers commonly interpret effect sizes by applying benchmarks proposed by 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 paper, 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, cost, and scalability into the process of interpreting the policy importance of effect sizes.

VERSION: August 2019

Kraft, Matthew. (2019). Interpreting Effect Sizes of Education Interventions. (EdWorkingPaper: 19-10). Retrieved from Annenberg Institute at Brown University: http://www.edworkingpapers.com/ai19-10

Interpreting Effect Sizes of Education Interventions

Matthew A. Kraft
Brown University

Updated: August 2019

Abstract

Researchers commonly interpret effect sizes by applying benchmarks proposed by 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 paper, 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, cost, and scalability into the process of interpreting the policy importance of effect sizes.

Correspondence regarding the article can be sent to Matthew Kraft at mkraft@brown.edu. Alvin Christian and Alex Bolves provided excellent research assistance. I am grateful to Matt Barnum, 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.

Interpreting Effect Sizes of Education Interventions

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 communicates or consumes 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 more intuitive outcomes such as infection rates or earnings, it remains difficult to compare the relative success of programs evaluated based on 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, Cook et al. (2015) find that integrating intensive individualized tutoring into the school day raised student achievement in math by 0.23 standard deviations (SD), while Frisvold (2015) finds that offering universal free school breakfasts increased achievement in math by 0.09 SD. Are the magnitudes of these impacts substantively meaningful? Should we conclude that individualized tutoring is a better math intervention than universal free breakfast? Answering these questions requires appropriate benchmarks and close attention to the study designs, costs, and scalability.

The default approach to evaluating the magnitude of effect sizes is to apply a set of benchmarks 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. A Google search for "effect size" reveals these benchmarks are ubiquitous online

and even featured on Wikipedia's entry for "Effect size." 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 find that education interventions often result in no effects or effects that would be 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 comparatively large relative to current alternatives. Grantmakers eschew investments in programs that deliver incremental gains in favor of interventions targeting alluringly large, but unrealistic, improvements.

In this paper, 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: 1) five broad guidelines with simple questions and corresponding interpretations for contextualizing effect sizes, and 2) a more structured schema for interpreting effects from a specific class of studies: causal analyses of education interventions with standardized achievement outcomes.

The paper contributes to the effect size literature in several ways. First, I update prior reviews (Coe, 2002; Bloom et al., 2008, Lipsey et al. 2012) with insights from a number of new articles (e.g. Cheung & Slavin, 2016; Simpson, 2017; Soland & Meng Thum, 2019; Lortie-Forgues & Inglis, 2019; Baird & Pane, 2019; Funder & Ozer, 2019; Schäfer & Schwarz, 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 under-recognized 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 describe why translating effects into more intuitive units such as months of learning or percentile changes is not an actual interpretation of the substantive significance of an effect, but can be a useful complementary approach. Next, I introduce the guidelines and schema for interpreting effect sizes, provide an example of how to apply them, 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*:

$$Effect\ Size = \frac{[Mean_1 - Mean_2]}{Standard\ Deviation} \qquad (1)$$

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 in order 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 (e.g. 2009) continue to incorporate these dated studies and ignore the importance of 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 4th, 5th and 8th graders received instruction in probability or cartography for three to four 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 (Gueron & Rolston, 2013; Murnane & Nelson, 2007; Angrist, 2004; Cook, 2001). 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 pre-registering 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. Quantitative research in education has evolved, but we have yet to update Cohen's benchmarks.

Current Approaches to Translating Effect Sizes

While Cohen's benchmarks continue to color our interpretation of effect sizes, scholars have increasingly adopted translational approaches to interpreting effect sizes. These approaches convert effect sizes onto more broadly familiar scales in an effort to provide more intuition about the importance of an effect. Several of these translational approaches are worth highlighting (for detailed descriptions of these techniques see Hill et al., 2008; Lipsey et al., 2012; Baird & Pane, 2019).

Months of learning: Converting effect sizes into months of learning is often favored by policymakers for its intuitive appeal. However, these estimates are highly sensitive to the large differences in learning rates across grade levels, making it an impractical approach for estimates that pool across grades (Baird & Pane, 2019).[3] Equally important, translating effect sizes into months of learning can be misleading because learning rates reflect influences from both inside and outside the classroom, as well as the natural developmental process.

Changes in percentile rank: This 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. However, the total percentile point change is sensitive to the starting percentile one chooses, so it is important to describe both the initial and post-intervention percentiles. For example, the effect of individualized tutoring (0.23 SD) is equivalent to moving male students in distressed Chicago high schools from the 50^{th} to the 59^{th} percentile of achievement.

Achievement gaps: Benchmarking against achievement gaps help to frame effects relative to policy-relevant metrics. For example, the 0.09 SD effect of universal free breakfast on math achievement represents 11 percent of the student-level Black-white achievement gap. Unfortunately, this framing can also mislead people to believe that an intervention would

decrease the Black-white achievement gap by this same magnitude. Whether interventions decrease achievement gaps depends on where they are targeted and their relative effects across different subgroups of students.

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 we are achieving currently within the education system. For example, a 0.09 SD effect is equivalent to the difference between an average teacher and a teacher at approximately the 73rd percentile in the distribution of teacher effectiveness, or between an average school and a school at roughly the 82nd percentile of school effectiveness. However, this approach is sensitive to the estimate one uses for the magnitude of teacher and school effects.

Translating effect sizes onto more intuitive scales can be a helpful, complementary approach to communicating about effect sizes when these conversions are applied with care. No single approach is uniformly better; their value depends on the audience one is trying to reach. But translations are not interpretations. They are simple unit conversions that leave the interpretation to the reader and allow considerable room for disagreement. Additionally, translational approaches assume all effect sizes are directly comparable rather than considering how study features might influence their magnitude. And they say nothing about program costs or scalability, which can have profound implications for understanding the policy-relevance of an effect. It is time we updated and expanded our 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.[4] 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 can be misleading when researchers do not make it clear whether the underlying relationship is correlational or causal. 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., Lipsey & Wilson, 1993; Hattie, 2009) 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, Shavelson, Hamilton, & Klein, 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 two to four times larger than effects on broad standardized state tests (Hill et al., 2008; Lipsey et al., 2012; Cheung & Slavin, 2016; Lynch et al., 2019). These larger effects on researcher-designed, specialized assessments can be misleading when they reflect narrow, non-transferable 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 (Kraft, 2015; Cook et al., 2015; Fryer, in press).

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 (Baily 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 very 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 of the measure, 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) show, measurement error can differ substantially across outcomes. They find that measurement error accounts for 17 percent of the variance in standardized test scores, but 84 percent 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 analysis 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 more effective among lower-achieving students (Paunesku et al., 2015, Yeager et al., in press). 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 holding students back 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). Researchers themselves often recruit participants that they most expect to benefit when first testing the potential efficacy of an intervention. 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, non-representative 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 judgement 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; Hedges, 2007; Dee & Dizon-Ross, 2019). 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: 1) the complete analytic (i.e., pooled) sample, 2) the control group sample, and 3) an estimate from a larger population.[5] For example, the effect of individualized tutoring in Cook et al. (2015) of 0.23 SD uses the control group sample. They also report effects scaled by the national distribution of test scores, which reduces 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 SD) than students in an unrestricted national sample. When baseline outcome measures are not available, it is preferable to use the SD 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 the Perry Preschool than for the Head Start program (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 tutoring, 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 contrasted with 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 do have access to programs, resources, or supports 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, the 0.23 SD effect on math achievement represents the effect of receiving tutoring. However, only 41 percent of all students who were randomly assigned to be offered tutoring took up this offer.[6] 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 an intervention to have smaller effect sizes than studies that report the effect of receiving an 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 upfront costs. Spending the marginal dollar on the most cost-effective program makes sense. Upfront fixed costs are also an important feature of education programs. The financial implications of reforms

that require large initial capital investments, such as modernizing school facilities, are very different from programs where costs can be amortized over longer periods and are flexible with scale, such as expanding school breakfast programs. Policymakers have to consider not only what works, but also how well it works relative to costs and what immediate financial investments are 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 non-monetary costs, such as the opportunity costs of educators' time, would go even farther 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 more costly 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: 1) program effects are often heterogeneous, and 2) programs are often difficult to replicate with fidelity at scale. As discussed above, impressive effects from non-representative 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.

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). Similarly, efforts to reduce class sizes statewide in California did not result in the large academic gains found in the Tennessee STAR class size experiment (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 include 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 effect 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).

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, are prohibitively costly, 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 fail to account for important contextual differences across studies or to reflect 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 solution to this tension is for researchers to identify benchmarks for specific classes of studies based on the distributions of effects from the relevant literature (e.g., Tanner-Smith, Durlak, & Marx, 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?" These sets of benchmarks would provide much-improved interpretations that we can and should refine based on the characteristics of individual studies and as more research becomes available.

The schema I propose provides new 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 (Chetty, Friedman, & Rockoff, 2014). 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 benchmarks for effect sizes from causal studies of pre-K–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 (see Appendix A for source data and coding details). 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 5th 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 percent of these achievement gains (Konstantopolus & Hedges, 2008; Chingos, Whitehurst, & Gallaher, 2015; Luyten, Merrell & Tymms, 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. Additionally, consider this: raising student achievement by 0.20 SD results in a 2 percent increase in annual lifetime earnings on average (Chetty, Friedman, & Rockoff, 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 rate of failure among interventions, given publication bias again null findings.

