How to Improve Education Outcomes Most Efficiently? A Comparison of 150 Interventions Using the New Learning-Adjusted Years of Schooling Metric

Noam Angrist, David K. Evans, Deon Filmer, Rachel Glennerster, F. Halsey Rogers, and Shwetlena Sabarwal

Abstract

Many low- and middle-income countries lag far behind high-income countries in educational access and student learning. Limited resources mean that policymakers must make tough choices about which investments to make to improve education. Although hundreds of education interventions have been rigorously evaluated, making comparisons between the results is challenging. Some studies report changes in years of schooling; others report changes in learning. Standard deviations, the metric typically used to report learning gains, measure gains relative to a local distribution of test scores. This metric makes it hard to judge if the gain is worth the cost in absolute terms. This paper proposes using learning-adjusted years of schooling (LAYS)—which combines access and quality and compares gains to an absolute, cross-country standard—as a new metric for reporting gains from education interventions. The paper applies LAYS to compare the effectiveness (and cost-effectiveness, where cost is available) of interventions from 150 impact evaluations across 46 countries. The results show that some of the most cost-effective programs deliver the equivalent of three additional years of high-quality schooling (that is, schooling at quality comparable to the highest-performing education systems) for just $100 per child—compared with zero years for other classes of interventions.

Keywords: Education; Cost-Benefit Analysis; Government Policies; Government Expenditures; Impact Evaluations

JEL: H43, H520, I2

Center for Global Development
www.cgdev.org

Working Paper 558
October 2020

How to Improve Education Outcomes Most Efficiently?
A Comparison of 150 Interventions Using the New
Learning-Adjusted Years of Schooling Metric

Noam Angrist
University of Oxford and the World Bank

David K. Evans
Center for Global Development

Deon Filmer
World Bank

Rachel Glennerster
United Kingdom Foreign, Commonwealth & Development Office (FCDO)

F. Halsey Rogers
World Bank

Shwetlena Sabarwal
World Bank

Correspondence to noam.angrist@bsg.ox.ac.uk

This work builds on a series of efforts by the authors to produce global public goods in education, including globally comparable test scores, global databases of education, and prior work developing the macro learning-adjusted years of schooling measure. The authors are grateful to the following colleagues for their contributions: Amina Mendez Acosta provided research assistance; Radhika Bula codified and structured data from the Abdul Latif Jameel Poverty Action Lab (J-PAL) database; Yilin Pan compiled data from the World Bank Strategic Impact Evaluation Fund; and Alaka Holla provided guidance on inclusion of studies from the SIEF database and input into categorizations of interventions and cost-effectiveness implications. The views expressed here are those of the authors and should not be attributed to their respective institutions. This work has been supported by a World Bank trust fund with the Republic of Korea (TF0B0356), acting through the Korea Development Institute (KDI) on the KDI School Partnership for Knowledge Creation and Sharing.

Noam Angrist, David K. Evans, Deon Filmer, Rachel Glennerster, F. Halsey Rogers, and Shwetlena Sabarwal, 2020. "How to Improve Education Outcomes Most Efficiently? A Comparison of 150 Interventions Using the New Learning-Adjusted Years of Schooling Metric." CGD Working Paper 558. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/how-improve-education-outcomes-most-efficiently-comparison-150-interventions-using-new

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

202.416.4000
(f) 202.416.4050

www.cgdev.org

Contents

1. Introduction

The average child in a low-income country is expected to attend 5.6 fewer years of school than a child in a high-income country (World Bank 2020).[1] By the age of 10, 90 percent of children in low-income countries still cannot read with comprehension, compared with only 9 percent in high-income countries (Azevedo et al. 2019). With limited resources, policymakers must make tough choices about what to invest in to improve education outcomes—from constructing schools to coaching teachers, from improving school management to deploying new educational software. Making these investment decisions requires comparable data on both the benefits and costs—i.e., the cost-effectiveness—of alternative approaches.

However, the impacts of educational interventions are often reported in ways that make these comparisons difficult. First, policymakers must choose between interventions that increase the number of years a child stays in school and investments that deliver increased learning during those years, without a good way of comparing progress against these alternative outcomes. But policymakers want a combination of the two. Politicians like Boris Johnson and advocates like Malala have called for an increase in the number of "years of quality education," a single concept that incorporates both quality and quantity dimensions (Crawfurd et al. 2020; McKeever 2020). There is evidence that some of the benefits of education, including economic growth, are more closely associated with learning (Hanushek and Woessmann 2012), whereas others are associated with years of schooling (Baird, McIntosh and Ozler 2011; De Neve et al. 2015; Duflo, Dupas, and Kremer 2017). These two dimensions of education cannot be considered entirely separately. Improving the quality of education has more impact if more children go to school for longer, and programs that increase years of schooling lead to more learning if the underlying education system is of a higher quality.

Second, gains are often reported in standard deviations of control-group test scores rather than against an absolute benchmark. Even if studies report absolute changes in test scores, these are not comparable across studies, because different studies use different tests. In countries with different levels of inequality in learning, the same absolute increase in average learning on the same test would generate very different standard-deviation improvements.

Third, current metrics make it hard to judge whether the results of a program are worth the cost. If $100 buys an additional 6 months of schooling for a child, is that a good buy if the quality of schooling is bad? Is $100 for an increase in test scores of 0.05 standard deviation a good investment? The answers depend on the underlying quality of the additional schooling in the first case and on the underlying heterogeneity in learning outcomes in the second.

[1] We calculate this based on a measure of expected years of schooling using source data from the UNESCO Institute for Statistics (UIS) compiled for the World Bank Human Capital Index (2020).

Comparing education gains across age groups and learning levels is methodologically challenging. The learning jump from single-digit subtraction to long division is inherently different from the jump between recognizing letters to being able to read a sentence. But if we conclude these are fundamentally different concepts that cannot be compared, we forfeit the ability to make comparisons across impact evaluations or advise policymakers on the most cost-effective approaches to improving education. When we compare studies using standard deviations as our metric, we impose the assumption that the difference in learning levels between the median and 66th-percentile student in a fourth-grade math class in Kenya is equivalent to the difference in learning levels between the median and 66th-percentile student in a twelfth-grade history class in Peru. A better and more transparent approach to comparing learning gains is to measure them against how long the average student in a high-performing education system would take to make this learning gain (at the appropriate age). This yields a plausible cardinal measure for comparing different types of learning gains: a gain that would take a student in a high-quality system twice as long to achieve is one with twice the educational value.

In this paper, we use this underlying concept to generate a way to measure cost-effectiveness that is closely tied to the policy objective of increasing years of schooling, adjusted for quality: learning-adjusted years of schooling, or LAYS.[2] We make a number of assumptions to apply this measure to evaluations of 150 education interventions across 46 low- and middle-income countries. For a subset of interventions for which cost data are available, we include cost-effectiveness analysis and comparisons. By setting out the benefits of using LAYS, we hope to encourage more researchers to express their results in LAYS. If researchers make greater use of standardized learning assessments, that will facilitate direct, more meaningful comparisons between studies and obviate the need for some of the assumptions made in this paper.

We find that while many interventions are not cost-effective, some of the most cost-effective interventions can deliver the equivalent of three years of high-quality education (i.e., three years of education in a high-performing country such as Singapore) for as little as $100 per child. This finding suggests that despite the huge challenges children and schools face in low- and middle- income countries, from poor health and nutrition of children to weakly performing teachers, the right investments can deliver huge returns, even against the benchmark of the best-performing systems. The three most cost-effective approaches are: targeted information campaigns on benefits, costs and quality; interventions to target teaching instruction by learning level rather than grade (such as "Teaching at the Right Level" interventions and tracking interventions); and improved pedagogy in the form of structured lesson plans with linked student materials, teacher professional development, and monitoring (which includes multi-faceted interventions such as Tusome in Kenya). In India, targeted instruction yields up to 4 additional learning-adjusted years of schooling per $100—

[2] In previous work, the concept of LAYS has been applied to characterizing the differences in quantity and quality of education across countries (Filmer et al. 2020), which we refer to in this study as "macro-LAYS."

a gain equivalent to the entire system-level education gap between India and Argentina.[3] Other interventions like providing school inputs alone (that is, without necessary complementary changes) perform poorly because they tend not to boost access or learning substantively. Shifting the marginal dollar of government expenditure from low-efficiency to high-efficiency educational investments could therefore yield very substantial benefits per dollar spent.

Any attempt to compare education gains across studies requires a number of assumptions. However, our results are robust to a series of tests and alternative choices in the construction of our measure, including alternative specifications of what constitutes high-quality learning, different scaling of test scores, tests for different distributions of performance within samples across countries, and different assumptions about learning in the absence of an intervention or policy change.

This work contributes to three major literatures. The first concerns the use of summary measures to inform policy analysis. Such measures have become foundational in public health, macroeconomics, and welfare analysis. In public health, such measures include Quality-Adjusted Life Years (QALY) and Disability-Adjusted Life Years (DALY), which were first introduced in the 1970s and early 1980s (Torrance et al. 1972; Zeckhauser et al. 1976; Pliskin et al. 1980). While DALYs rely on many assumptions, today they are used widely as the reference standard in cost-effectiveness analysis (Murray et al. 1996; Drummond et al. 2015). In macroeconomics, summary measures such as Gross Domestic Product (GDP) have enabled researchers and policymakers to deepen understanding of economic forces and informed real-time policy responses to economic shocks. Our work introduces a related summary measure for education.

We also contribute to the literature synthesizing results from rigorous impact evaluations in education. Previous reviews examining the impact of educational interventions in low- and middle-income countries include Glewwe et al. 2013, Kremer, Brannen, and Glennerster 2013, Krishnaratne et al. 2013, Evans and Popova 2016a, Ganimian and Murnane 2016, and others. Our analysis takes this literature a step further by using a metric (LAYS) that increases comparability of results across studies. It also updates that literature with recent evaluations and provides cost-effectiveness analysis for more studies than previous work has covered.

Finally, we relate to a literature attempting to inform government intervention through the use of cost-effectiveness and cost-benefit analysis across a very broad range of potential government interventions. Much of this literature conducts cost-effectiveness analysis, but in different ways. For example, higher education analyses typically report the cost per

---

[3] This calibration does not imply that this intervention would necessarily close the gap between country-level education systems, since many interventions are less effective at scale and political economy factors may impede effectiveness at the system-level. Rather, this comparison is meant to illustrate and calibrate the magnitude of effects.

enrollment (Kane 1994; Dynarski 2000) and early childhood education studies often report a social benefit-cost ratio (Heckman et al. 2010). Hendren and Sprung-Keyser (2020) propose a unified analysis using a new measure of Marginal Value of Public Funds (MVPF) and compare benefit and cost information (expressed in monetary terms) to prioritize among 133 social policies in the United States. Their analysis reveals that direct investment by governments in low-income children's health and education in the United States has historically had the highest return on investment and that many such policies pay for themselves. Our study similarly demonstrates that there are investments in education interventions in low- and middle-income countries that can deliver large gains at relatively low cost, even when compared against a benchmark of education gains made by children in high-income countries.

This work, like other syntheses and summary measures, has limitations. First, outputs of this type are only as good as the inputs, and in this case data are still limited, especially on costs. Many education interventions have yet to be evaluated rigorously. As data inputs improve and the range of evaluated interventions expands over time, the outputs of comparative analysis will also improve. Second, in many cases studies report learning outcomes only in standard deviations. We therefore have to use assumptions about the distribution of learning levels in the study area to translate the study findings into LAYS. However, if future studies use common, standardized tests, future comparisons will allow relaxation of that assumption. Third, because both impact estimates and cost measurement are measured with imprecision (Evans and Popova 2016b), it would be unwise to focus on small differences in cost-effectiveness. Rather, this analysis aims to inform broad trade-offs at the aggregate level in cases where there are large, consistent differences. For example, we consistently see that as a cost-effective tool for improving LAYS, cash transfers rank lower than investments in early childhood development. This pattern is robust to method, data inputs, and study or country contexts. Fourth, while access to school and learning proxied by test score performance capture important components of education, they do not capture all aspects of education (like socioemotional learning). However, the combination of these measures represents an improvement over the status quo where typically only one measure is used. Fifth, context matters. Even for the most cost-effective interventions, policymakers will have to consider whether contextual conditions support local adaptation of an intervention (Bates and Glennerster 2017).

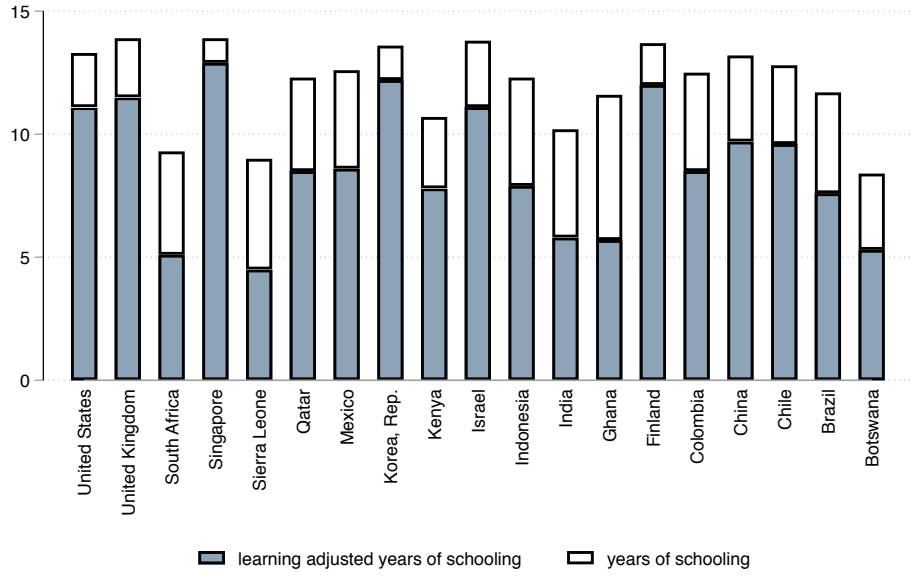
The rest of paper proceeds as follows. Section 2 provides a framework for the learning-adjusted years of schooling. Section 3 describes the set of studies and data included in the analysis of education policies and interventions. Section 4 presents the results in terms of both effectiveness and cost-effectiveness. Section 5 provides a series of robustness tests, and Section 6 concludes.

2. Learning-Adjusted Years of Schooling Framework

Learning-adjusted years of schooling for a given country—what we call macro-LAYS—are the product of years of schooling and a measure of schooling quality (Filmer et al. 2020).

Specifically, they are produced by scaling the country's average schooling by its test-score performance relative to a global high-performance benchmark.[4] Figure 1 shows an example using data from the World Bank Human Capital Index. For example, Singapore's average student test scores are closer to the high-performance benchmark than any other country's scores. As a result, its 14 average years of schooling are discounted only slightly, to 13 LAYS. By contrast, South Africa has 10 years of schooling but only about 5 LAYS, because its test scores are only about half of Singapore's. In other words, macro-LAYS are produced at the country level by adjusting average schooling in a given country by the amount of learning in that country (relative to a high-performance benchmark). Expressing national education levels in terms of macro-LAYS provide a unified and user-friendly measure for a variety of education outcomes.

Figure 1. Years of Schooling and Learning-Adjusted Years of Schooling (Macro-LAYS)



Note: Schooling data is based on UNESCO expected years of schooling and learning data is based on Harmonized Learning Outcomes (HLO).

Source: The Human Capital Index is described in Kraay (2019) and is based on Angrist et al. (2019) learning data and UIS enrollment data.

[4] The high-performance benchmark used in the World Bank Human Capital Index is an artificial benchmark of high performance of 625 as defined by the international assessment Trends in Mathematics and Science Study (TIMSS), which was chosen because that benchmark is stable over time and is apolitical (Kraay 2019). Other high-performance benchmarks can also be used to construct LAYS estimates. For example, we can use the top-performing country. If Singapore is the highest-performing country in a given year, we can express every country's LAYS in Singapore-equivalent years. That is, we could say that a student in South Africa achieves 10 years of schooling, but 5 years of Singapore-quality schooling.

In this section, we show how LAYS can also be used at the micro level to compare education interventions. The number of rigorous studies evaluating the effect of interventions on educational outcomes is growing, with nearly 300 impact evaluations focused on learning outcomes in low- and middle-income countries as of 2016 (World Bank 2018). An improved and more comparable metric would enable better evidence synthesis and clearer policy recommendations. As described below, we aim to address many of the challenges that limit current comparisons—mostly notably, that access impacts and learning impacts are often discussed separately, and that learning gains can be expressed only relative to local performance. We do this by expressing education outcomes from interventions and policies in terms of LAYS units that offer a single, internationally comparable, and policy-salient metric. Hereafter, we will refer to LAYS gained from an intervention or policy as micro-LAYS.

If impact evaluation studies tested students, and reported results, against internationally agreed test scores such as the Trends in International Mathematics and Science Study (TIMMS), the Programme for International Student Assessment (PISA) or the Early Grade Reading Assessment (EGRA), the translation into LAYS would be straightforward. This is what we would hope to see in future studies. However, this is currently not the norm, and therefore a number of assumptions are needed to translate existing studies into LAYS. To ensure a coherent unifying approach, the micro-LAYS methodology invokes assumptions similar to those used in constructing country-level macro-LAYS estimates. In this section, we outline the approach to producing micro-LAYS for studies that report effects on schooling participation, such as attendance or years of school gained, and subsequently for studies that report effects on learning outcomes.

2.1 Micro-LAYS using schooling participation estimates

When studies report effects on schooling participation, micro-LAYS are the product of: (1) the access gains resulting from the intervention and (2) the schooling quality in the country where the intervention took place, measured relative to a global benchmark of high performance. We then multiply these gains by the duration over which the effects of the intervention are expected to persist. The construction of micro-LAYS derived from impacts on schooling participation, denoted by superscript *p,* can be expressed as follows:

$$\mathrm{LAYS}^{p} = \gamma_{\mathrm{i}} * \mathrm{t} * \mathrm{L}_{\mathrm{i}}^{\mathrm{h}}$$

where L_i^h is a measure of learning for a cohort of students in country *i* relative to a high-performance learning benchmark *h*, such that $\mathrm{L}_{\mathrm{i}}^{\mathrm{h}} = \frac{\mathrm{L_i}}{\mathrm{L_h}}$. Because interventions differ in the duration of their impacts, we include a multiplicative factor *t* that represents how long the intervention effects γ are expected to persist.

In our analysis, we explore various options for the time over which the intervention is expected to be effective. These include per single year *(t=1)*; the length of the evaluation (*g*); and the remaining school life expectancy (*s*). Consider a case where schools are built in a remote area of Afghanistan and we observe that the intervention delivers on average an additional year of globally benchmarked high-quality schooling per child over the course of

an evaluation. If we assume that students will stay in school once the school is built and that the quality of schooling remains constant, we can then adjust this estimate by the remaining school life expectancy (i.e., the number of grades in a given school system minus the grade at which the students entered the school), because we expect students to continue to benefit even after the evaluation period. If students entered in grade three and there are seven grades in primary school, then we would simply multiply the additional year of high-quality schooling by four. Thus, the intervention produced four years of high-quality schooling. In our main analysis we restrict parameter *t* to the observed gains over the length of the evaluation (*t*=*g*).

2.2 Micro-LAYS using learning estimates

When studies report effects on learning gains, we first express the learning gains from the intervention in terms of a quantity measure, the equivalent years of schooling gained in a given country with "business as usual" learning. For example, if students learned 0.25 standard deviation per year as a result of an intervention in South Africa, and if students typically learn 0.25 standard deviation per year in a given year in South Africa, then students will have learned a year's worth of South African schooling as a result of the intervention. Second, we then apply a global quality-adjustment factor to derive the corresponding LAYS. For example, if South African students learn half as much as the high-quality benchmark on an international test, we adjust the one year's worth of South African schooling to reflect that it is worth half a year of globally benchmarked high-quality schooling. In the third and final step, we introduce a multiplicative factor for the period of time over which effects are expected to last. As an example, if students had fallen behind grade level and an intervention enables students to catch up to grade level, they might now benefit from day-to-day schooling for the remaining school life expectancy.

Formally, we first express the intervention's learning impacts in terms of equivalent years of school gained. Drawing on the methodology used by Evans and Yuan (2019), we derive equivalent years of schooling, *e*, by expressing learning gained relative to learning in the status quo:

$$e = \frac{\beta_i^{\sigma,test}}{\delta_{i,x}^{\sigma,test}}$$

where β is the learning gain produced by the intervention expressed in standard deviations (σ) per year in country *i; test* denotes the test used to measure learning; δ is the status-quo learning rate in standard deviations (σ) per year in country *i;* and *x* denotes the population for which this status-quo learning rate is calculated. This population *x* could represent the control group of the same study from which the β estimates are drawn; alternatively, *x* could be the student population in country *i,* in which case δ becomes the average learning trajectory for the country as a whole. The choice of *x* will affect our interpretation. If we choose the control group, then the resulting value for equivalent years of schooling gained is relevant to the study sample only. If we choose national-average learning trajectories, we can interpret the value as the equivalent years of schooling gained at the national level. In this

paper's main calculations, we use national-level learning trajectories *n*, and in the robustness section, we explore the trade-offs involved in using a different measure of status quo learning.

We estimate micro-LAYS derived from impacts on learning, denoted by superscript *l*. To derive these estimates, we adjust equivalent years of schooling, *e*, gained in country *i* by the quality of learning L_i^h in that country relative to learning in a high-performance global benchmark country *h*:

$$\text{LAYS}^{\,l} = \overbrace{\frac{\beta_i^{\sigma,\text{test}}}{\delta_{i,n}^{\sigma,\text{test}}}}^{\text{equivalent years of school}} * \overbrace{L_i^h}^{\text{learning adjustment}}$$

We substitute in terms for $L_i^h = \frac{L_i}{L_h}$. This is analogous to the quality adjustment used in macro-LAYS. We further specify that both the numerator and denominator of the learning-adjustment term are expressed in terms of standard deviations (σ) on a *test* that is representative at national level *n* for each country, such that:

$$\text{LAYS}^{\,l} = \overbrace{\frac{\beta_i^{\sigma,\text{test}}}{\delta_{i,n}^{\sigma,\text{test}}}}^{\text{equivalent years of school}} * \overbrace{\frac{L_{i,n}^{\sigma,\text{test}}}{L_{h,n}^{\sigma,\text{test}}}}^{\text{learning adjustment}}$$

For the next step we invoke two assumptions. First, we assume that learning is constant along a local trajectory. This assumption, explored in depth in Filmer et al. (2020) for macro-LAYS, enables conversion of relative *levels* $L_i^h = \frac{L_i}{L_h}$ into relative *rates* $L_i^h = \frac{\delta_i}{\delta_h}$, since the relationship is constant. (This assumption is discussed in more detail in section 5.) Second, we assume that standard deviations across tests and samples are comparable. This assumption, though not trivial, is not novel: it is implicitly invoked any time standard-deviation effect sizes are compared across studies, which is the dominant practice in the literature on education interventions. We note that this assumption is most robust when learning gains in a given study are based on similar tests to the ones used in computing the learning-adjustment factor. We further explore robustness to this assumption in section 5. These assumptions simplify our conversion to:

$$\text{LAYS}^{\,l} = \overbrace{\frac{\beta_i}{\delta_{i,n}}}^{\text{equivalent years of school}} * \overbrace{\frac{\delta_{i,n}}{\delta_{h,n}}}^{\text{learning adjustment}}$$

The $\delta_{i,n}$ terms cancel, and we are left with the expression

$$\text{LAYS}^{\,l} = \frac{\beta_i}{\delta_{h,n}}$$

This expression produces an intuitive metric: the years of *h*-quality learning from the intervention. For example, assume that an intervention in South Africa yields 0.25σ per year of learning ($\beta_{South\ Africa} = 0.25$), and that in Singapore, a high-performance benchmark

on international learning assessments, students learn 1σ over the course of a given year ($\delta_{Singapore} = 1$). Then we have 0.25 LAYS^{l}; in other words, the intervention enabled South African students to gain a quarter of a year's worth of Singaporean-quality schooling.

Finally, we apply a multiplicative factor *t* for the length of time over which the intervention is expected to have lasting effects:

$$\text{LAYS}^{l} = \frac{\beta_{i}}{\delta_{h,n}} * t$$

As with micro-LAYS based on participation estimates, *t* can take on a few values: a single year, such that *t=1*; the length of the evaluation, *g*; and the remaining school life expectancy *s*. As an example, if students had fallen behind grade level and an intervention enables students to catch up to grade level, they might now benefit from day-to-day schooling for the remaining school life expectancy, *s*.[5] We find that micro-LAYS are robust to a range of sensitivity and robustness tests, outlined in section 5.

2.3 Putting micro-LAYS estimates together

In summary, both participation- and learning-based LAYS tell us that a given intervention in a given country produces a certain number of years' worth of globally benchmarked high-quality learning. Thanks to this common unit, the impacts of studies that measure these two different types of outcomes can be directly compared, as we will illustrate below. One challenge in assembling these micro-LAYS estimates is how to handle a study that reports impacts on both participation and learning. If we sum the estimates, we will double-count in cases where gains in learning resulted directly from gains in participation or where gains in learning led to gains in participation (e.g., because students had a greater incentive to attend schools that delivered more learning). As an alternative to adding the two estimates, we could choose to use only estimates from either participation or learning. However, under this approach we would be assuming that one is the central output, and that the other outcome dimension is largely captured within that central output. Instead, for the purposes of this paper, we use the LAYS impact that is greater—whether that was obtained through schooling or learning increases—for each evaluation. This approach places *a priori* equal weight on schooling and learning, introduces no new assumptions, and avoids double-counting.

3. Data and Analysis Framework

We now compare impact estimates from over 150 evaluations of various education interventions in 46 countries, using the unified measure. In our comparison, we highlight findings from a subset of studies that have comparable cost data and that therefore allow us to compare cost-effectiveness of interventions. We examine how many LAYS each policy or

[5] The value of *t*, the length of time an intervention's effect is expected to last, might vary by intervention and apply differently to quality improvements versus quantity improvements.

intervention delivers; how cost-effective those gains are; and how much of the gap between quality-adjusted years of schooling and actual years of schooling that intervention would close if it were scaled up, assuming that the effectiveness of the intervention remained constant. This assumption of scalability is not trivial, given that effectiveness at system scale is often substantially lower than effectiveness in even a large pilot study; we therefore carry out this calculation as a calibration exercise rather than a simulation exercise.

We start with over 300 studies compiled from a database produced by Evans and Yuan (2019), which draws studies from a wide range of reviews (Evans and Popova 2016a; Glewwe et al. 2013; Kremer, Brannen and Glennerster 2013; Krishnaratne et al. 2013; Ganimian and Murnane 2016). We then add 13 studies from the World Bank Strategic Impact Evaluation Fund (SIEF), as well as four additional recent studies that have rigorous evaluation methodologies and high-quality impact and cost data (Romero, Sandefur, and Sandholtz 2020; Sabates et al. 2018; Piper et al. 2018; Eble et al. 2020).

Our inclusion criteria are that the study should be based on a credible causal inference strategy, using either randomized controlled trials or quasi-experimental methods, such as differences-in-differences, instrumental variables, regression discontinuity, fixed effects, or propensity score matching. (To aggregate across outcomes, we code outcomes such that positive impacts always represent an improvement; for example, a reduction in absenteeism is coded as an increase in attendance.) In addition, for studies that report impact on learning, we start with impact estimates that can be expressed in terms of standard deviations. The list of studies included in our analysis is illustrative rather than exhaustive, and in the future, we aim to continue adding more studies and build as comprehensive a database of education interventions as possible. In total, after applying our inclusion criteria, we analyze data from over 150 impact estimates across 46 low- and middle-income countries.[6]

In this analysis, we make several choices for parameters and data inputs. First, we assume that the intervention's effects last only for the duration of the evaluation, since this is a known quantity and requires no further assumptions. In Appendix Figure B1 we explore the alternative assumption that the impacts last just one year. A second choice that we make is to set the high-quality benchmark learning rate equal to 0.8 standard deviations per year. This is a conservative estimate for high rates of learning, drawn from year-on-year learning gains in high-performing education systems, policy-relevant differences across education systems, and standard benchmarks. We choose an artificial benchmark for this analysis, because unlike the actual learning rates of high performers, such as Finland or Singapore, it has the advantage of being stable over time and of being apolitical. This approach to defining high-quality learning *rates* is similar to the one used to define the high-performance benchmark learning *level* in the World Bank Human Capital Index, which sets the benchmark at 625 on the scale of TIMSS and the Programme for International Student Assessment (PISA). In the

[6] This approach has uses beyond comparison of effects of impact evaluations. LAYS is a unit of measurement and therefore any result, including descriptive and observational results, can technically be expressed in terms of LAYS.

robustness section, we explore four plausible approaches to validate this 0.8-standard-deviation high-performance benchmark.

We calculate how much to adjust improvements in access for the level of learning (i.e., the learning adjustment rate) using Harmonized Learning Outcomes (HLO), which are globally comparable measures of learning introduced by Angrist et al. (2019). We choose HLO data over alternative test score data for various reasons. First, these data enable us to use the same learning scale for interventions from 164 countries across the world, a wide range of countries from which we also draw impact evaluation education estimates. Second, since these data are used in the World Bank Human Capital Index (HCI), this enables us to produce micro-LAYS that map directly to the macro-LAYS in the HCI. The international tests of student learning that are included in the HLO data are often scaled to a mean of 500 and standard deviation of 100. For micro-LAYS, we also derive a learning scale whose lower limit plausibly represents zero learning. We use data from early grade reading assessments (EGRA), where underlying test items have a plausible zero: no reading comprehension. In Appendix Figure B2 we show that the HLO score that corresponds to a floor of zero reading comprehension is 300. In accordance with this, in our analysis we scale the HLO data with a linear transformation of 300. In section 5, we further explore the sensitivity of results to the score scale.

4. Results

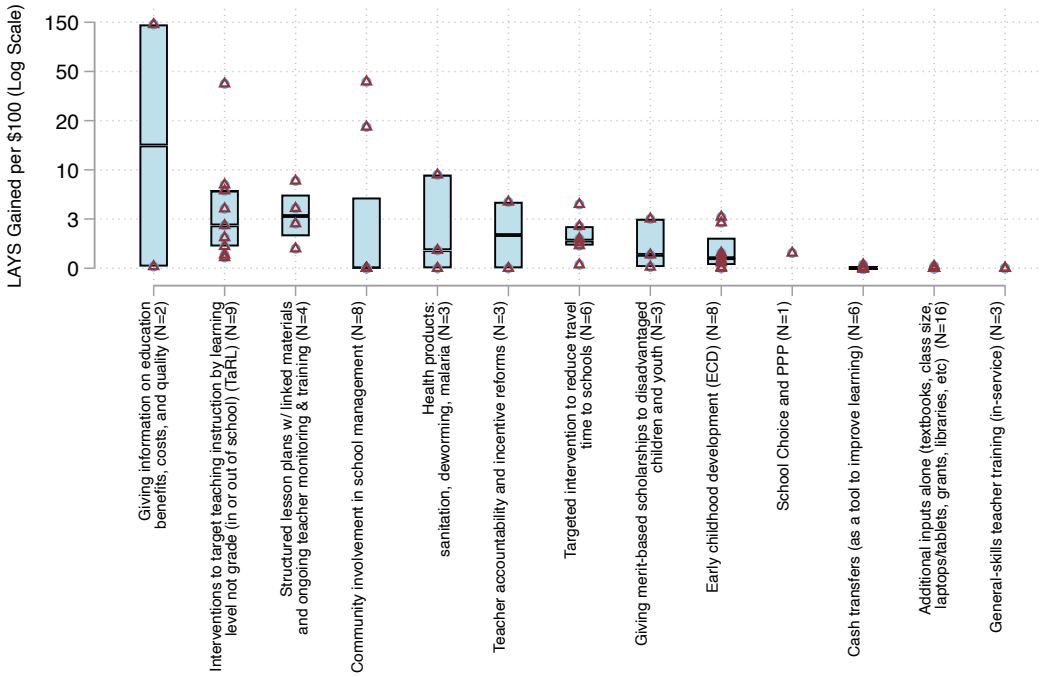
4.1 Aggregate categories of policies and interventions

We are interested in comparing the impacts of aggregate categories of policies and interventions. To this end, we summarize results by category of education intervention, such as Early Childhood Development (ECD) or instruction targeted to the child's level of learning rather than grade level. Intervention categories are based on original study designations, with a few adjustments. First, we recategorize technology interventions into either "targeted instruction" or "additional inputs alone" based on whether they involved adaptive software or were largely a hardware-based intervention. Second, we classify interventions for ECD that involved building or opening of schools or classrooms as "targeted intervention to reduce travel time to schools." Third, we define teacher training interventions narrowly. Many interventions include training of teachers; for this analysis, when a program provides materials to help teachers target instruction to the level of the child and also provides training to those teachers, we classify that as a "targeted instruction" intervention. "General-skills teacher training" captures only general teacher training programs without other major elements. Fourth, for interventions with multiple components, we selected the central component and used that. Later in the paper, we examine individual studies and so will characterize them more precisely.

Comparative information on effectiveness will be most useful to policymakers when it incorporates information about cost. Therefore, we start by analyzing cost-effectiveness of policies and interventions with a subset of studies where cost information is available.

Figure 2 shows the LAYS gained per $100 per child. To calculate this, we divide the per-student gains by the per-student costs, so the figure shows LAYS gained per student per $100 spent per student. Typical spending in education systems ranges between $208 per student in Sub-Saharan Africa to $7,908 in East Asia in primary school in terms of 2013 PPP USD (Bashir et al. 2018). Therefore cost-effectiveness per $100 per student is a relevant unit for many systems, along the lines of status-quo spending benchmarks, even at the lower tail of spending. Of course, the share of overall investment that $100 represents will depend substantially on context.

Figure 2. Learning-Adjusted Years of School (LAYS) Gained per $100 by Category



Notes: Each category of education intervention shows the learning-adjusted years of school (LAYS) gained from a given intervention or policy. Each red triangle represents a cost-effectiveness estimate. The boxplot is ordered from largest to smallest mean effects and the shaded boxplot describes the 25th and 75th percentile, with whiskers at upper and lower fences at a distance of 1.5 times the interquartile range above and below the nearest quartile. The y-axis is reported on a natural log scale.

The top performers, ranked by mean effect size, are: targeted information campaigns on benefits, costs and quality; interventions to target teaching instruction by learning level rather than grade (such as "Teaching at the Right Level" interventions and tracking interventions); improved pedagogy in the form of structured lesson plans with linked materials and monitoring (which includes combination interventions such as Tusome in Kenya); community involvement in school management (such as training for community members); health products (such as anti-malarial or deworming pills); scholarships for disadvantaged groups; teacher accountability and incentives (such as camera monitoring of teacher

attendance or merit based pay); targeted interventions to reduce travel time to school (for example, constructing schools in remote underserved areas); merit-based scholarships provided to disadvantaged children and youth; early childhood development (ECD); and school choice and public-private partnerships (such as voucher schemes). The last three categories—cash transfers, additional inputs alone (such as textbooks, technology hardware, uniforms, school grants, or reducing class size without complementary reforms), and general skills teacher training—have zero effect on LAYS.

Some of these categories have moderate effects in absolute terms, but are extremely cheap, making them very cost-effective; an example is providing information on the returns to schooling. Other interventions are highly effective in absolute terms, but are expensive, and are thus moderately cost-effective; these include school construction to reduce travel times to school as well as scholarship schemes. Moreover, we observe that some categories have low variance—as in the case of class-size reductions and additional inputs, which are tightly concentrated around zero—while other categories have high variance. An example of the high-variance group is information campaigns on the costs and benefits of education: some of the impact estimates for this category are around zero, while others are at the highest end of the spectrum. Structured lesson plans produce large gains with relatively low variation, whereas community involvement has a lower average effect but high variation. This indicates that when considering interventions, we should consider not only the average effect but also the variance. This further points to the importance of contextual relevance: some interventions have similar effects across contexts, while others work extremely well in one context but not in others.

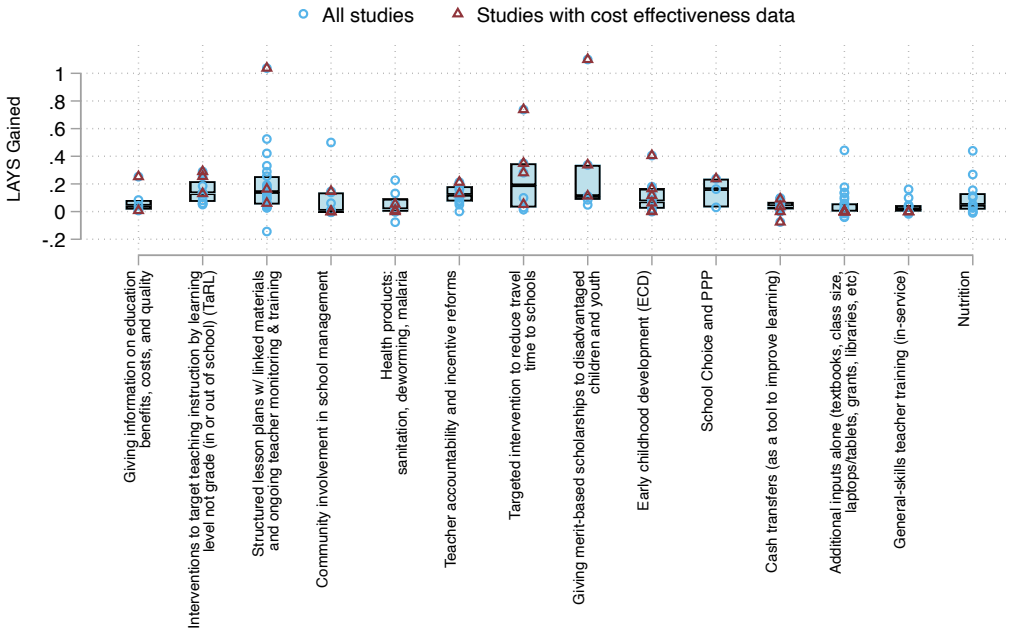
Moreover, context is essential to consider across all categories regardless of variation. For example, early childhood development might be most effective in contexts with strong primary education systems where these early investments translate into preparedness for primary school, thus enabling dynamic complementarities (Johnson and Jackson 2019); providing information on the returns to education may be highly cost-effective in one country but ineffective in a context where those returns are well known; and similarly, a deworming program is unlikely to be cost-effective in a place with low levels of intestinal worms.

It is important to consider these results in the context of how governments typically spend their budgets. They make substantial investments in textbooks, technology hardware, uniforms, school grants, class-size reductions, and general-skills teacher training. When not well integrated with other interventions, these categories of interventions consistently produce almost no effect. By contrast, investments such as early childhood development and merit-based scholarships to disadvantaged students can yield gains of up to 3 additional LAYS per $100 per child. To this end, shifting the marginal dollar of government investment from status-quo spending to more efficient educational investment could substantially improve education outcomes. These implications are consistent with the findings of prior cost-effectiveness reviews, such as Kremer, Brannen, and Glennerster (2013).

Beyond reinforcing findings of prior reviews, our unified analysis also reveals new insights. One is that many interventions that increase participation in schooling are less cost-effective than interventions that improve the *productivity* of schooling—that is, the amount of actual learning gained while in school. For example, prior reviews have shown that cash transfers can increase schooling. However, those results have not been compared to those of interventions that improve learning directly. We find that cash transfers are not a cost-effective tool to improve LAYS, since they yield gains in schooling in systems with low-quality education and do so without improving learning across the studies in our sample, all at relatively high cost. By contrast, some policies that improve the *productivity* of each year of schooling, such as targeting instruction to a child's learning level or structured lesson plans, can yield on average of around 3 additional LAYS per $100. This does not imply that cash transfers are not a useful tool to improve social welfare in general; indeed, research has shown they can be highly effective in achieving their primary aim of reducing poverty and increasing consumption (Fiszbein et al. 2009; Haushofer and Shapiro 2016). Rather, these results suggest that if the goal of governments is to improve quality education, cash transfers might not be the most efficient tool for this specific purpose. More broadly, our analysis reveals the importance of focusing on policies and interventions that improve the productivity of schooling, rather than solely providing additional schooling.

In Figure 3 we show the full set of 150 studies from 46 countries, with the subset of impact evaluations with cost-effectiveness data highlighted. The first important takeaway from this figure is that, by and large, the subset of interventions with cost-effectiveness data are not systematically biased towards high or low impacts, although within each category the studies with cost-effectiveness data may skew one way or the other. This figure further enables us to assess LAYS gains in absolute terms, rather than per $100, and decompose whether an intervention is cost-effective due to being effective, cheap, or both. For example, health products are moderately effective in improving outcomes, with up to 0.2 LAYS gains per child, but are cheap. Thus, in Figure 2 we see the modest absolute gains translate into up to 10 LAYS gains per $100 per child, marking these health interventions as highly cost-effective. Other interventions are highly effective but expensive, and therefore less cost-effective. Giving merit-based scholarships can yield up to 1 LAYS, but since this policy is relatively expensive, it is less cost-effective than the health programs, delivering 3 LAYS per $100 at the upper end of the distribution of studies. Finally, Figure 3 also includes a new category: nutrition interventions (such as school feeding), which did not enter the cost-effectiveness analysis in Figure 2 due to a lack of cost data. We observe that school feeding is relatively effective in improving LAYS, although with high variance. Taken together with the findings in Figure 2, this indicates that health and nutrition interventions can translate into meaningful education outcomes.

Figure 3. Learning-Adjusted Years of School (LAYS) Gained by Intervention Category



Notes: Each category of education intervention shows the learning-adjusted years of school (LAYS) gained from a given intervention or policy across over 150 impact estimates in 46 countries. The boxplot describes the 25th and 75th percentile, with whiskers at upper and lower fences at a distance of 1.5 times the interquartile range above and below the nearest quartile. The boxplot is ordered in the same order as Figure 2 to provide a direct analogy, with the exception of the "nutrition" category which has no cost-effectiveness data and does not appear in Figure 2.

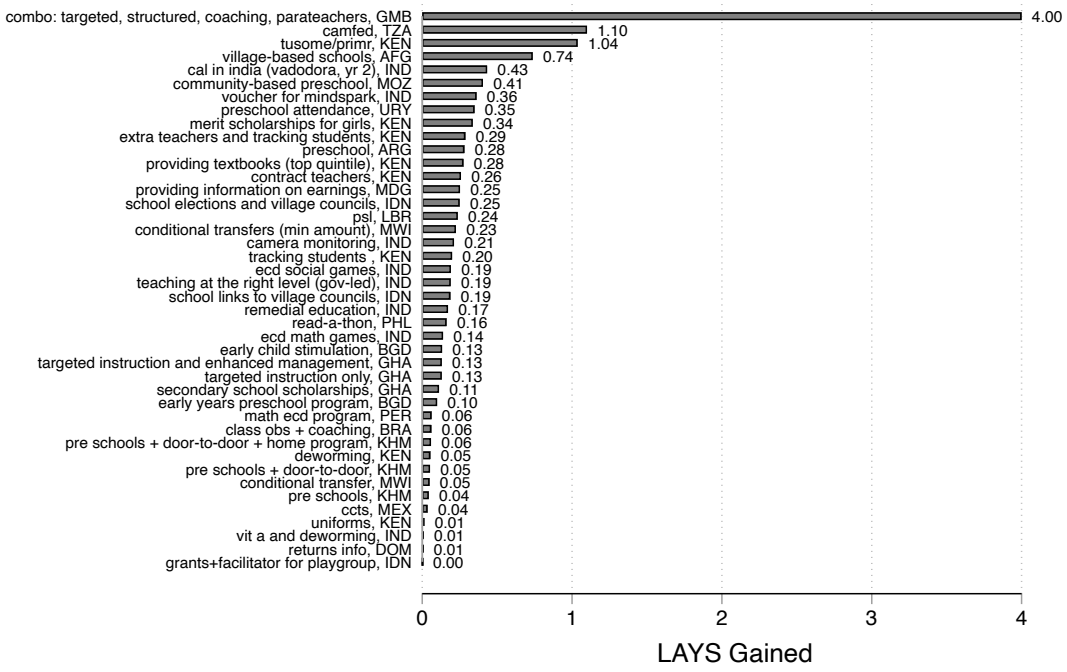
4.2 Specific cost-effectiveness studies

4.2.1 Effectiveness and cost-effectiveness

Next, we examine specific interventions to explore the degree to which aggregate patterns might parallel more granular ones or reveal underlying heterogeneity. Figure 4 shows results for absolute LAYS gained by intervention and country for the studies that include cost-effectiveness data. The top ten performers are: a combined intervention with improved pedagogy, para-teachers and targeted instruction in The Gambia (4 LAYS); the Campaign for Female Education (Camfed) program in Tanzania—a program that provides scholarships for girls along with school materials and training for teachers and parents (1.1 LAYS); Tusome (the Kiswahili word for "Let's Read") in Kenya—a program that provides textbooks, teacher coaching, and teacher training (1.04 LAYS); building village-based schools in Afghanistan (0.74 LAYS); computer-assisted learning (CAL) in India (0.43 LAYS); community-based preschools in Mozambique (0.41 LAYS); vouchers for mind-spark adaptive learning software in India (0.36 LAYS); preschool attendance in Uruguay (0.35 LAYS); merit scholarships for girls in Kenya (0.34 LAYS); and ability grouping using extra

teachers in Kenya (0.29 LAYS). By contrast, about half of all interventions produce no significant effects; those interventions are not included in the figure.

Figure 4. Learning-Adjusted Years of Schooling (LAYS)

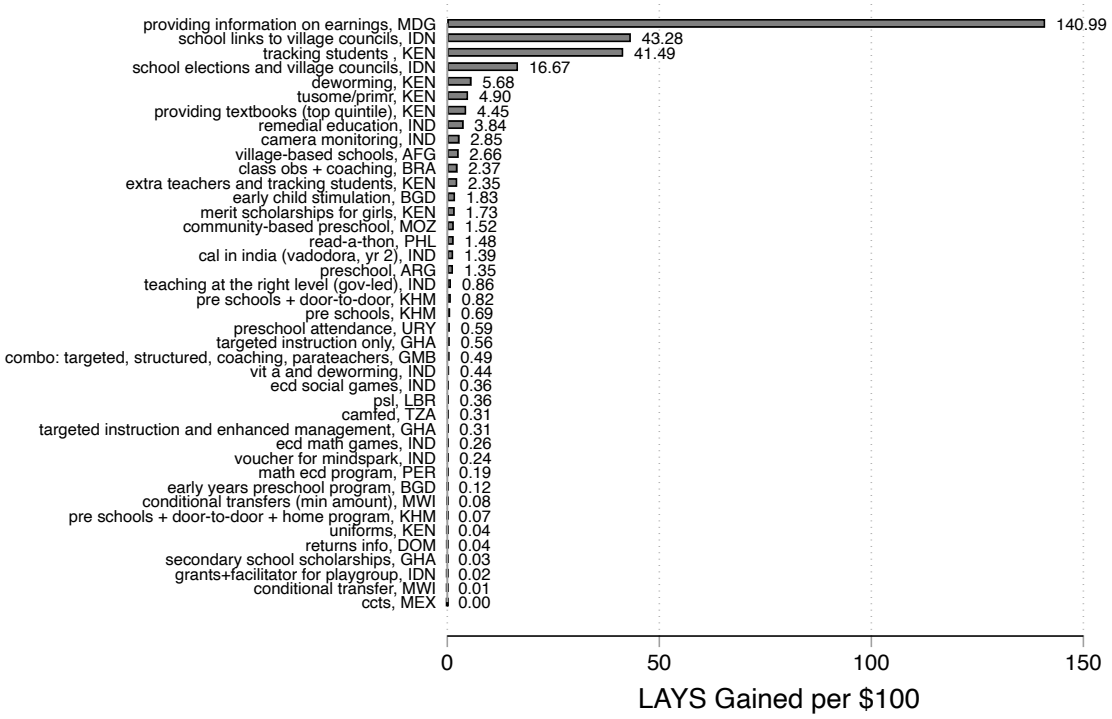


Notes: We do not include interventions with null impacts, which are not cost-effective by definition.

These findings point to a few overall lessons. Among the most effective programs—in this illustrative sample of studies—are: multidimensional programs (a combined intervention in The Gambia, Camfed in Tanzania, and Tusome in Kenya); programs that are carefully targeted to a local need, such as scholarships (for girls), information (when returns to schooling are not widely known), and school construction in under-serviced remote areas; pedagogical instruction that is pitched to students' levels of learning, not based on a rote syllabus or an over-ambitious curriculum; and programs that facilitate early childhood development.

Figure 5 shows cost-effectiveness estimates for these interventions, expressed in LAYS per $100. When we take cost into account, several new interventions join the list of top performers: provision of information on the returns to schooling in Madagascar; school links to village councils in Indonesia, remedial education in India, camera-based monitoring of teachers with incentives in India, deworming in Kenya, and classroom observation and coaching in Brazil. By contrast, public-private partnerships, scholarship programs, targeted school construction and access, and technology-assisted adaptive instruction (such as Mindspark in India) drop down the list because of their higher costs, although these programs continue to be cost-effective, with LAYS per $100 gained ranging between 0.24 to 2.66.

Figure 5. Learning-Adjusted Years of Schooling (LAYS) per $100



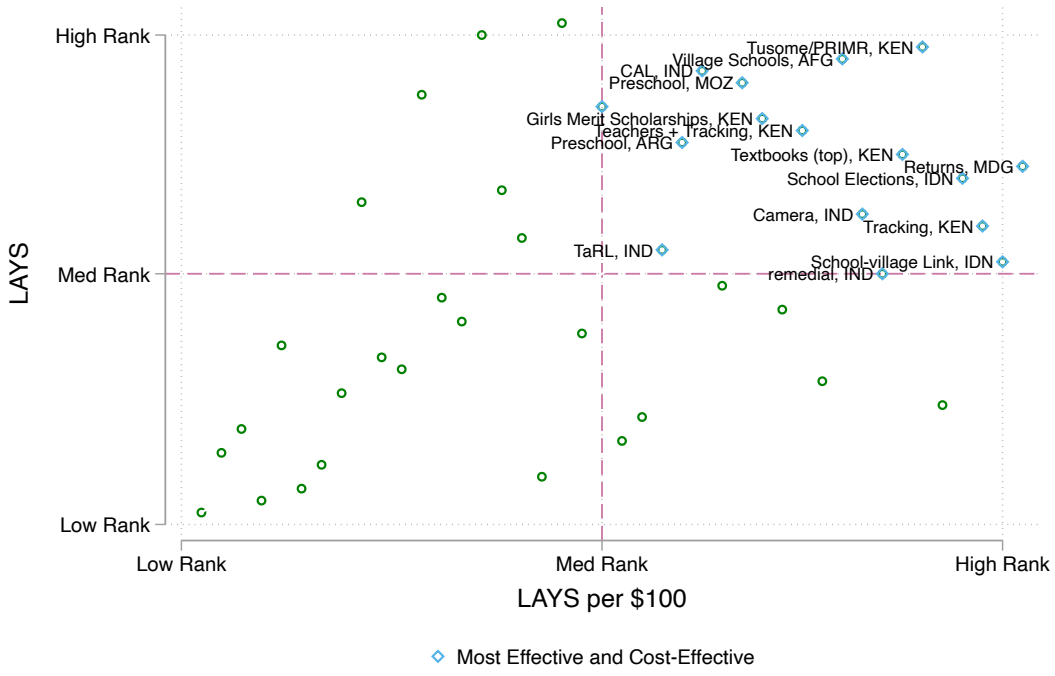
Notes: We omit interventions with a null effect.

It is important to highlight that the cost-effectiveness of some interventions is an order of magnitude greater than the median. These highly cost-effective interventions include providing information on the returns to schooling in Madagascar, creating school links to village councils in Indonesia, and grouping students by ability level in Kenya. These interventions stand out for being both effective and extremely cheap.

The upper-right quadrant of Figure 6 highlights the interventions that are both effective and cost-effective. The programs that do well on both measures include: targeted information (on future earnings) and targeted scholarships (for girls); accountability reforms, such as camera monitoring of teachers with incentives, and school elections and community engagement; instruction targeted to student levels through pedagogical interventions, grouping, and technology; school construction in remote areas that otherwise lack school access; structured pedagogy interventions, and early childhood development programs.

Overall, this exploration of specific interventions reveals broadly consistent patterns with the aggregate categories in Figures 2 and 3. Rather than delivering precise estimates or identifying specific interventions to invest in, this analysis as most useful for the aggregate patterns that it reinforces, such as the relative efficiency of input-alone interventions versus targeting instruction to children's level. Results for these aggregate categories of policies should help to inform prioritization by governments among specific policies for improving quality education.

Figure 6. Effective and Cost-Effective Interventions

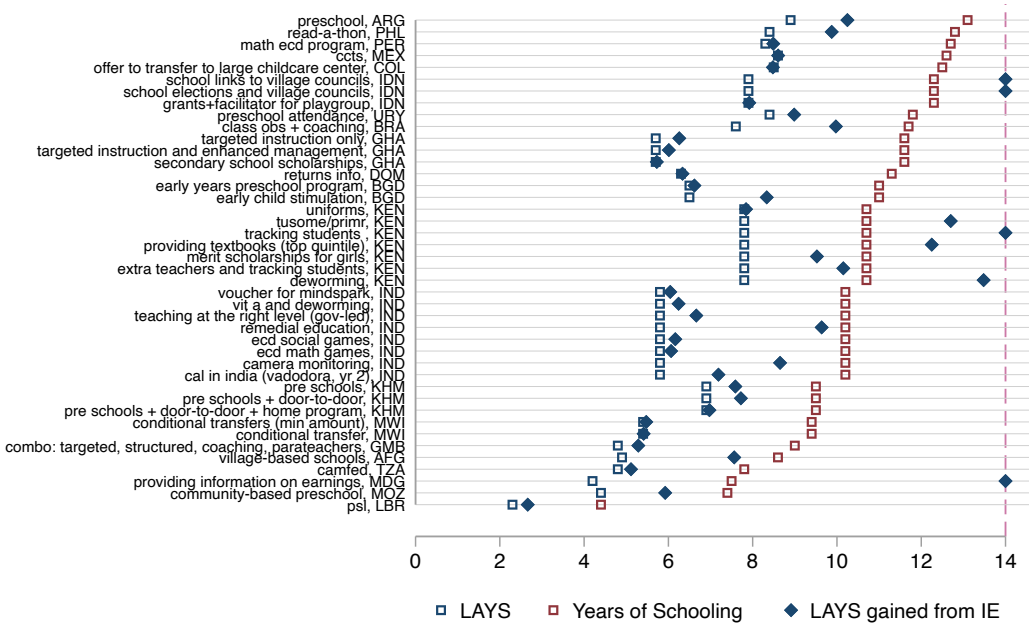


Notes: "High rank" means more effective or cost-effective, respectively.

4.2.2 Calibrating gains from specific interventions and policies to system-level gaps

What could the macro, systemwide effects of these highly effective interventions be? To answer this, we explore first how many LAYS a given intervention could contribute toward closing the gap to achieve a full and globally benchmarked quality education at scale in a given country—assuming, as mentioned before, that the nationally scaled-up version of the program were as effective as the evaluated version. Of course, this is rarely the case, and this exercise is meant as a calibration rather than as a simulation. An alternative approach would be to apply a "discount rate" to intervention effectiveness as they go to scale. In essence, in this exercise we map micro-LAYS onto macro-LAYS. Figure 7 takes cost-effectiveness into account, making the interventions comparable by showing the gap that each could close at a cost of $100 per child. It shows that highly cost-effective programs like Tusome and ability grouping could substantially narrow the learning gap separating the children in Kenya from their peers in higher-quality systems. Moreover, policies that improve the productivity of each year of schooling, such as targeting instruction to a child's learning level, can yield up to 4 additional LAYS per $100 in India. This calibration reveals that shifting the marginal dollar of government expenditure from low-efficiency education investments to high-efficiency educational investments could help countries make the most out of the years of education they are offering. While this cost analysis is useful for comparing interventions on a common scale, both the cost and the effectiveness of interventions can change at varying scales of implementation.

Figure 7. LAYS Gained per $100 per Intervention, Calibrated to Country-Level LAYS Gaps



Notes: This calibration assumes no loss of effectiveness once an intervention operates at national scale, which often is not the case. Alternative calibrations could apply a discount factor to account for weaker effects at scale. For the purposes of this excerise, which are designed only as a calibration of effect sizes, we provide a single estimate. We include years of schooling and learning-adjusted years of schooling (LAYS) from publicly available data used in the World Bank's Human Capital Index for each country. The LAYS gained from the impact evaluation (IE) indicates how much a given intervention or policy helps a country close its country-specific LAYS gap as well as bridge the global LAYS gap. The dashed red line at 14 years of schooling indicates the "distance to the frontier" as defined by the HCI as 14 years of high-quality schooling. Where the LAYS gained from the IE result in a LAYS estimate that exceeds the global benchmark of 14 of high quality schooling, we set the LAYS gained from IE estimate to the value needed to close the global LAYS gap fully.

5. Robustness

In this section, we present sensitivity analyses of our assumptions and parameter choices. We focus on four main areas: the high-quality learning benchmark, scaling of the learning assessments, standard deviations across tests and samples, and status-quo learning trajectories.

5.1 High-quality learning benchmark

We use 0.8σ as a benchmark for high-performing learning rates. As noted above, this value is an artificial high-performance benchmark chosen because it is stable (unlike benchmarks based on actual performance of leading countries) and non-political. This approach to defining high-quality learning *rates* is similar to the approach to defining the high-performance benchmark learning *level* in the World Bank Human Capital Index (Kraay

2019). We explore three approaches to validating this high-performance benchmark: (a) average annual learning trajectories in high-performance cases; (b) policy-relevant learning changes; and (c) rules of thumb and a range of effect sizes in reviews of multiple studies.

The first approach draws on high-performance learning trajectories. