Interpreting Effect Sizes of Education Interventions

Matthew A. Kraft
Brown University

December 2018

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 often large in the context of field-based education interventions. This focus on magnitude also obscures important differences in study features, program costs, and scalability. In this paper, I propose a new framework for interpreting effect sizes of education interventions, which consists of five broadly applicable guidelines and a detailed schema for interpreting effects from causal studies with standardized achievement outcomes. The schema introduces new effect-size and cost benchmarks, while also considering program scalability. Together, the framework provides scholars and research consumers with an empirically-based, practical approach for interpreting the policy importance of effect sizes from education interventions.

Suggested Citation:
Kraft, M.A. (2018). Interpreting Effect Sizes of Education Interventions. *Brown University Working Paper*.

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, Brooks Bowden, Carrie Conaway, Thomas Dee, Angela Duckworth, Avi Feller, Dan Goldhaber, Michael Goldstein, Jonathan Guryan, Doug Harris, Susanna Loeb, Richard Murnane, Lindsay Page, Todd Rogers, Nathan Schwartz, John Tyler, Dylan Wiliams, and David Yeager for their feedback on earlier drafts.

The ability to make empirical analyses accessible and meaningful for broad audiences is a critical skill in academia. However, this can be a substantial challenge when outcomes are measured in unintuitive units such as scale score points. Even when outcomes are measured in more familiar units such as GPA or days absent, it remains difficult to compare the relative success of programs that are evaluated using 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 intervention than universal free breakfast? Answering these questions requires realistic benchmarks and close attention to the study designs, costs, and potential scalability.

In this paper, I develop a new framework for interpreting effect sizes of education interventions 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) a set of five broad guidelines with simple questions and corresponding interpretations for contextualizing effect sizes, and 2) a detailed schema specifically for interpreting effects from causal studies with standardized achievement outcomes. The schema integrates new empirically-based benchmarks for effect sizes and costs

into a joint matrix for evaluating the importance of research findings. Together, these guidelines and schema build on a range of insights from the effect-size literature that have often been considered in isolation.[1] They also highlight the under-recognized importance of program scalability and political feasibility for interpreting the policy relevance of research findings.

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).[2] These standards are based on small social psychology lab experiments from the 1960s performed largely on undergraduates. Cohen's conventions continue to be taught and used widely across the social sciences today, although 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 potentially effective programs or practices which have traditionally small, but actually meaningful effects. Grant makers dismiss interventions that produce more incremental gains in favor of interventions targeting alluringly large, but unrealistic, improvements.

[1] For example, prior studies have focused on interpreting effect sizes statistics (Rosenthal, Rosnow, & Rubin, 2000; Hedges, 2008), illustrating how research designs influence effect sizes (Cheung & Slavin, 2016; Simpson, 2017), using empirical benchmarks for interpreting effect sizes (Bloom et al., 2008; Lipsey et al., 2012), considering cost-effectiveness (Duncan & Magnuson, 2007; Harris, 2009; Levin & Belfield, 2015), and interpreting effect sizes in the field of child development research (McCartney & Rosenthal, 2000).

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

In recent years, the field has made progress towards recalibrating scholars' expectations for effect sizes of education interventions. The What Works Clearinghouse (WWC), a federal repository of "gold-standard" evidence on education programs, now characterizes effect sizes of 0.25 SD or larger are "substantively important" (WWC, 2014 p.23). Recent meta-analyses of education interventions suggest even this revised benchmark reflects inflated expectations for well-designed field experiments (Cheung & Slavin, 2016; Fryer, 2017). Education interventions often fail or have quite small effects. Less than one out of every six education programs that won scale-up grants from the federal Investing in Innovation Fund (i3) produced statistically significant positive impacts (n=67). Moreover, interpreting the relevance of an effect size for policy and practice requires that we consider a program's cost relative to its benefits as well as its potential to scale under ordinary circumstances.

In what follows, I provide a brief summary of the evolution of education research, which serves to illuminate the origins of many common misinterpretations of effect sizes. Next, I introduce the guidelines and schema for interpreting effect sizes and conclude by discussing the implications of the proposed guidelines for policy and practice.

Effect Sizes and the Evolution of Education Research

Until the mid-20th century, researchers often evaluated the importance of findings from education research 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 finding. As the importance of interpreting the magnitude of findings in education research became more widely recognized, scholars began reporting results as effect sizes, a standardized measure of differences in means. 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 and remain the most common benchmarks for interpreting the standardized effect size statistic, or Cohen's *d*, today.

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 Mark Lipsey and David Wilson (1993) found a mean effect size of precisely 0.5 SD. However, many of the research studies included in the 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). Even recent reviews, such as those by Hattie (2009), continue to incorporate these dated studies and suggest that large effect sizes are common 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 effect sizes, despite early objections (Slavin, 1987), and remain influential today.

At the turn of the 21st century, a growing emphasis on causal inference across the social sciences began to reshape quantitative education research (Gueron & Rolston, 2013). Scholars called for wider adoption of causal methods in education research (Angrist, 2004; Cook, 2001; Murnane & Nelson, 2007). The newly established Institute of Education Sciences (IES) provided substantial federal funding for large-scale randomized field trials and the U.S. Department of Education increasingly required rigorous evaluations of grant-funded programs. Findings from this new generation of causal field-based studies were uniformly more modest in size. For example, Mark Lipsey and his colleagues (2012) found an average effect size of only 0.28 SD among a sample of 124 randomized trials. This was half of the average effect size Lipsey had found in the meta-analysis he conducted with Wilson in 1993. Quantitative research in education has evolved, yet we continue to interpret effect sizes using outdated and "somewhat arbitrary" benchmarks (p.146, Cohen, 1962) from tightly-controlled lab settings. 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 which represent the standardized mean difference between treatment and control groups in randomized controlled trials (RCT). Random assignment eliminates any systematic differences between groups so any subsequent differences can be attributed to the

intervention.[3] However, education researchers also report descriptive differences in means, changes in group performance over time, and estimates from multiple regression models as effect sizes.

Any relationship between two variables, such as height and achievement, can be converted into an effect size. Correlation coefficients themselves are a type of effect size. These descriptive effect sizes provide useful information, but can be misleading when researchers do not make it clear whether their underlying relationship is correlational or causal. Taller students have higher achievement because they are older, on average, not because of their stature. Interpreting spurious relationships as causal effects can misinform policymakers' decisions.

Knowing whether an effect size represents a causal or correlational relationship also matters for interpreting its magnitude. The meta-analytic reviews by Lipsey and his colleagues (1993, 2012) illustrate how correlational relationships are, on average, larger than corresponding causal relationships. It is incumbent on researchers reporting effect sizes to clarify which type of effect size their statistic describes, and it is important that research consumers do not assume effect sizes represent causal relationships.

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

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

Expect effect sizes to be larger for descriptive and correlational studies than causal studies.

2) The magnitude of effect sizes depends on what outcomes are evaluated and when these outcomes are measured

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

Studies are more likely to find larger effects on outcomes that are easier to change, proximal to the intervention, and administered soon after the intervention is completed (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). In the context of education, outcomes that reflect short-term decision making and effort, such as passing a class in a summer credit recovery program, are easier to influence than outcomes that are the culmination of years of decisions and effort, such as high-school graduation. 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' self-reported or subjectively assessed 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 (Lipsey et al., 2012; Cheung & Slavin, 2016). 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 daily tutoring over an entire school year (Kraft, 2015; Cook et al., 2015; Fryer, in press).

When an outcome is measured is also directly related to 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, a study of the effect of attending high-performing charter high schools in Boston using lottery admissions shows

larger effects on contemporaneous achievement outcomes, but relatively more moderate effects on college going outcomes (Angrist et al., 2016). 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 impact an outcome. The further down this causal chain, the smaller the effect sizes are likely to be.

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.*

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.*

3) Effect sizes are impacted by subjective decisions researchers make about the study design and analyses

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 for historically marginalized students (Paunesku et al., 2015). 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, while more targeted interventions, such as holding students back a grade, can only be studied among restricted samples (Greenberg & Abenavoli, 2017). Targeted interventions generally produce larger effect sizes than universal ones for two reasons: 1) they narrow the population of study participants to those most likely to benefit, and 2) there is less variation in outcomes among the subset of eligible participants in targeted interventions relative to universal interventions (more on this point below).

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, creating selection bias (Allcott, 2015). Researchers themselves often recruit participants that they most expect to benefit when first testing the potential efficacy of an intervention. From a policy perspective, it might make sense to target the group most likely to benefit from a program given limited resources. However, we are interested typically in a program's impact implemented at scale for all students under standard conditions.

ASK: *Are study participants a broad sample or a sub-group most likely to benefit from the intervention?*

INTERPRET: *Expect studies with more targeted samples to have larger effect sizes.*

Estimating the standard deviation of the outcome

Small details about how researchers decide to standardize an estimated difference in means can have large consequences for the magnitude of the corresponding effect size. The sample researchers use to calculate the standard deviation of the outcome will affect the effect size. Three common approaches are to use: 1) the complete analytic (i.e., pooled) sample, 2) the

control group sample only, and 3) an estimate from a larger population.[4] In most cases, the choice is a subjective decision. 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 measures of outcomes 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.[5]

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

INTERPRET*: Expect studies that use more homogeneous and less representative samples to have smaller variation in their outcomes and, thus, larger effect sizes.*

The treatment intensity difference between treatment and the control groups

For RCTs, the treatment intensity — or the contrast between the experiences of the treatment and control groups — plays an important role in determining effect sizes. For example, some 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). With more recent studies, such as the nationally representative Head Start Impact Study, a sizable minority

[4] 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.

[5] The variability, and thus the SD, of an outcome measure is also affected by its reliability. Less reliable measures combine actual scores with more measurement error that further extends the range and variability of these measures. Any increase in the variability of an outcome measure, such as that caused by measurement error, will result in systematically smaller effect sizes.

of students in the control group were also enrolled in center-based care (Puma et al., 2010). Here the difference in child-care experiences between the treatment and control groups was less pronounced because some children in the control group also received center-based care. This weaker treatment intensity in the Head Start study is one of several important differences that may explain why Perry Preschool had substantially larger effects on high school graduation rates than Head Start.

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 two different approaches to managing student behavior, a new approach contrasted with the current approach or "business as usual." 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 group that did receive the intervention from the group that did not?*

INTERPRET: *Expect studies to have larger effect sizes when control groups do not have access to resources or supports similar to the intervention.*

The question the treatment effect answers

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 in this context 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 and total upfront costs. Studies increasingly include a back-of-the envelope estimate of per-participant costs, which serves to contextualize the return of an education intervention. Comprehensive cost-effectiveness analyses that account for both monetary and non-monetary costs, such as the opportunity costs of educators' time, provide policymakers with valuable evidence for making difficult decisions with limited resources.

[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.

Several recent studies of programs designed to provide personalized information to parents and students illustrate the importance of considering costs. Doss, Fahle, Loeb and York (2018) find that a text-messaging program designed to help parents of preschoolers support their children's literacy development raised literacy achievement by 0.11 to 0.15 SD. Rogers and Feller (2018) find that delivering personalized information to parents about their students' absenteeism reduced total absences by 6 percent. Page and Gehlbach (2017) found that sending rising college freshman personalized text messages about navigating the transition to college increased the likelihood these freshmen enrolled by 3.3 percentage points. While the magnitude of all these findings are meaningful in their own right, they become that much more impressive considering they cost only $1 to $15 per student.

Upfront fixed costs are also a key and variable 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.

Considering costs is a central and often overlooked element of interpreting effect sizes. Spending the marginal dollar on the most cost-effective program make sense. 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 that require lower costs to implement are more impressive, even when they may be considered small.*

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 an effect size, the greater its 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 i3-funded education interventions chosen because of their prior evidence of success and potential for implementation at scale, only 12 were scaled successfully (Boulay et al., 2018). Efforts to reduce class sizes and introduce teacher coaching programs statewide have not resulted in the large gains documented in the research literature, likely because of program modifications caused by budget restrictions and hiring difficulties (Jepsen & Rivkin, 2009; Lockwood, McCombs & Marsh, 2010).

The challenge posed by taking programs to scale is largely proportional to the degree of behavioral change required to implement a program. More technical and top-down interventions that require limited implementation by personnel are often easier to scale. Examples include

financial incentives for recruiting teachers in shortage areas and hard-to-staff schools, 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 adopting a new school-wide approach to behavioral supports for students, establishing professional learning communities among teachers, and implementing new curricula.

Political feasibility also plays an important role in determining scalability. For example, incentive schemes that leverage loss aversion — where teachers have to return bonuses they are paid in advance if students do not make achievement gains — have demonstrated impressive effects, but are unlikely to be implemented in practice (Fryer et al, 2012). Interventions often stall when they face opposition from organized constituencies. Efforts to use high-stakes teacher evaluation systems to remove low-performing teachers and pay large bonuses to highly-effective teachers have largely been undercut by strong political opposition (Kraft, 2018).

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 scale with fidelity and maintain their effectiveness 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, such as those above, can be applied widely and flexibly, but require considerable attention to detail and may result in subjective interpretations. Fixed benchmarks are easy to apply and provide unambiguous answers, but fail to account for important contextual differences across studies or to consider program costs and scalability. Like Cohen, I recognize the "risk inherent in offering conventional [benchmarks]," but agree there is "more to be gained than lost" (1998, p.25). When faced with complex information and competing demands on our time, we rely on heuristic shortcuts for interpreting effect sizes. The persistent application of Cohen's benchmarks, despite repeated critiques, suggests that little short of a simple alternative heuristic can dislodge these unrealistic expectations.

The schema I propose provides new benchmarks for causal studies that assess program effects measured by student achievement on broad standardized tests, while also considering program costs and scalability. The effect-size benchmarks are most appropriate for studies of upper elementary, middle, and high school students for whom annual learning gains are at least broadly comparable across grades. Effect-size benchmarks for achievement among younger students should likely be adjusted upwards to reflect the substantially larger annual learning gains children make early in their development (Bloom et al, 2008).

The motivation for a focus on causal studies with achievement outcomes is threefold. First, the focus serves to narrow contextual differences that make benchmarks impractical when considering a more diverse body of research. Second, effects on standardized achievement tests

have become a common standard by which education interventions are judged. These outcomes are collected annually for 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 achievement outcomes on which to base new benchmarks.

A New Approach

There is a growing consensus among researchers that effects that are small by Cohen's standards are often large and meaningful in the context of education interventions. Scholars have proposed that effect sizes of 0.20 or 0.25 SD should be considered "of policy interest" (Hedges & Hedberg, 2007, p.77), "substantively important" (WWC, 2014, p.23) or to have "educational significance" (Bloom et al., 2008, p.295). Lipsey and his colleagues assert unequivocally that effect sizes of 0.25 SD in education research should be considered "large" (Lipsey et al., 2012, p.4). More recent meta-analyses of field experiments in education demonstrate that it is particularly challenging to raise academic achievement. Cheung and Slavin (2016) find average effect sizes on academic achievement of 0.16 SD among 197 RCTs, while Fryer (2017) finds average effect sizes of 0.05 SD in math and 0.07 SD in reading based on 105 school-based RCTs.

I propose the following effect-sizes benchmarks for causal studies evaluating effects on student achievement among upper elementary, middle and high school students: less than 0.05 is Small, 0.05 to less than 0.20 is Medium, and 0.20 or greater is Large. To be clear, these are subjective but not arbitrary benchmarks. They are easy heuristics to remember that reflect the findings of recent meta-analyses. To illustrate this, I describe the distribution 481 effect sizes from 242 RCTs of education interventions with achievement outcomes in Table 1. The proposed

benchmarks represent reasonably approximate middle cutpoints in the distribution (45th and 76th percentiles) and are in proportion with other established empirical reference points such as the size of annual student learning gains across grade levels and the magnitude of teacher and school effects.

Table 1. Empirical Distributions of Effect Sizes and Costs from Education Interventions

Percentile	ES	Per-Pupil Cost
1st	-0.28	$18
10th	-0.05	$77
20th	-0.01	$121
30th	0.01	$210
40th	0.04	$301
50th	0.07	$882
60th	0.11	$1,468
70th	0.16	$3,150
80th	0.24	$7,259
90th	0.40	$15,530
99th	0.90	$61,248
N	481 (242 studies)	68

Notes: ES = Effect Size.
Source: Effect sizes are based on gains in test scores from Boulay et al. (2018) and Fryer (2017). 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, and Elson (2007), Levin et al. (2012), and Hollands et al. (2016).

If calling an effect size of 0.20 SD large seems overly enthusiastic, consider this: studies show that raising student achievement by 0.20 SD results in a 2 percent increase in annual lifetime earnings on average (Chetty, Friedman, & Rockoff, 2014) and is equivalent to approximately one fourth of the Black-White achievement gap (Bloom et al., 2008). Others might object to characterizing a 0.05 SD effect as moderate, but raising academic achievement is difficult. Over 22% of the effect sizes in the distribution shown in Table 1 are 0 SD or smaller, with many more failing to obtain traditional levels of statistical significance. 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 a fraction of these achievement gains (Chingos, Whitehurst, & Gallaher, 2015). As Prentice and Miller (1992) argued over 25 years ago, small effect sizes on broad outcomes that are difficult to move should be considered impressive, especially when they are the result of relatively minimal interventions.

However, 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 per-pupil cost benchmarks where less than $500 is Low, $500 to under $4,000 is Moderate, and $4,000 or greater is High.[7] Similar to the effect-size benchmarks, these subjective conventions provide simple heuristics that are directly informed by an empirical distribution of the per-pupil costs of 68 education interventions as well as other empirical reference points such as the average per-pupil expenditures in U.S. public schools (~$12,000). They also represent similar, reasonably approximate middle cutpoints in the cost per-pupil distribution shown in Table 1 (43rd and the 74th percentiles).

Table 2. A Schema for Interpreting Effect Sizes from Causal Studies with Achievement Outcomes

		Cost-Effectiveness Ratio (ES/Cost)				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

[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.

The combination of the proposed effect-size and cost benchmarks results in a 3x3 matrix where each cell classifies an effect size as a simple cost-effectiveness ratio. The matrix helps to clarify two key insights about interpreting effect sizes by illustrating how effect sizes and costs interact. Large effect sizes 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 different combinations of effect sizes and costs that share the same color on a common downward slopping diagonal.

The third and final step is assessing whether an intervention is easy, reasonable, or hard to scale. Here there are no clear benchmarks to apply. Instead, this step requires the subjective judgement of the interpreter following the guidance I provide above. Reasonable people will disagree about program scalability. The larger point is to interject 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 can be fully considered and better addressed.

An Example

Consider, for example, the effects of individualized tutorials (0.23 SD) and universal free breakfast (0.09 SD). Cook et al. (2015) report that the annual cost of individualized tutorials is more than $2,500 per student. 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). Simply considering costs illustrates how the smaller effect size of universal free breakfast is, from a policy standpoint, equally if perhaps not more impressive than the large effect of individualized tutorials. Universal free breakfast produces a

medium effect size at a very low cost compared to individualized tutoring with a large effect size at a moderate cost.

Incorporating scalability serves to further illustrate how smaller effect sizes can be more meaningful than larger ones. Implementing individualized tutorials requires schools to reorganize their schedule to incorporate tutoring into the school day. Schools would need to recruit, select, train and support a corps of tutors, while also adapting an existing tutorial curriculum or developing their own. Much of the effect of tutoring depends on the quality of the tutors. I would characterize these implementation challenges as non-trivial, but reasonable, given they don't require major behavioral changes on the part of 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

Evidence from rigorous evaluations of education interventions is necessary for evidence-based policy and practice, but it is not sufficient. Scholars and research consumers need to be able to interpret this evidence and judge its importance. This article provides broad guidelines and a more detailed schema to aid in the process of evaluating findings reported as effect sizes. Although the proposed schema is intended for causal studies with achievement outcomes, it provides a blueprint that could be adapted easily for other types of studies and outcomes. Together, this framework illustrates why effect sizes judged to be small by Cohen's standards

deserve to be discussed, not dismissed. Education interventions in the field often have small effects or no effect at all. We need to update our expectations as well as look beyond impacts to consider program costs and scalability. Effect sizes that are equal in magnitude are rarely equal in importance.

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.

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. and Morales, M., 2015. A benefit-cost analysis of City Connects. *Center for Benefit-Cost Studies in Education: Teachers College, Columbia University*.

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. (1977). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, N.J.: Lawrence Erlbaum.

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*

Doss, C., Fahle, E. M., Loeb, S., & York, B. N. (2018). More than just a nudge: Supporting kindergarten parents with differentiated and personalized text-messages. *Journal of Human Resources*, 0317-8637R.

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*.

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.

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., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87.

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. and 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.

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

Jacob, R., Armstrong, C., Bowden, A.B. and 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.

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.

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. and Hanisch-Cerda, B. (2012). Cost–effectiveness analysis of interventions that improve high school completion. *Teacher College, Columbia University*.

Levin, H.M., Catlin, D. and 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*.

Lockwood, J. R., McCombs, J. S., & Marsh, J. (2010). Linking reading coaches and student achievement: Evidence from Florida middle schools. *Educational Evaluation and Policy Analysis*, *32*(3), 372-388.

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.

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, *3*(4), 2332858417749220.

Patall, E. A., Cooper, H., & Allen, A. B. (2010). Extending the school day or school year: A systematic review of research (1985–2009). *Review of Educational Research*, 80(3), 401-436.

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.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological*

Bulletin, *112*(1), 160.

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*.

Rogers, T., & Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, *2*(5), 335.

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.

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

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.

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

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

What Works Clearinghouse. (2014). WWC procedures and standards handbook (Version 3.0). *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.*

York, B. N., Loeb, S., & Doss, C. (2018). One step at a time: The effects of an early literacy text messaging program for parents of preschoolers. *Journal of Human Resources*, 0517-8756R.