Adapting the Benchmarks

The proposed benchmarks provide a general heuristic for interpreting effect sizes from causal studies of education interventions with pre-K–12 achievement outcomes. In Table 1, I explore how we might adapt these benchmarks to account for effect size heterogeneity across subjects, grades, and select study characteristics. 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 pre-kindergarten through 3rd grade. This is evident in Figure 1, which depicts the median and interquartile range (25^{th} to 75^{th} percentiles) of effect sizes in math (Panel A) and reading (Panel B) across grade levels (see Appendix Table B1 for specific statistics).

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 pre-kindergarten (median 0.12 SD), and are similar in magnitude to effect sizes in reading across 4^{th} to 12^{th} grade (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 pre-kindergarten and lower elementary grades in reading. One might adjust benchmarks for evaluating effect sizes on assessments of early literacy upward to, say, 0.10 and 0.30 SD.

Similar to prior studies, we find further evidence that larger studies with broad achievement measures produce 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). And effect sizes on broad achievement measures 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 pre-register their research design and report their findings, have a median effect size of 0.03 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 pre-registered and examine broad achievement measures.

Incorporating Costs and Scalability into Policy Decisions

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 effect size benchmarks with a corresponding set of empirically-based per-pupil cost benchmarks: less than $500 is *Low*, $500 to under $4,000 is *Moderate*, and $4,000 or greater is *High* (see Appendix Table C1 for more details).[7] Given that these cost benchmarks are derived from a sample of only 68 education interventions, they should be viewed as only a rough guide for classifying effect sizes into the simple cost-effectiveness ratios shown in this 3x3 matrix.

The matrix 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 as policy decisions depend on local priorities, resources, and politics as well.

The last 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 judgement of the interpreter following the guidance I provide above. Reasonable people will disagree about program scalability. 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- 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.

An Example

Consider, for example, the previously cited studies evaluating the effects of universal free breakfast (0.09 SD) and individualized tutoring (0.23 SD). In many ways, these studies share similar core features. Both studies employ causal methods and examine 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 analyze sizable samples of over 2,000 students in grades (4^{th}/5^{th} vs. 9^{th}/10^{th}) where there are few systematic differences in the average effect size of education interventions.

However, differences in sample characteristics and analytical approaches, costs, and scalability all indicate these effect sizes might be more similar in practical importance than their magnitudes suggest. Cook et al. (2015) target their tutoring study to male youth of middling achievement in distressed Chicago high schools, a narrow population for which the intervention is specifically designed and in which there is less variance in outcomes. They also focus on the effect of *receiving* tutoring, whereas Frisvold (2015) reports 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, 2017). Cook et al. (2015) report 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 non-trivial, but reasonable, given they don't 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 the existing equipment in school cafeterias. 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.

Conclusion

Rigorous evaluations of education interventions are necessary for evidence-based policy and practice, but they are not sufficient. To inform policy, scholars and policymakers must be able to interpret findings 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 analysis, all shape the magnitude of program effects. This article provides broad guidelines for incorporating study features into the interpretation process. It also proposes a new, more detailed schema with empirical benchmarks that reflect how the vast majority of education interventions fail or only produce effects that would be judged as small by Cohen's

standards. Interpreting the policy relevance of effects requires that we update our expectations as well as consider program costs and scalability. Effect sizes that are equal in magnitude are rarely equal in importance.

Endnotes

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

[2] For example, prior studies have focused on defining effect sizes (Kelley & Preacher, 2012), calculating effect sizes (Rosenthal, Rosnow, & Rubin, 2000; Hedges, 2008; Soland & Meng 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 (Lipsey et al., 2012; Baird & Pane, 2019), 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] For example, 2nd graders typically make average gains of 1.00 SD in math over the course of the school year, while 9th 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 2nd grader (0.2 SD * [9 months / 1.00 SD annual gain]) but over 7 months for a 9th grader (0.2 SD * [9 months / 0.25 SD gain]).

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

[5] 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.

[6] 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.

[7] 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., Fullerton, J. B., Kane, T. J., Pathak, P. A., & Walters, C. R. (2011). Student achievement in Massachusetts' charter schools. *Cambridge, MA: Center for Education Policy Research at Harvard University*.

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., ... & 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*.

Bowden, A.B., Belfield, C.R., Levin, H.M., Shand, R., Wang, A. & Morales, M., 2015. A benefit-cost analysis of City Connects. *Center for Benefit-Cost Studies in Education: Teachers College, Columbia University*.

