How Big Are Effect Sizes in International Education Studies?

David K. Evans and Fei Yuan

Abstract

In recent years, a growing literature has measured the impact of education interventions in low- and middle-income countries on both access and learning outcomes. But interpretation of those effect sizes as large or small tends to rely on benchmarks developed by a psychologist in the United States in the 1960s. In this paper, we demonstrate the distribution of standardized effect sizes on learning and access from hundreds of studies from low- and middle-income countries. We identify a median effect size of 0.10 standard deviations on learning and 0.06 standard deviations on access. Effect sizes are similar for randomized controlled trials and for quasi-experimental studies. They are much larger for small-scale studies than for large-scale studies. Understanding the distribution of existing effects can help researchers and policymakers to situate new findings within the distribution of current knowledge.

Keywords: education; development; standard deviations; measurement

JEL: I21; I25; O15

Center for Global Development
www.cgdev.org

Working Paper 545
August 2020

How Big Are Effect Sizes in International Education Studies?

David K. Evans
Center for Global Development
devans@cgdev.org

Fei Yuan
Harvard Graduate School of Education
fyuan@g.harvard.edu

Author order is alphabetical. The authors thank Lee Crawfurd, Susannah Hares, Matthew Kraft, Justin Sandefur, Eric Taylor, and Eva Vivalt for helpful comments and Amina Mendez Acosta for helpful research assistance.

The Center for Global Development is grateful for contributions from the Bill & Melinda Gates Foundation in support of this work.

David K. Evans and Fei Yuan, 2020. "How Big Are Effect Sizes in International Education Studies?" CGD Working Paper 545. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/how-big-are-effect-sizes-international-education-studies

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

The Center for Global Development works to reduce global poverty and improve lives through innovative economic research that drives better policy and practice by the world's top decision makers. Use and dissemination of this Working Paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License.

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors, funders of the Center for Global Development, or the authors' respective organizations.

Introduction

What are the best strategies to improve access to school and improve learning in school in low- and middle-income countries? Recent years have seen a rapid increase in rigorous evaluations of interventions intended to expand access and improve learning outcomes (World Bank, 2018). But researchers often interpret the size of these impacts based on benchmarks proposed by a U.S.-based psychologist more than fifty years ago (Cohen, 1969).

In this paper, we present the distribution of effect sizes from hundreds of studies that evaluate the impact of educational interventions on students' access to schooling and 225 studies that evaluate the impact on students' learning in school. This builds on Kraft (2020), which carries out a similar exercise for studies with learning outcomes in the high-income countries. We further provide the distribution of effect sizes across different study designs (randomized controlled trials versus quasi-experimental studies), the scale of the study, and the specific enrollment and learning outcome (e.g., school attendance versus school dropout, or math versus reading). This distribution of effects provides researchers with a simple way to situate impacts of new studies relative to what is known about how to expand access and increase learning. It also anchors expectations about the impact of future interventions. The current distribution of effect sizes does not rule out dramatically larger effect sizes of innovative education interventions in the future, but it does help researchers and policy makers to understand what a dramatically larger effect size would be.

Specifically, we draw on a large database of 156 randomized controlled trials (RCTs) and 143 quasi-experimental studies with learning or access outcomes, drawn from previous synthesis reviews of education in low- and middle-income countries. We then standardize effect sizes across studies. We find that across 130 RCTs that report learning outcomes, the median effect size is 0.10 standard deviations (SDs), and the 90th percentile is 0.38 SDs. In his work in high-income countries, Kraft (2020) draws on 747 randomized controlled trials of education interventions with test score outcomes and identifies a median impact of 0.10 SDs across education interventions. Smaller studies (with fewer than 500 students) report point estimates that are on average twice the size of larger studies (with more than 5,000 students): 0.10 SDs versus 0.05 SDs. For access outcomes, the median effect size is 0.07 SDs across 74 RCTs (and the effect is similar for 150 total studies). The 90th percentile is 0.30 SDs. The distribution of access effect sizes is similar whether the outcome is enrollment or attendance. Our sample of quasi-experimental studies show a similar pattern of results.

This distribution of effects contrasts with the benchmarks proposed by Cohen (1969), that a small effect size is 0.2 SDs, a medium effect size is 0.5 SDs, and a large effect size is 0.8 SDs. Those effect sizes were developed based on a small sample of social psychology lab experiments in the United States in the 1960s, mostly with undergraduate students (Kraft, 2020), so their relevance to education impact evaluations in basic education today in either the high-income countries or – much less – in low- and middle-income countries is questionable.

