The Effect-Size Benchmark That Matters Most: Education Interventions Often Fail

Matthew A. Kraft[1]

It is a healthy exercise to debate the merits of using effect-size benchmarks to interpret research findings. However, these debates obscure a more central insight that emerges from empirical distributions of effect-size estimates in the literature: Efforts to improve education often fail to move the needle. I find that 36% of effect sizes from randomized control trials of education interventions with standardized achievement outcomes are less than 0.05 *SD*. Publication bias surely masks many more failed efforts from our view. Recognizing the frequency of these failures should be at the core of any approach to interpreting the policy relevance of effect sizes. We can aim high without dismissing as trivial those effects sizes that represent more incremental improvement.

Keywords: achievement; educational policy; effect size; research utilization

We need to reorient how we view success in the education sector. As researchers, policymakers, and practitioners, we have high hopes that new reforms will unlock students' full potential and prepare them to thrive as adults. We are naturally disappointed when policies or programs fail or fall short of these aspirational goals. But improving education is difficult work, particularly in high-income countries with well-established K–12 systems such as the United States. Schooling is a complex endeavor where societal inequities, political systems, institutional structures, and individual actors exert strong forces over how educational initiatives are design and implemented. We are, as Tyack and Cuban (1995) famously wrote, "Tinkering Toward Utopia."

In this article, I replicate the analyses in Kraft (2020) using an expanded data set of over 3,000 effect sizes and respond to critiques raised in Simpson (2021). I argue that debates about what constitutes a small, medium, or large effect size in education are not sufficiently anchored to the fact that education interventions often fail to move the needle on standardized tests of student achievement. In both the original analyses and the expanded data set, I find that 36% of effect sizes from randomized control trials (RCTs) of education interventions with standardized achievement outcomes are smaller than 0.05. Accounting for publication bias would certainly raise this percentage further. This is the benchmark that matters most. Failing to recognize the frequency of failure causes us to hold outsized, unrealistic norms for what counts as a policy relevant effect.

Effect Sizes in an Expanded Sample

In "Interpreting Effect Sizes of Education Interventions" (IESEI; Kraft, 2020), I synthesized the literature into five broad guidelines for interpreting the policy relevance of effect sizes. I then attempt to uproot the long-standing use of Cohen's benchmarks in education research by proposing alternative benchmarks for effect sizes from RCTs of education interventions with standardized test outcomes. Here, I start by exploring whether we might further refine the empirical benchmarks I proposed as more data have become available. I replicate the main analyses using an expanded sample of 3,426 effect sizes (for details, see the Data Appendix available on the journal website).

The addition of over 75% more effect-size estimates to the data set does little to shift the overall empirical distribution or patterns revealed in Kraft (2020). For example, percentiles of the overall distribution remain remarkably stable (30th percentile in expanded dataset = 0.02 vs. 0.02 in original; 50th percentile = 0.10 vs. 0.10; 70th percentile = 0.21 vs. 0.21). The only notable change is a longer right-hand tail of the effect-size distribution in math that raises the median slightly (50th 0.11 vs. 0.07) but pushes out the 90th percentile substantially (0.78 vs. 0.37). This appears driven by the addition of small-sample studies aimed at improving early math skills in numeracy, arithmetic, and fractions that use more narrowly focused tests. Collapsing the data by calculating a simple average across effect-size estimates within

[1]Brown University, Providence, RI

Educational Researcher, Vol. 52 No. 3, pp. 183–187
DOI: 10.3102/0013189X231155154

each of the 973 studies results in a very similar distribution (30th percentile = 0.04, 50th percentile = 0.12, 70th percentile = 0.25).

The studies included in my original and expanded samples are almost entirely from high-income countries. Evans and Yuan (2022) provided complementary evidence on the distribution of effect sizes across 96 education impact evaluations examining learning outcomes in low- and middle-income countries. Strikingly, they also find a median effect size of 0.10 *SD* (30th percentile = 0.03, 70th percentile = 0.18). Overall, both my expanded sample and evidence from Evans and Yuan suggest the baseline benchmarks I proposed in IESEI (Kraft, 2020) remain reasonable. This, of course, might change over time as research progresses.

The Critique

Simpson's (2021) technical comment on IESEI (Kraft, 2020), titled "Benchmarking a Misnomer" (BAM), focused on my proposed effect-size benchmarks. The critique is twofold: Any attempt to use benchmarks is misguided given that effect sizes from different studies are essentially incomparable and that my proposed benchmarks are too low. BAM provides a welcome opportunity to engage in the spirit of scholarly debate on each of these thorny issues. Simpson was kind enough to share an early draft, and I reciprocated by sharing my data. We have since engaged in an amiable correspondence that has revealed much common ground. For example, we both share a healthy suspicion of "league tables" that rank different education interventions based on the strength of their associations with any manner of outcomes. Although his critiques have pushed my thinking in important ways, I remain convinced that the baseline benchmarks and interpretation process I proposed in IESEI are a productive approach for characterizing the magnitude of effect sizes from causal studies of education interventions with standardized achievement outcomes.

Are Effect-Size Benchmarks Ever Appropriate?

Theoretical critiques against the use of effect-size benchmarks are well established in the literature (e.g., Glass et al., 1981; Kelley & Preacher, 2012). The central argument is this: Research designs differ across a multitude of factors that render effect sizes across studies incomparable and benchmarks inappropriate. BAM's (Simpson, 2021) critique has merit—converting treatment estimates into effect sizes does not magically allow for a perfect apples-to-apples comparison. In IESEI (Kraft, 2020), I highlighted how a range of study features influence effect sizes including the research design; what, when, and how outcomes are measured; the sample; the approach to standardizing; the treatment-control contrast; and the type of treatment effect estimated. Where we disagree is what to do about this challenge.

Theoretical critiques such as BAM's (Simpson, 2021) have had little success dislodging the widespread use of Cohen's benchmarks despite Cohen himself explaining that his benchmarks are somewhat arbitrary:

> These proposed conventions were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible. The values chosen had no more reliable basis than my own intuition. (Cohen, 1988, p. 532)

An orthodoxy that says we can never compare effect sizes or, by extension, any empirical estimates unless every feature of a study is identical loses sight of the essential purpose of statistics. Mathematics has inviolable laws—the area of a circle is always equal to πr^2. Statistics embraces variability in the name of synthesizing large amounts of information, uncovering patterns, and making broader inferences. Excommunicating the effect-size statistic from the tool kit of social scientists because no study is a pure apples-to-apples comparison would weaken evidence-based policymaking and leave us with worse alternatives. Surely we can gain at least some insights by comparing the noisy signals of estimated program effects across different studies.

My proposed approach to interpreting effect sizes aims to find a constructive path forward by combining a "focus-narrowing" and "adjustment" approach (to borrow BAM's useful language; Simpson, 2021). The baseline benchmarks are specific to studies of education interventions with causal research designs that examine standardized achievement outcomes. I explain further how we should adjust these baseline benchmarks up or down based on specific study features with the intention of contextualizing estimates relative to their own study context. Benchmarks such as these are pragmatic and make the interpretation process accessible to a wide audience.

The alternative—relying on "professional judgment in context" (Simpson, 2019, p. 105)—is an abstract ideal that is fraught with even greater subjective biases and infeasible for many consumers of research. Relying on professional judgment alone, even among academics, can result in widely different normative interpretations of effects. This alternative also places the onus on policymakers and educators to analyze full academic texts that are often inaccessible because of arcane writing, technical complexity, and paywalls. Instead, we should use effect-size benchmarks (for specific bodies of literature with similar research designs and outcomes) as empirically informed starting points that can be modified based on professional judgment.

Are IESEI's Baseline Benchmarks Too Low?

BAM (Simpson, 2021) argues that the baseline benchmarks I propose are too low because of how I chose the cut points in the empirical distribution and because I did not take the absolute value of effect sizes or remove nonsignificant effect sizes from the sample. These technical critiques and our subsequent exchanges have helped to sharpen my thinking about the interplay between magnitude, sign, and significance when interpreting effect sizes. However, I remain convinced that considering where an effect size falls in the full distribution of unadjusted effect-size estimates, regardless of their precision, is a productive approach for evaluating the policy importance of a finding. BAM's (Simpson, 2021) technical adjustments would substantially distort this raw empirical distribution, obscuring the first-order insight that many education interventions fail. If anything, empirical benchmarks may set the bar too high because of publication bias and estimation error.

Selecting empirical cut points for benchmarks. To be clear, the benchmarks I propose in IESEI (Kraft, 2020) are subjective and conceptually focused on interpreting the policy relevance of positive effect sizes. They were informed by the overall empirical distribution of effect sizes I constructed and the heuristic value of landmark numbers and established empirical reference points (e.g., the size of annual learning gains across grade levels and the magnitude of teacher and school effects).[1] This is why I do not report or use the exact empirical values of the tercile cut points in IESEI (Kraft, 2020).

The alignment of the benchmarks with approximate terciles of the full empirical distribution is misleading because it appears that I chose the benchmarks strictly based on these cut points. This would make little sense, as BAM (Simpson, 2021) rightly points out, given that even extreme negative effect sizes would fall into the *small* category. My focus was on interpreting the policy relevance of positive effect sizes given that negative effect sizes are likely "off the table" from a policy perspective. I should have stated explicitly what I assumed would be implied—that the range of small (positive) effects has an implicit lower bound of 0.[2]

Taking the absolute value of effect sizes. Taking the absolute value of effect sizes would be appropriate if my primary purpose for proposing empirical benchmarks was to inform power analyses, as was Cohen's (1988). From the perspective of statistical power, the sign of an effect is inconsequential. However, the stated purpose of IESEI (Kraft, 2020) is to inform how we can judge the *policy importance* of an effect size. Here, the sign of an effect is critical.

BAM's (Simpson, 2021) argument for removing the signs of effect sizes even when constructing benchmarks for policy relevance stems from the idea that the direction of an effect is arbitrary. This is true for studies where there are two treatment arms and no traditional control group such as the example BAM (Simpson, 2021) points readers to, Agodini and Harris (2010).[3] But the idea that "randomized control trials of the type in the dataset are normally symmetrical with respect to treatment" (Simpson, 2021) is not an accurate characterization of the literature. Most modern RCTs in the field contrast outcomes for a treatment group to those from a control group that experiences "business as usual" even if the RCT includes multiple treatment arms. For example, I found that 94% of the treatment-control contrasts reported in studies reviewed by Fryer (2017) have a treatment group that did something new and a control group that engaged in standard practice prior to the intervention. The design of the Agodini and Harris (2010) is an exception. I agree with BAM (Simpson, 2021) that what "business as usual" actually means is highly variable across studies and can shape the magnitude of effect sizes as I describe in IESEI (Kraft, 2020). However, the direction of an effect in the typical RCT design is clear from a policy perspective: Did the intervention increase student outcomes relative to current practice?

Most fundamentally, taking the absolute value of effect-size estimates would prevent us from judging where in the full empirical distribution an effect-size estimate falls. Knowing the relative position of an effect-size estimate is precisely the simple eyeball test we need to see that even small positive effect-size estimates rank relatively high in the distribution of education interventions that have been rigorously evaluated.

Ignoring nonsignificant effect-size estimates. BAM's (Simpson, 2021) assertion that I should drop nonsignificant effect-size estimates is partly a product of the unfortunate semantics of the term "effect size." Effect sizes are simply point estimates that have been standardized to be on a common unit scale. However, "effect-size estimates" is commonly abbreviated as "effect sizes," which then become "effects" in shorthand, suggesting it also implies statistical significance. My proposed benchmarks are for characterizing the policy importance of these estimates and should be paired with information about statistical significance when interpreting their relevance. Both the magnitude and precision of effect-size estimates matter.

BAM (Simpson, 2021) argues one should benchmark effect-size estimates relative only to other interventions that have produced statistically significant effects. This approach is not wrong; it just answers a fundamentally different question from the one I am focused on. From a policy standpoint, I view it as more instructive to set benchmarks relative to effect-size estimates from all interventions studied in the literature. We can't know ex ante if an intervention will produce a significant effect, so why make our sample for estimating empirical benchmarks conditional on results that are only known ex post? There are likely many education interventions that have effects near zero. It is difficult and costly to conduct studies that are large enough to provide the statistical power needed to distinguish these small effects from zero. Excluding nonsignificant estimates when constructing a counterfactual distribution of effect sizes would disproportionally remove smaller, less precisely estimated effect sizes, exaggerating the dispersion of the distribution and our expectations.

Publication Bias and Estimation Error

Two primary concerns make me think that, if anything, the baseline benchmarks I proposed could be too high. A large literature documents the prevalence of publication bias in academic research, with some arguing this phenomenon is particularly acute in the social sciences (Rothstein et al., 2005). Here, I use the term "publication bias" to encompass the pattern where the direction, size, and statistical significance of research findings influence whether a study is published in a mainstream academic journal. These biases arise from a number of factors, including (a) scholars selectively choosing not to write or submit manuscripts based on negative, small, or imprecise effects; (b) scholars selectively choosing not to report individual effects that are negative, small, or imprecise; (c) academic journals selectively choosing not to publish studies with negative, small, or imprecise effects; and (d) studies with negative, small, or imprecise effects being published in less well-known journals that may be harder to find in a review (Chan et al., 2004; Dickerson, 2005).

Analyses of publication bias in the experimental economics and psychology literatures suggest that significant findings are more likely to be published than insignificant results by multiple orders of magnitude (Andrews & Kasy, 2019). Although precisely quantifying the nature and extent of publication bias is challenging, the data described in Table 1 exhibit two patterns suggestive of this bias (Rothstein et al., 2005). First, effects from smaller studies are, on average, more than 10 times larger than effects from larger studies. This is consistent with a pattern where

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		
	Multiple ESs per study	Single average ES per study	Math	Reading	≤ 100	101 to 250	251 to 500	501 to 2,000	> 2,000	Broad	Narrow	DoE studies
Mean	0.18	0.20	0.24	0.16	0.40	0.25	0.18	0.10	0.04	0.17	0.27	0.04
Standard deviation	0.33	0.29	0.44	0.27	0.44	0.45	0.28	0.16	0.13	0.32	0.38	0.16
Mean (weighted)	0.05	0.06	0.06	0.06	0.40	0.24	0.18	0.09	0.03	0.05	0.15	0.02
P1	–0.33	–0.29	–0.67	–0.30	–0.44	–0.71	–0.25	–0.21	–0.23	–0.37	–0.20	–0.38
P10	–0.07	–0.04	–0.08	–0.07	–0.08	–0.13	–0.06	–0.06	–0.06	–0.08	–0.04	–0.12
P20	–0.01	0.01	–0.01	–0.01	0.05	–0.03	0.01	–0.01	–0.03	–0.01	0.02	–0.06
P30	0.02	0.04	0.03	0.03	0.14	0.04	0.06	0.02	–0.01	0.02	0.06	–0.04
P40	0.06	0.08	0.07	0.06	0.25	0.11	0.10	0.05	0.01	0.06	0.10	0.01
P50	0.10	0.12	0.11	0.10	0.36	0.17	0.14	0.07	0.03	0.10	0.16	0.03
P60	0.15	0.17	0.18	0.15	0.48	0.24	0.18	0.10	0.05	0.14	0.24	0.05
P70	0.21	0.25	0.29	0.20	0.57	0.35	0.22	0.13	0.07	0.20	0.32	0.08
P80	0.33	0.35	0.49	0.30	0.70	0.47	0.29	0.18	0.10	0.31	0.48	0.14
P90	0.57	0.55	0.78	0.49	0.91	0.76	0.43	0.27	0.15	0.55	0.74	0.19
P99	1.37	1.27	1.98	1.07	1.65	1.98	1.09	0.77	0.55	1.37	2.09	0.66
k (number of effect sizes)	3,426	973	1,011	2,178	504	759	593	896	480	3,057	369	194
n (number of studies)	973	973	396	659	194	194	203	286	208	954	91	56

Note. The vast majority of the standardized achievement outcomes (93%) are based on math and English language arts test scores, with the remaining based on science, social studies, or general achievement. Weights are based on sample size for weighted mean estimates. For details about data sources, see Appendix available on the journal website. DoE = U.S. Department of Education; ES = effect size. P1, P10, P20 . . . = percentile values in the overall distribution.

underpowered studies that find small estimated effects are less likely to be published because they lack the sample size necessary to distinguish small effects from zero. Second, I find much smaller effects among a subsample of studies commissioned by the U.S. Department of Education that required scholars to submit reports with their findings (50th percentile = 0.03 for Department of Education studies vs. 50th 0.10 for full sample; see also Lortie-Forgues & Inglis, 2019). Scholars have found this pattern of smaller average effects among samples of preregistered studies and funded trials relative to published academic articles across multiple disciplines (DellaVigna & Linos, 2022; Kaplan & Irvin, 2015). These patterns are not proof positive of publication bias and could instead reflect systematic differences in studies that are correlated with sample size and the nature of larger-scale efficacy trials typically funded by governments. It would seem unlikely, though, that publication bias did not skew the distribution of published effect sizes toward larger positive estimates at least to some degree.

A second concern is that estimation error inherent in the research process causes distributions of effect-size estimates to appear wider than the underlying distribution of true effects. While individual effect sizes from causal studies can be considered unbiased, the joint distribution of these individual estimates reflects both true variation and error variance from random draws of positive or negative error terms (Morris, 1983; for applied examples, see Jackson & Mackevicius, 2021; von Hippple & Bellows, 2018).[4] This estimation error inflates the variance of an observed empirical distribution, giving the appearance that empirical benchmarks should be farther away from the center of the distribution. Future approaches to updating these and other benchmarks might benefit from applying Bayesian shrinkage methods to estimate the degree to which these empirical distributions overestimate true underlying variation in effect sizes.[5]

The Path Forward

I continue to see the empirical benchmarks I propose in IESEI (Kraft, 2020) as a productive baseline for interpreting effect sizes from causal research on education interventions with standardized achievement outcomes. Debating their merit is worthwhile because they are empirically informed but ultimately subjective and focus on interpreting the policy relevance of positive effect sizes. I encourage scholars to update my proposed effect-size benchmarks as science advances, but not by narrowly using terciles of the full empirical distribution. Instead, I would continue to anchor on key empirical reference points in the literature and consider whether the relative rank of my proposed benchmarks shift upward (implying they may be too large) or downward (implying they may be too small) in the full distribution as new estimates are added.

At the same time, the debate about what is a small, medium, and large effect obscures a more important insight that should be

our North Star: Efforts to improve education often fail to improve student outcomes, particularly student achievement. Holding outsized expectations keeps us focused exclusively on the next silver bullet. We can aim high without dismissing the unglamorous but essential work of incremental improvements. This North Star, alongside the costs and scalability of an intervention, should be at the core of any approach to interpreting the policy relevance of effect sizes.

ORCID ID

Matthew A. Kraft https://orcid.org/0000-0002-3889-8413

NOTES

I am deeply grateful for the support that Joshua Bleiberg, Alex Bolves, Alvin Christian, Virginia Lovison, and Melissa Lyon provided in compiling the expanded effect-size data set. All errors and omissions are my own.

[1]This language comes directly from the early working paper version of IESEI (Kraft, 2020) available at https://edworkingpapers.com/sites/default/files/kraft_2018_interpreting_effect_sizes%20%282%29.pdf.

[2]BAM's (Simpson, 2021) critique does raise the question of whether the effect-size benchmarks I propose should be applied symmetrically to negative effect sizes. I see two possible approaches. One could adopt different cut points for negative effect sizes to reflect the fact that empirical effect-size distributions are rarely symmetric or centered on the origin. One could also use negative cut points that are symmetric with the ones I propose, which is appealing for its simplicity (i.e., –0.20 and –0.05 for the negative effect sizes and 0.05 and 0.20 for positive effect sizes).

[3]I have removed this study from the expanded sample of effect sizes used in this article.

[4]Formally, effect estimates can be written as $\hat{\beta}_j = \beta_j + \varepsilon_j$, where β_j is the true causal effect from study j and ε_j is independent mean-zero estimation error. Thus, $E[\hat{\beta}_j] = \beta_j$ but $Var(\beta_j) = \theta^2_{\beta_j} + \frac{1}{J} * \sum_j \theta^2_{\varepsilon_j} > \theta^2_{\beta}$, where θ^2_{β} is the true effect variance and $\theta^2_{\varepsilon_j}$ is the variance of ε_j.

[5]This is not possible in my original or expanded effect size samples given that not all the sources we used to collect effect sizes reported their associated standard errors or specific p-values.

REFERENCES

Agodini, R., & Harris, B. (2010). An experimental evaluation of four elementary school math curricula. *Journal of Research on Educational Effectiveness, 3*(3), 199–253.

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–2794.

Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, *291*(20), 2457–2465.

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

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116.

Dickerson, K. (2005). Publication bias: Recognizing the problem, understandings its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment, and adjustments* (pp. 11–34). Wiley

Evans, D. K., & Yuan, F. (2022). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis, 44*(3), 532–540.

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

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

Jackson, C. K., & Mackevicius, C. (2021). *The distribution of school spending impacts* (No. w28517). National Bureau of Economic Research.

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PloS One*, *10*(8), Article e0132382. https://doi.org/10.1371/journal.pone.0132382

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

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

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

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*(381), 47–55.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.

Simpson, A. (2019). Separating arguments from conclusions: The mistaken role of effect size in educational policy research. *Educational Research and Evaluation*, *25*(1–2), 99–109.

Simpson, A. (2021). Benchmarking a misnomer: A note on "Interpreting effect sizes in education interventions." *Educational Researcher*. Advance online publication. https://doi.org/10.3102/0013189X20985448

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia*. Harvard University Press.

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review*, *64*, 298–312.

AUTHOR

MATTHEW A. KRAFT, EdD, is an associate professor of education and economics at Brown University, PO Box 1938 Providence RI, 02912; *mkraft@brown.edu.* His research focuses on efforts to improve educator and organizational effectiveness in K–12 public schools.

Manuscript received April 6, 2021
Revisions received May 12, 2022, and December 14, 2022
Accepted December 21, 2022