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*.

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), 2633-79.

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.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences* (1st ed.). New York: Academic Press.

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

Cook, T. D. (2001). Sciencephobia. *Education Next*, *1*(3). Retrieved from educationnext.org/

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 *Handbook of Economic Field Experiments* (Vol. 2, pp. 95-322). North-Holland.

Fryer Jr, R. G., & Noveck, M. H. (in press). High-Dosage Tutoring and Reading Achievement: Evidence from New York City. *Journal of Labor Economics*.

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.* Beverly Hills, CA: 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.

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.

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*. Abingdon: 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.

Hollands, F. M., Kieffer, M. J., Shand, R., Pan, Y., Cheng, H., & Levin, H. M. (2016). Cost-effectiveness analysis of early reading programs: A demonstration with recommendations for future research. *Journal of Research on Educational Effectiveness*, *9*(1), 30-53.

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

Jacob, R., Armstrong, C., Bowden, A. B., & Pan, Y. (2016). Leveraging volunteers: An experimental evaluation of a tutoring program for struggling readers. *Journal of Research on Educational Effectiveness*, *9*(sup1), pp.67-92.

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.

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 Hess R. & McShane, M. (Editors). Bush-Obama School Reform: Lessons Learned. Harvard Education Press. 69-84.

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.

Levin, H. M., Belfield, C., Hollands, F., Bowden, A.B., Cheng, H., Shand, R., Pan, Y., & Hanisch-Cerda, B. (2012). Cost–effectiveness analysis of interventions that improve high school completion. *Teacher College, Columbia University*.

Levin, H. M., Catlin, D., & Elson, A. (2007). Costs of implementing adolescent literacy programs. Informed choices for struggling adolescent readers: A research-based guide to instructional programs and practices, 61-91.

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

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & 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. & Inglish, 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., Shapiro, G., Broene, P., ... & Ciarico, J. (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*, 813.

Schwartz, A. E., & Rothbart, M. W. (2017). Let Them Eat Lunch: The Impact of Universal Free Meals on Student Performance. Working Paper.

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). Retrieved from Annenberg Institute at Brown University: http://www.edworkingpapers.com/ai19-100

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised 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). Retrieved from Annenberg Institute at Brown University: 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.

Washington State Institute for Public Policy. *Pre-K to 12 education benefit-cost/meta-analytic results.* Olympia, WA: Author. Information retrieved 2018, November.

Yeager, D. S., Hanselman, P., Walton, G. M., Crosnoe, R., Muller, C. L., Tipton, E., … Dweck, C. S. (Forthcoming). A national experiment reveals where a growth mindset improves achievement. *Nature*.

Tables

Table 1. *Empirical Distributions of Effect Sizes from Randomized Control Trials of Education Interventions with Standardized Achievement Outcomes*

	Overall	Subject		Sample Size					Scope of Test		DoE Studies
		Math	Reading	≤100	101 to 250	251 to 500	501 to 2,000	>2,000	Broad	Narrow	
Mean	0.16	0.11	0.17	0.30	0.16	0.16	0.10	0.05	0.14	0.25	0.03
Mean (weighted)	0.04	0.03	0.05	0.29	0.15	0.16	0.10	0.02	0.04	0.08	0.02
Std	0.28	0.22	0.29	0.41	0.29	0.22	0.15	0.11	0.24	0.44	0.16
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 (# of effect sizes)	1942	588	1260	408	452	328	395	327	1352	243	139
n (# of studies)	747	314	495	202	169	173	181	124	527	91	49

Notes: A majority of the standardized achievement outcomes (95%) are based on math and ELA test scores, with the remaining based on science, social studies, or general achievement. Weights are based on sample size for weighted mean estimates. See Appendix A for details about data sources

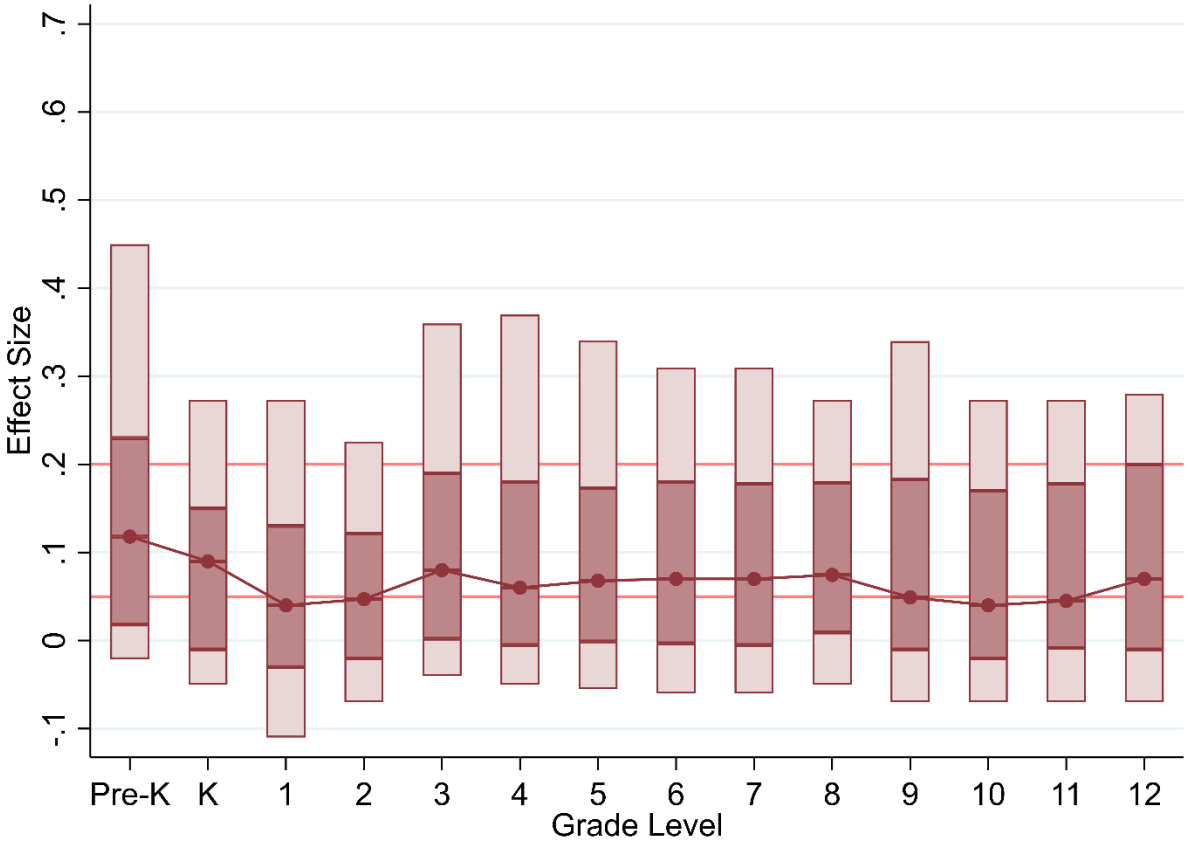
Table 2. *A Schema for Interpreting Effect Sizes from Causal Studies of Education Interventions with Standardized Achievement Outcomes*

		Cost-Effectiveness Ratio				Scalability
		Cost Per Pupil				
		Low (< $500)	Moderate ($500 to <$4,000)	High ($4,000 or >)		
Effect Size	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 or >)	**Large ES / Low Cost**	**Large ES / Moderate Cost**	**Large ES / High Cost**		**Hard to Scale**

Notes: ES = Effect Size. Green and red shading represent higher and lower cost-effectiveness ratios, respectively.

Figures

Panel A. Math



Panel B. Reading

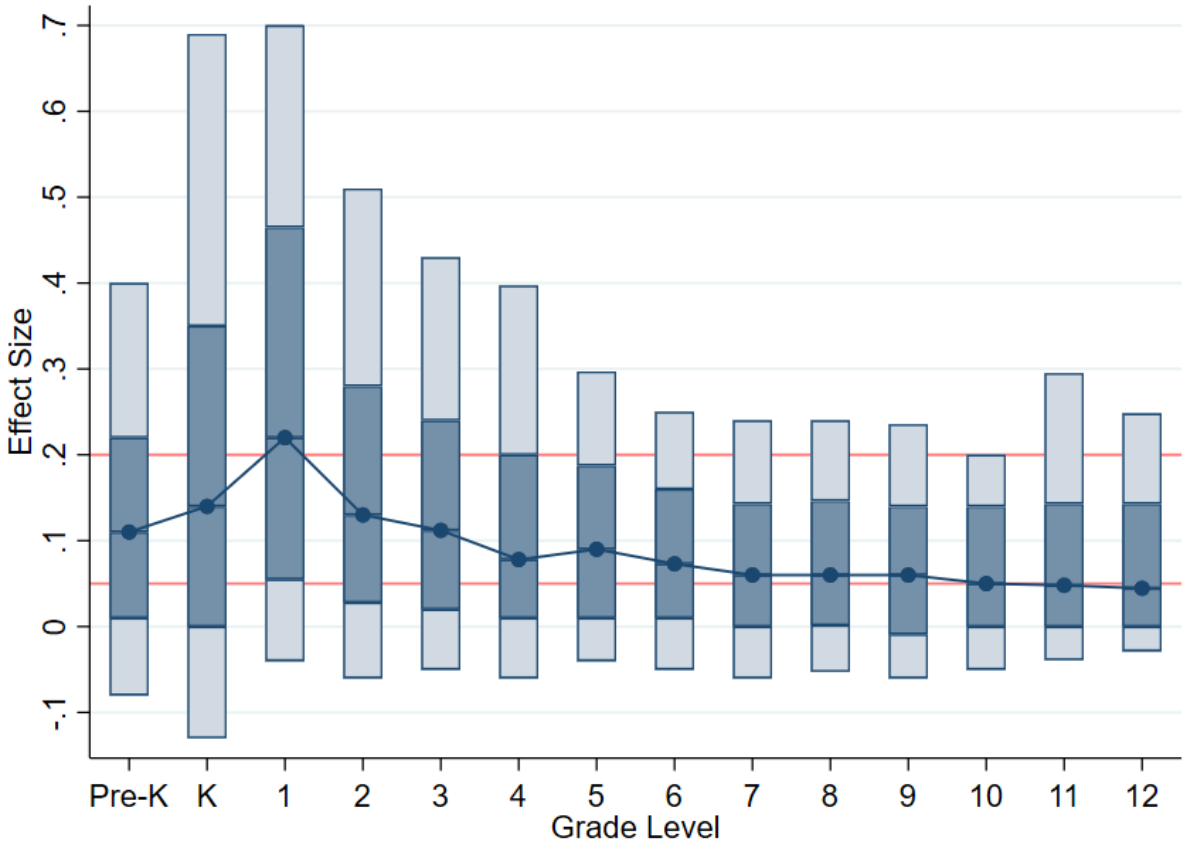


Figure 1. The distribution of effect sizes from RCTs of education interventions with student achievement outcomes by subject and grade level.

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

Appendix A

Effect Size Source Data and Coding

A. Data

I use six main sources to collect effect size outcomes from randomized controlled trials in education. These sources include five reports conducted by third parties which evaluate the impact of interventions studied using public grant money (4 US based; 1 UK based) and a textbook focused on implementing rigorous field experiments. The minimum effect sizes from each study range from -1.00 to -0.47 while the maximum effect sizes range from 0.51 to 2.85. In Table A1, I provide detailed descriptions of each data source with further summary statistics.

B. Sample

Drawing on these data sources, I restrict my analytic sample to include only effect sizes from studies that are 1) education interventions, 2) randomized controlled trials, and 3) use as the outcome a standardized test. These restrictions result in an analytic sample of 1,942 effect sizes from an initial sample of 2,528. Over 98 percent of these effects are estimated from student-level data, suggesting they overwhelmingly reflect student-level standard deviations.

Studies often reported effect sizes across a range of standardized and unstandardized measures. My research team and I excluded all non-test outcomes and any test outcome that was not standardized (such as researcher designed instruments). We used a unique study ID based on publication year and author last name(s) to remove duplicates introduced as a result of one study being reported in multiple sources. Table A1 provides the final counts for the number of effect sizes (n) and studies (k) in each source and also includes effect sizes for the mean, 33rd, 50th, and 66th percentile values for each data set.

Table A1. *Description of sources used to collect effect size outcomes*

Source	Description	Effect Sizes	Studies	Mean	Mean (weighted)	Percentiles		
						33rd	50th	66th
Handbook of Field Experiments, Vol. 2	Handbook of Field Experiments, Volume Two explains how to conduct experimental research, presents a catalog of research to date, and describes which areas remain to be explored. Chapter two looks at the findings from 196 randomized field experiments specifically in education.	347	190	0.15	0.04	0.04	0.10	0.16
Best Evidence Encyclopedia	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 under funding from the Institute of Education Sciences, U.S. Department of Education. It is intended to give educators and researchers fair and useful information about the strength of the evidence supporting a variety of programs available for students in grades K-12.	871	379	0.14	0.09	0.05	0.11	0.18
IES WWC Database	The What Works Clearinghouse is an investment of the Institute of Education Sciences (IES) within the U.S. Department of Education that was established in 2002. The work of the WWC is managed by a team of staff at IES and conducted under a set of contracts held by several leading firms with expertise in education, research methodology, and the dissemination of education research.	506	162	0.21	0.01	0.04	0.12	0.22
IES Commissioned RCTs 2002-2013	This report published by the Coalition for Evidence-Based Policy highlights key findings from 90 interventions that have been evaluated in IES-commissioned RCTs.	85	28	0.01	0.01	-0.05	0.02	0.05
Investing in Innovation Evaluations	Evaluations from the Investing in Innovation Fund, which provides competitive grants to local education agencies and nonprofits to implement and evaluate educational interventions. All interventions are evaluated by outside organizations.	54	21	0.06	0.03	-0.03	0.04	0.12
Education Endowment Foundation	The Education Endowment Foundation was established in 2011 by The Sutton Trust, as a lead charity in partnership with Impetus Trust (now part of Impetus - The Private Equity Foundation) with a £125m founding grant from the Department for Education. The EEF and Sutton Trust are, together, the UK government-designated What Works Centre for Education.	79	72	0.20	0.07	0.07	0.17	0.22

Notes: Duplicate studies and effect sizes from different data sources were dropped.

C. Codes

After compiling the analytic sample, my research team and I coded these data for a range of characteristics including study sample size, grade level, subject, and whether a test was narrow or broad.

We created indicator variables to identify which grades each study focused on. Many of the interventions ranged across multiple grades and only presented overall effect sizes. In these cases, we included effects sizes in all grade-level groups that are represented in each sample. In cases where effect sizes were listed separately by grade, they are included as separate observations. The result is that many effect sizes are not mutually exclusive by grade across the sample. Of the 1,942 effect sizes in the analytic sample, 1,017 are for single grades, 299 are with two grades, 226 are with three grades, 167 are with four grades, and the remaining 233 are associated with four or more grades.

Following Hill et al. (2007), we distinguished between standardized tests that cover a broad subject matter and more narrow standardized tests. Studies often reported effect sizes for broad overall test scores and for scores from more narrow subdomains. To ensure these non-independent effect sizes were not double counted, we included only the overall standardized score when the overall effect sizes and subdomain effect sizes were both reported. Table A2 provides several examples of how we coded studies as broad and narrow.

Table A2. *Examples of broad and narrow standardized outcomes*

Broad Standardized Measures	Narrow Standardized Measures
Test of Preschool Early Literacy (TOPEL) Comprehensive Score	TOPEL Phonological Awareness TOPEL Print Knowledge TOPEL Definitional Vocabulary
Gates Macginitie Total Score	Gates MacGinitie Vocabulary Gates MacGinitie Comprehension Gates MacGinitie Word Decoding
California Achievement Test (CAT): Total mathematics subscore	CAT: Mathematics application CAT: Mathematics computation CAT: Mathematics concepts
Woodcock Johnson-III Math Score	WJ-III Math Fluency WJ-III Quant Concepts WJ-III Math Reasoning
- *No associated broad measure* -	Virginia Standards of Learning Algebra I Test McGraw-Hill Algebra Proficiency Exam Test of Economic Literacy

Appendix B

Appendix Table B1. *Empirical Distributions of Effect Sizes from Randomized Control Trials of Education Interventions with Standardized Achievement Outcomes by Grade and Subject*

	Pre-K	K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12
	Panel A: Math													
Mean	0.18	0.08	0.06	0.06	0.11	0.11	0.10	0.10	0.11	0.11	0.09	0.08	0.08	0.09
Mean (weighted)	0.07	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.00	-0.01	-0.01	-0.01
Std	0.27	0.14	0.15	0.13	0.16	0.20	0.18	0.17	0.19	0.17	0.22	0.23	0.19	0.21
P1	-0.24	-0.33	-0.23	-0.22	-0.20	-0.30	-0.30	-0.29	-0.30	-0.29	-0.44	-0.78	-0.78	-0.78
P10	-0.02	-0.05	-0.11	-0.07	-0.04	-0.05	-0.05	-0.06	-0.06	-0.05	-0.07	-0.07	-0.07	-0.07
P20	0.01	-0.02	-0.05	-0.03	-0.01	-0.02	-0.02	-0.02	-0.02	0.00	-0.03	-0.04	-0.03	-0.03
P30	0.03	0.01	-0.02	-0.01	0.02	0.01	0.01	0.02	0.01	0.03	0.00	0.00	0.00	0.00
P40	0.07	0.04	0.02	0.02	0.04	0.03	0.04	0.04	0.04	0.05	0.02	0.02	0.02	0.03
P50	0.12	0.09	0.04	0.05	0.08	0.06	0.07	0.07	0.07	0.07	0.05	0.04	0.04	0.07
P60	0.16	0.11	0.08	0.09	0.11	0.10	0.10	0.10	0.10	0.11	0.09	0.09	0.09	0.14
P70	0.22	0.13	0.11	0.11	0.15	0.15	0.13	0.13	0.15	0.15	0.16	0.14	0.16	0.18
P80	0.30	0.19	0.15	0.15	0.23	0.22	0.21	0.20	0.22	0.21	0.21	0.20	0.21	0.21
P90	0.45	0.27	0.27	0.23	0.36	0.37	0.34	0.31	0.31	0.27	0.34	0.27	0.27	0.28
P99	1.40	0.40	0.48	0.40	0.75	0.77	0.75	0.91	0.91	0.91	0.73	1.09	0.62	0.62
k (# effect sizes)	61	81	138	164	156	209	204	158	153	172	114	81	72	61
n (# of studies)	22	49	69	80	95	123	123	87	77	90	65	48	39	34
	Panel B: Reading													
Mean	0.13	0.21	0.29	0.17	0.15	0.12	0.11	0.10	0.09	0.09	0.08	0.08	0.11	0.10
Mean (weighted)	0.10	0.02	0.04	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	-0.01
Std	0.22	0.41	0.39	0.24	0.21	0.23	0.15	0.14	0.14	0.15	0.16	0.17	0.20	0.20
P1	-0.38	-0.54	-0.45	-0.38	-0.38	-0.38	-0.15	-0.18	-0.21	-0.21	-0.24	-0.21	-0.17	-0.17
P10	-0.08	-0.13	-0.04	-0.06	-0.05	-0.06	-0.04	-0.05	-0.06	-0.05	-0.06	-0.05	-0.04	-0.03
P20	-0.03	-0.02	0.03	0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	0.00
P30	0.02	0.03	0.09	0.05	0.03	0.02	0.03	0.02	0.01	0.01	0.00	0.01	0.01	0.01
P40	0.06	0.08	0.16	0.08	0.07	0.04	0.05	0.05	0.04	0.04	0.03	0.03	0.03	0.03

P50	0.11	0.14	0.22	0.13	0.11	0.08	0.09	0.07	0.06	0.06	0.06	0.05	0.05	0.04
P60	0.15	0.20	0.29	0.19	0.15	0.12	0.12	0.11	0.09	0.10	0.10	0.09	0.10	0.10
P70	0.18	0.29	0.41	0.24	0.20	0.17	0.16	0.14	0.12	0.13	0.12	0.12	0.13	0.13
P80	0.27	0.43	0.52	0.32	0.28	0.23	0.22	0.18	0.18	0.18	0.15	0.15	0.16	0.15
P90	0.40	0.69	0.70	0.51	0.43	0.40	0.30	0.25	0.24	0.24	0.24	0.20	0.29	0.25
P99	0.87	1.73	1.73	0.87	0.73	0.93	0.57	0.59	0.58	0.59	0.75	0.93	0.93	0.93
k (# effect sizes)	221	321	364	286	263	215	220	193	185	164	100	57	34	30
n (# of studies)	65	104	167	139	145	130	121	110	96	84	54	34	20	18

Notes: Effect sizes across grades are not mutually exclusive as many studies present one effect sizes pooled across multiple grades. Weights are based on sample size for weighted mean estimates.

Appendix C

Cost Data

I report per-pupil costs in 2016 dollars from 68 education interventions. My research team and I gathered information on the costs of education interventions from the Washington State Institute for Public Policy (WSIPP) and selected studies that included information on intervention costs. Relevant to policymakers, we included costs from a broad range of education interventions that many states are currently considering, such as full-day kindergarten and teacher performance pay programs.

Appendix Table C1. *Empirical Distributions of Program Costs from Education Interventions*

Percentile	Per-Pupil Cost
Mean	$4,752
Std	$9,720
P1	$18
P10	$77
P20	$121
P30	$210
P40	$301
P50	$882
P60	$1,468
P70	$3,150
P80	$7,259
P90	$15,530
P99	$61,248
n	68

Notes: Costs are calculated in 2016 dollars based on interventions from the Washington State Institute for Public Policy (2018), Harris (2009), Cook et al. (2015), Bowden et al. (2015), Jacob et al. (2016), Levin, Catlin, & Elson (2007), Levin et al. (2012), and Hollands et al. (2016).

A. WSIPP

Approximately three-quarters of my data on costs come from the WSIPP, a nonpartisan public research group whose purpose is to identify evidence-based policies that Washington

State can implement to improve statewide outcomes and efficiently use taxpayer dollars. The organization conducts meta-analyses of a range of Pre-K–12 interventions and reports how much it would cost to implement a particular intervention in the state of Washington.

B. Other

The rest of my cost data comes from Harris (2009), Cook et al. (2015), Bowden et al. (2015), Jacob et al. (2016), Levin, Catlin, and Elson (2007), Levin et al. (2012), and Hollands et al. (2016). I use these studies because they report costs for varied interventions like tutoring for struggling elementary school students (Jacob et al., 2016) to more broad service programs that address academic, health, emotional, and family needs (Bowden et al., 2015).