This study contributes to the growing body of synthesis work on the impact of education interventions in low- and middle-income countries (Evans & Popova, 2016; Kremer et al., 2013; Snilstveit et al., 2016), as well as to work seeking to characterize how large the impact of education interventions are relative to benchmarks such as the amount of learning usually gained during a year of schooling (Evans & Yuan, 2019a) or the difference in learning levels between rich and poor countries (Angrist et al., 2020; Filmer et al., 2020).

Data and analysis

We use a database of education impact evaluation studies to collect effect sizes. An earlier database compiled by Evans and Popova (2016) gathered all studies included in 10 systematic reviews of evidence on how to improve learning and access to education in low- and middle-income countries. That database was updated for Evans and Yuan (2019b) through a Google Scholar search in 2017-2018 as well as a review of working papers published on the websites of organizations that regularly carry out research in low- and middle-income countries (e.g., the World Bank, the International Initiative for Impact Evaluation, and the Jameel Abdul Latif Poverty Action Lab, among others). The updated database includes an initial sample of 518 studies. Using this database, we restrict our analytical sample to include effect sizes from studies that evaluate (1) direct education interventions (such as teacher professional development and providing learning materials), as well as two other classes of interventions that commonly report educational outcomes: (2) health interventions (such as providing deworming drugs and micronutrients) and (3) safety net interventions (such as cash transfers). We only include studies that use an experimental design (RCT) or a quasi-experimental design (specifically, difference-in-differences, regression discontinuity, instrumental variable, or propensity score matching) in the evaluation of effects. We excluded studies which reported effects only for boys or for girls. These restrictions yield a sample of 336 studies.

Not all studies reported effect sizes. For studies that only reported point estimates, we convert them into standardized effect sizes or Cohen's *d*, following Borenstein et al.(2009).

$$d = \frac{D}{S_{pooled}}$$

where D is the raw mean difference between a treatment group and a control group at follow-up, and S_{pooled} is the pooled standard deviation for the treatment and control groups combined. Studies without sufficient data for us to calculate the standard effect sizes were also excluded, bringing us to a final sample of 299 studies.

Besides standardizing effect sizes, we extracted data for a range of characteristics including the country and region where the intervention took place, the evaluation method, the sample size, and a list of measured outcomes.[1] Table 1 presents the summary statistics of our

[1] We follow the World Bank (2020) classification of region groups.

analytical sample. Our final analytical sample consists of 299 studies with 1,266 effect sizes from 52 low- and middle-income countries. By evaluation method, our sample includes 827 effect sizes from 156 RCT studies and 439 effect sizes from 143 quasi-experimental studies. We also categorized measured outcomes into two broad types: learning (which includes test scores of any subject, composite test scores, or passing a test) and access (which includes enrollment, attendance, dropout and years of schooling). Two thirds of effect sizes measure the impact on a learning outcome. Sub-Saharan Africa has the largest number of effect sizes (441), while Europe and Central Africa (7) and Middle East and North Africa (10) have the fewest.

Results

We present the results from 156 randomized controlled trials as our primary sample because of the potentially higher precision of those estimates. Across the RCTs that measure learning outcomes, we find a median impact of 0.10 SDs (Table 2 Panel A).[2] The median impact is smaller for math assessments (0.07 SDs) than for reading assessments (0.13 SDs). Studies that report a composite test score (or that do not report the subject) fall in between. For small studies with under 500 participating students, median impacts are 0.10 SDs. For large studies, with more than 5,000 students, the median impact is half that, at 0.05 SDs. This confirms the widespread belief that implementing effective programs at a pilot scale is easier than doing so at a national scale. The distribution of impacts is comparable (if slightly smaller) for quasi-experimental studies (Appendix Table A1), including the effect of larger impacts for the smallest scale studies and smaller impacts for the largest scale studies.

For RCTs that report impacts on access, the median impact is smaller: 0.07 SDs (Table 2 Panel B). Studies commonly report one or more of three access outcomes: enrollment (0.06 SDs), attendance (0.08 SDs), and dropout (0.05 SDs). The differences across outcomes are modest, but the slightly higher median for attendance may reflect greater ease in boosting student participation at the intensive margin than the extensive margin. The gap between small-scale and large-scale studies is even more striking with access outcomes: for studies with fewer than 500 participants, the median impact is 0.12 SDs, whereas for the largest studies (more than 5,000 students), the median impact is just 0.03 SDs. Quasi-experimental studies again show slightly smaller effects and a similar pattern vis-a-vis the scale of the program (Appendix Table A1).

Standard deviations are not always comparable across studies. As Singh (2015a, 2015b) shows, standard deviations will vary both across populations and across classes of tests. Thus, comparing SDs across contexts and tests should be treated with caution. In this study, we divide tests by subject to increase comparability. We also examine the relationship between scale of intervention and effect size in the three countries in our sample with the most estimates from randomized controlled trials, in order to enhance comparability across populations: Kenya (172 estimates), India (147 estimates), and China (104 estimates). For

[2] The mean is 0.14, which is close to the mean impact across a collection of international development impact evaluations (including but not limited to education) of 0.12, as reported in Vivalt (forthcoming).

access estimates, we observe a negative relationship for all three countries; it is statistically significant in Kenya and China (Figure 1 Panel A). For learning estimates, we observe a negative, statistically significant relationship for Kenya and China, with no clear correlation for India (Figure 1 Panel B).

Finally, we examine whether there are apparent differences in the distribution of effect sizes across regions (Appendix Table A2). We find, in the four regions with a reasonable sample of studies, both the most RCTs and largest median impacts on learning in Sub-Saharan Africa (0.13 SDs). Other regions have smaller medians: 0.08 SDs in South Asia, 0.09 SDs in Latin America and the Caribbean, and 0.08 SDs in East Asia and the Pacific. Europe and Central Asia and the Middle East and North Africa have too few studies to provide reliable estimates.

Discussion

In this short paper, we provide a distribution of studies that measure the impact of education interventions on learning and access in low- and middle-income countries. These data can help to situate future studies among the distribution of existing work. This is not a normative distribution. There is a large gap between student access and student learning in low-income countries versus high-income countries (Filmer et al., 2020), and one can reasonably argue that closing that gap will require either much larger effect sizes or a great many reforms that deliver effect sizes of the type that we observe. But merely calling for "transformative" education interventions will not by itself deliver student access and learning gains that are dramatically outside of the distribution we observe.

Our finding that – among RCTs – effect sizes are double in the smallest studies relative to the larger studies for learning and quadruple for access also encourage caution when policy makers encounter pilot results with impressive effect sizes. It is possible to improve both access and learning at scale, but usually the improvements are smaller than those observed in pilots.[3]

Our analysis yields recommendations for researchers. First, because SDs are not always comparable across studies, benchmarking effect sizes against real world metrics can enhance interpretation. For example, a recent study of a public-private partnership for primary education in Liberia yielded an effect size of 0.16 SDs after three years, which the authors emphasize is "equivalent to 4 words per minute additional reading fluency" (Romero & Sandefur, 2019). Having an older sister boosts Kenyan children's language and motor development by 0.1 SDs, which the authors highlight as the same as the difference between "children of primary-educated and secondary-educated mothers" (Jakiela et al., 2020). Second, many tests used in development studies are designed by the research team instead of

[3] Alternatively, we cannot rule out that this effect could be driven by selective publication bias, that success in publishing the results of small-scale evaluations is more sensitive to effect size than publishing larger-scale evaluations.

using standardized tests, and little is reported about what exactly is measured. Developing comparable measures across studies will require greater reporting precision by researchers.

One limitation of this work is that we do not incorporate cost effectiveness. The interpretation of an effect size is mediated by the cost of the intervention. Unfortunately, few studies report cost effectiveness into their analysis. An analysis of 76 RCTs in low- and middle-income countries found that nearly half reported no details on costs, and most of the others had minimal information (McEwan, 2015). A more recent analysis of recent education research from Africa found that less than a third of studies comprehensively reported costs (Evans & Mendez Acosta, 2020). As more studies report cost data in comparable ways, it will be possible to supplement this analysis with costs.

Because the distribution we demonstrate is so far removed from the benchmarks proposed by Cohen (1969), many readers may be tempted to despair that impacts of interventions tested thus far have been so small. First, comparing the literacy skills of adults with different years of schooling in five low- and middle-income countries (Bolivia, Colombia, Ghana, Kenya, and Vietnam), we infer that students gain between 0.15 and 0.21 SDs of literacy during the course of a school year (Evans & Yuan, 2019a). Thus, a large effect on learning according to our benchmark (larger than 0.16 SDs) is the equivalent to how much a student might learn in a full year of business-as-usual schooling. Second, average effects can mask large effects for subsets of students. An average effect of 0.16 SDs (again, a large effect size) may mean that an intervention had no impact for three-quarters of the students but an impact of 0.64 SDs for a subset of students, which would be considered a success by virtually any metric (Gelman, 2020). Many programs aim to improve outcomes for the students struggling the most, so such an outcome may represent a program success. Third, there are programs that do have exceptionally large program impacts. For example, a literacy program that provides intensive training and materials to teachers in Uganda to help them teacher literacy in children's mother tongue increased reading scores by 0.64 SDs and writing scores by 0.45 SDs (Kerwin & Thornton, forthcoming). High-intensity learning camps boosted language scores by 0.70 SDs in one state in India (Banerjee et al., 2016). Scholarships to secondary school in Ghana boosted enrollment by 0.56 SDs (Duflo et al., 2019). Providing take-home meals to students in Uganda, conditional on their attendance, boosted enrollment by 0.42 standard deviations (Alderman et al., 2012). These are all above the 90th percentile of the distribution of effect sizes. Interventions can and should still aim to achieve large changes in learning and access, but it is important to understand the full distribution.

These results suggest that small changes to existing interventions are unlikely to produce radical changes in closing access gaps or boosting learning. The interventions in our sample include everything from traditional interventions (like providing school inputs and training teachers) to twenty-first century interventions (using education technology). Closing education gaps between high- and low-income countries will require both an array of interventions and creative thinking to relax constraints more effectively than in the past.

References

Alderman, H., Gilligan, D. O., & Lehrer, K. (2012). The Impact of Food for Education Programs on School Participation in Northern Uganda. *Economic Development and Cultural Change*, 61(1), 187–218. https://doi.org/10.1086/666949

Angrist, N., Evans, D. K., Filmer, D. P., Glennerster, R., Rogers, H., & Sabarwal, S. (2020). *A New Micro Measure for Education Interventions: Learning-Adjusted Years of Schooling (LAYS)* [Working Paper].

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M., & Walton, M. (2016). *Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India* (Working Paper No. 22746). National Bureau of Economic Research. https://doi.org/10.3386/w22746

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470743386

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

Duflo, E., Dupas, P., & Kremer, M. (2019). *The Impact of Free Secondary Education: Experimental Evidence from Ghana*. Massachusetts Institute of Technology Working Paper Cambridge. https://web.stanford.edu/~pdupas/DDK_GhanaScholarships.pdf

Evans, D., & Mendez Acosta, A. (2020). Education in Africa: What Are We Learning? Center for Global Development working paper 542.

Evans, D., & Popova, A. (2016). What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews. *The World Bank Research Observer*, *31*(2), 242–270. https://doi.org/10.1093/wbro/lkw004

Evans, D., & Yuan, F. (2019a). *Equivalent Years of Schooling : A Metric to Communicate Learning Gains in Concrete Terms* (No. WPS8752; pp. 1–52). The World Bank. http://documents.worldbank.org/curated/en/123371550594320297/Equivalent-Years-of-Schooling-A-Metric-to-Communicate-Learning-Gains-in-Concrete-Terms

Evans, D., & Yuan, F. (2019b). *What We Learn about Girls' Education from Interventions that Do Not Focus on Girls* (No. WPS8944; pp. 1–45). The World Bank. http://documents.worldbank.org/curated/en/243741563805734157/What-We-Learn-about-Girls-Education-from-Interventions-that-Do-Not-Focus-on-Girls

Filmer, D., Rogers, H., Angrist, N., & Sabarwal, S. (2020). Learning-adjusted years of schooling (LAYS): Defining a new macro measure of education. *Economics of Education Review*, 101971. https://doi.org/10.1016/j.econedurev.2020.101971

Gelman, A. (2020). *Understanding the "average treatment effect" number*. https://statmodeling.stat.columbia.edu/2020/06/30/understanding-the-average-treatment-effect-number/

Jakiela, P., Ozier, O., Fernald, L. C. H., & Knauer, H. A. (2020). Big Sisters [Working Paper]. http://www.pamjakiela.com/bigsisters.pdf

Kerwin, J. T., & Thornton, R. L. (forthcoming). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. *The Review of Economics and Statistics*, 1–45. https://doi.org/10.1162/rest_a_00911

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kremer, M., Brannen, C., & Glennerster, R. (2013). The Challenge of Education and Learning in the Developing World. *Science*, *340*(6130), 297–300. https://doi.org/10.1126/science.1235350

McEwan, P. J. (2015). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research*, 85(3), 353–394. https://doi.org/10.3102/0034654314553127

Romero, M., & Sandefur, J. (2019). Beyond Short-term Learning Gains: The Impact of Outsourcing Schools in Liberia after Three Years. Center for Global Development working paper 521. https://www.cgdev.org/publication/beyond-short-term-learning-gains-impact-outsourcing-schools-liberia-after-three-years

Singh, A. (2015a, January 13). How standard is a standard deviation? A cautionary note on using SDs to compare across impact evaluations in education [Development Impact - World Bank Blog]. https://blogs.worldbank.org/impactevaluations/how-standard-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations

Singh, A. (2015b). Private school effects in urban and rural India: Panel estimates at primary and secondary school ages. *Journal of Development Economics*, *113*, 16–32. https://doi.org/10.1016/j.jdeveco.2014.10.004

Snilstveit, B., Stevenson, J., Menon, R., Philips, D., Gallagher, E., Geelen, M., Jobse, H., Schimdt, T., & Jimenez, E. (2016). *The impact of education programmes on learning and school participationin low- and middle-income countries* (3ie Systematic Review Summary 7). International Initiative for Impact Evaluation (3ie). https://www.3ieimpact.org/sites/default/files/2019-05/srs7-education-report.pdf

Vivalt, E. (forthcoming). How Much Can We Generalize from Impact Evaluations? *s.* http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf

World Bank. (2018). *World Development Report 2018: Learning to Realize Education's Promise*. https://www.worldbank.org/en/publication/wdr2018

World Bank. (2020). *World Bank Country and Lending Groups*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Tables

Table 1. Summary statistics

	All	RCTs	Quasi-Experimental studies
Number of studies	299	156	143
Number of effect sizes	1,266	827	439
Number of learning effect sizes	842	582	260
Number of access effect sizes	424	245	179
Number of countries	52	38	41
By region (number of studies)			
Sub-Saharan Africa	85	51	34
South Asia	52	32	20
Latin America and the Caribbean	104	37	67
East Asia and Pacific	56	36	21
Europe and Central Asia	3	0	3
Middle East and North Africa	1	1	0
By region (effect sizes)			
Sub-Saharan Africa	441	337	104
South Asia	255	171	84
Latin America and the Caribbean	344	156	188
East Asia and Pacific	209	153	56
Europe and Central Asia	7	0	7
Middle East and North Africa	10	10	0

Notes: Region groups follows World Bank (2020). Source: Authors' construction based on effects gathered from Evans and Popova (2016) and Evans and Yuan (2019).

Table 2. Distribution of impacts across randomized controlled trials

Panel A: Learning effect sizes

	Overall	Subject			Sample Size				
		Math	Reading	Other Test Score	<=500	501-1000	1001-3000	3001-5000	>5,000
Mean	0.14	0.09	0.20	0.12	0.18	0.14	0.17	0.10	0.09
Std	0.20	0.17	0.27	0.18	0.23	0.21	0.22	0.19	0.15
P1	-0.28	-0.36	-0.22	-0.48	-0.14	-0.57	-0.22	-0.61	-0.12
P10	-0.04	-0.05	-0.04	-0.06	-0.06	-0.06	-0.05	-0.08	-0.02
P20	0.00	-0.02	0.01	-0.01	-0.02	0.00	0.01	-0.02	0.00
P30	0.03	0.01	0.04	0.03	0.02	0.05	0.06	0.02	0.01
P40	0.06	0.04	0.08	0.07	0.05	0.10	0.10	0.05	0.00
P50	0.10	0.07	0.13	0.11	0.10	0.13	0.13	0.08	0.05
P60	0.14	0.10	0.18	0.15	0.18	0.16	0.16	0.14	0.08
P70	0.17	0.13	0.24	0.19	0.30	0.25	0.21	0.17	0.12
P80	0.25	0.16	0.36	0.24	0.37	0.29	0.29	0.21	0.15
P90	0.38	0.27	0.75	0.35	0.57	0.37	0.45	0.17	0.23
P99	0.84	0.69	0.91	0.58	0.83	0.75	0.90	0.70	0.69
# of effect sizes	582	158	239	139	69	85	209	62	157
# of studies	130	61	61	50	21	27	49	20	37

Notes: Test score includes composite test score and any test score that does not specify subject. Source: Authors' construction based on effects gathered from Evans and Popova (2016) and Evans and Yuan (2019).

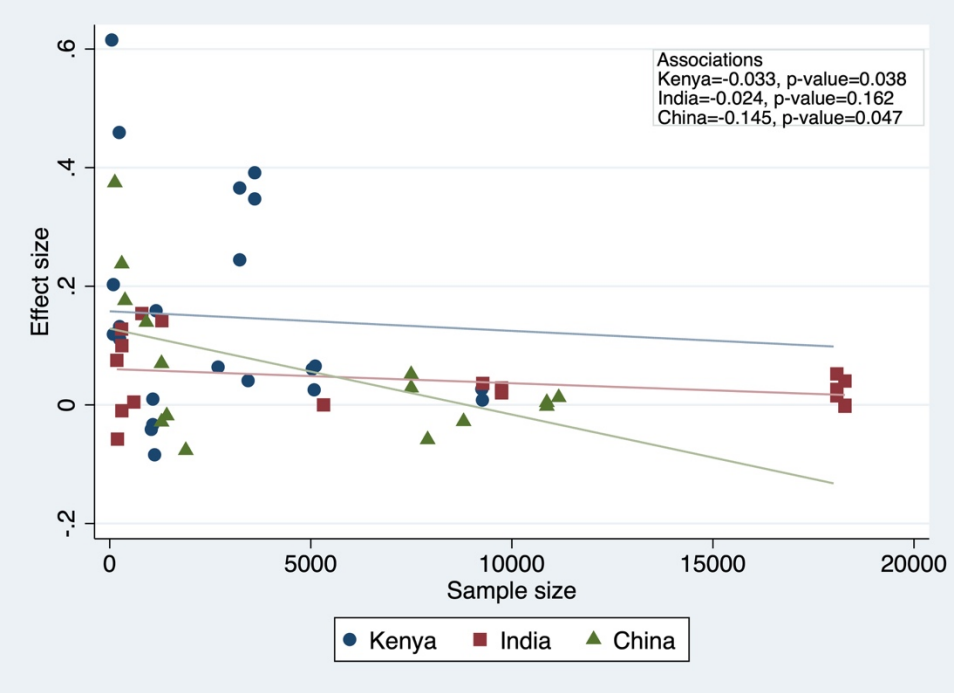
Panel B: Access effect sizes

	Overall	Outcome			Sample Size				
		Enrollment	Attendance	Dropout (Absolute value)	<=500	501-1000	1001-3000	3001-5000	>5,000
Mean	0.10	0.11	0.14	0.05	0.16	0.11	0.06	0.11	0.03
Std	0.14	0.17	0.22	0.09	0.19	0.08	0.11	0.11	0.04
P1	-0.12	-0.10	-0.06	-0.11	-0.14	-0.10	-0.08	-0.05	-0.06
P10	-0.03	-0.03	0.00	-0.07	-0.06	0.01	-0.06	0.03	0.00
P20	0.00	0.00	0.01	-0.02	0.00	0.03	-0.02	0.04	0.00
P30	0.03	0.03	0.03	0.02	0.05	0.05	0.00	0.05	0.01
P40	0.05	0.05	0.06	0.03	0.08	0.08	0.02	0.07	0.02
P50	0.07	0.06	0.08	0.05	0.12	0.12	0.05	0.08	0.03
P60	0.09	0.08	0.11	0.06	0.14	0.14	0.08	0.09	0.03
P70	0.12	0.11	0.14	0.08	0.22	0.16	0.09	0.11	0.05
P80	0.16	0.18	0.18	0.10	0.35	0.18	0.11	0.16	0.06
P90	0.27	0.37	0.36	0.14	0.46	0.21	0.16	0.35	0.08
P99	0.62	0.80	1.25	0.30	0.80	0.29	0.64	0.39	0.14
# of effect sizes	245	75	103	33	76	39	55	25	50
# of studies	74	33	43	15	25	16	19	14	21

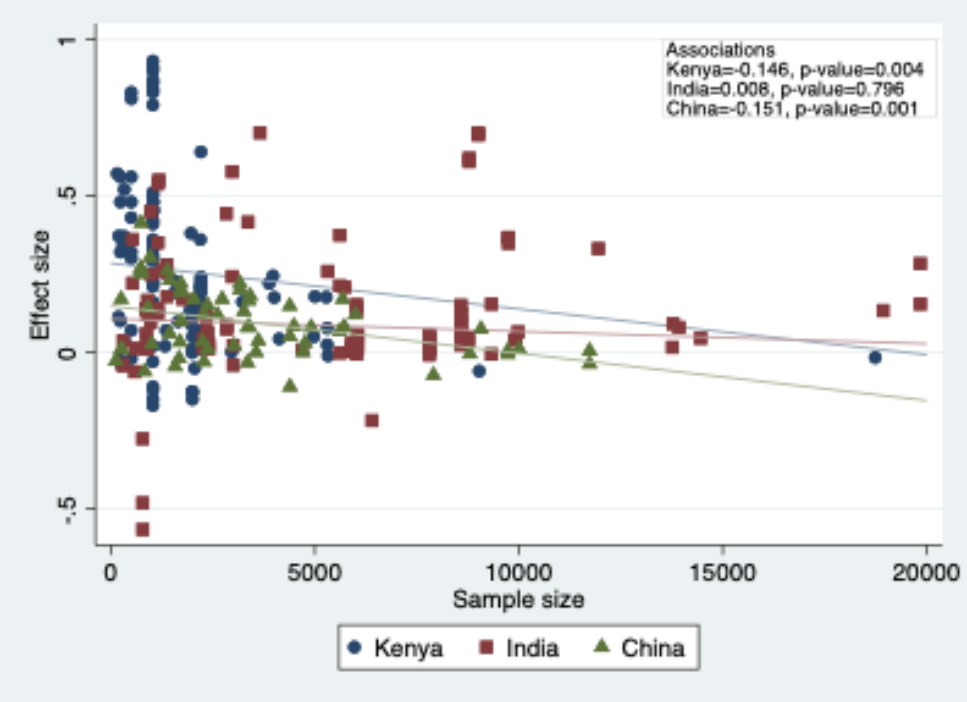
Notes: The effect sizes of dropout are reported in absolute value, i.e. the effect of decreasing dropout. Source: Authors' construction based on effects gathered from Evans and Popova (2016) and Evans and Yuan (2019).

Figure 1. The association between sample size and effect size for randomized controlled trials in China, India, and Kenya

Panel A: Access estimates



Panel B: Learning estimates



Source: Authors' construction based on effects gathered from Evans and Popova (2016) and Evans and Yuan (2019).

Appendix

Table A1. Distribution of impacts across quasi-experimental studies

Panel A: Learning effect sizes

	Overall	Subject			Sample size				
		Math	Reading	Other Test Score	<=1000	1001-3000	3001-5000	5001-10,000	>10,000
Mean	0.16	0.13	0.17	0.15	0.28	0.15	0.16	0.02	0.06
Std	0.48	0.52	0.56	0.30	0.51	0.26	0.39	0.22	0.71
P1	-0.78	-2.62	-0.78	-0.63	-0.78	-0.20	-0.33	-0.72	-3.59
P10	-0.12	-0.14	-0.19	-0.01	-0.20	-0.08	-0.12	-0.22	-0.01
P20	-0.02	-0.05	-0.08	0.01	-0.08	-0.01	-0.07	-0.02	0.02
P30	0.01	0.00	0.00	0.03	0.06	0.01	-0.03	0.01	0.02
P40	0.05	0.03	0.02	0.06	0.12	0.04	0.00	0.05	0.04
P50	0.08	0.05	0.07	0.10	0.22	0.05	0.02	0.11	0.05
P60	0.12	0.09	0.16	0.12	0.30	0.08	0.13	0.12	0.07
P70	0.19	0.17	0.26	0.16	0.46	0.15	0.18	0.15	0.10
P80	0.35	0.35	0.58	0.20	0.61	0.35	0.26	0.16	0.16
P90	0.63	0.47	0.82	0.50	0.84	0.63	0.79	0.17	0.45
P99	1.73	1.75	1.49	1.59	1.94	0.86	1.59	0.20	1.75
# of effect sizes	260	73	107	49	76	64	41	22	57
# of studies	95	49	49	27	26	23	16	11	31

Notes: Other test score includes composite test score and any test score that does not specify subject. Source: Authors' construction based on effects gathered from Evans and Popova (2016) and Evans and Yuan (2019).

Panel B: Access effect sizes

	Overall	Subject			Sample size				
		Enrollment	Attendance	Dropout (Absolute value)	<=1000	1001-3000	3001-5000	5001-10,000	>10,000
Mean	0.13	0.17	0.12	0.11	0.27	0.13	0.10	0.04	0.06
Std	0.40	0.52	0.15	0.31	0.81	0.27	0.15	0.06	0.21
P1	-0.67	-0.12	-0.05	-0.24	-0.80	-0.11	-0.02	-0.05	-0.16
P10	-0.02	-0.01	0.01	-0.02	-0.12	-0.03	-0.01	0.00	0.00
P20	0.01	0.01	0.02	0.00	0.02	0.01	0.02	0.00	0.00
P30	0.02	0.03	0.05	0.00	0.06	0.03	0.03	0.03	0.02
P40	0.04	0.05	0.06	0.02	0.10	0.05	0.04	0.03	0.02
P50	0.05	0.06	0.06	0.02	0.15	0.07	0.07	0.04	0.02
P60	0.07	0.08	0.09	0.03	0.20	0.09	0.08	0.04	0.04
P70	0.10	0.12	0.10	0.04	0.23	0.13	0.12	0.05	0.05
P80	0.17	0.17	0.23	0.13	0.44	0.20	0.13	0.06	0.07
P90	0.27	0.27	0.31	0.30	0.56	0.30	0.23	0.10	0.12
P99	1.98	4.42	0.68	1.40	4.42	1.98	0.68	0.22	1.40
# of effect sizes	179	88	27	22	33	57	19	23	47
# of studies	76	45	21	18	17	20	12	16	29

Notes: The effect sizes of dropout are reported in absolute value, i.e. the effect of decreasing dropout. Source: Authors' construction based on effects gathered from Evans and Popova (2016) and Evans and Yuan (2019).

Table A2. Distribution of impacts by region

Panel A: Learning effect sizes

Region		Mean	Std	P1	P10	P25	P50	P75	P90	P99	# of effect sizes	# of studies
Sub-Saharan Africa	RCT	0.17	0.23	-0.17	-0.06	0.01	0.13	0.27	0.48	0.90	236	45
	Quasi-experimental	0.36	0.47	-0.78	-0.07	0.07	0.34	0.63	0.84	1.94	73	23
South Asia	RCT	0.14	0.21	-0.48	-0.01	0.01	0.08	0.21	0.45	0.70	134	29
	Quasi-experimental	0.12	0.33	-0.72	-0.15	-0.05	0.06	0.23	0.61	1.13	38	12
Latin America and the Caribbean	RCT	0.12	0.26	-0.36	-0.03	0.03	0.09	0.15	0.29	0.77	112	25
	Quasi-experimental	0.07	0.22	-0.42	-0.09	0.00	0.05	0.13	0.25	0.47	105	44
East Asia and Pacific	RCT	0.11	0.14	-0.11	-0.04	0.01	0.08	0.17	0.26	0.75	99	31
	Quasi-experimental	0.13	0.92	-3.59	-0.15	0.01	0.06	0.35	1.41	1.75	39	14
Europe and Central Asia	RCT	-	-	-	-	-	-	-	-	-	0	0
	Quasi-experimental	-0.09	0.14	-0.20	-0.20	-0.20	-0.16	0.02	0.10	0.10	5	2
Middle East and North Africa	RCT	-	-	-	-	-	-	-	-	-	1	1
	Quasi-experimental	-	-	-	-	-	-	-	-	-	0	0

Notes: Region groups follows World Bank (2020). Source: Authors' construction based on effects gathered from Evans and Popova (2016) and Evans and Yuan (2019).

Panel B: Access effect sizes

Region		Mean	Std	P1	P10	P25	P50	P75	P90	P99	# of effect sizes	# of studies
Sub-Saharan Africa	RCT	0.12	0.17	-0.08	-0.03	0.01	0.06	0.17	0.40	0.62	101	27
	Quasi-experimental	0.16	0.36	-0.05	0.01	0.03	0.06	0.14	0.30	1.98	31	16
South Asia	RCT	0.08	0.09	-0.06	0.00	0.02	0.05	0.12	0.16	0.46	37	12
	Quasi-experimental	0.06	0.10	-0.24	-0.03	0.01	0.04	0.09	0.21	0.45	46	14
Latin America and the Caribbean	RCT	0.14	0.30	-0.09	-0.05	0.03	0.08	0.11	0.21	1.46	44	20
	Quasi-experimental	0.10	0.25	-0.80	-0.01	0.02	0.06	0.13	0.27	1.40	83	36
East Asia and Pacific	RCT	0.07	0.11	-0.14	-0.07	0.00	0.07	0.14	0.21	0.37	54	14
	Quasi-experimental	0.41	1.05	-0.01	0.01	0.04	0.13	0.22	0.68	4.42	17	9
Europe and Central Asia	RCT	-	-	-	-	-	-	-	-	-	0	0
	Quasi-experimental	-0.04	0.09	-0.11	-0.11	-0.11	-0.04	0.02	0.02	0.02	2	1
Middle East and North Africa	RCT	0.13	0.11	0.01	0.01	0.05	0.11	0.16	0.30	0.30	9	1
	Quasi-experimental	-	-	-	-	-	-	-	-	-	0	0

Notes: Region groups follows World Bank (2020). Source: Authors' construction based on effects gathered from Evans and Popova (2016) and Evans and Yuan (2019).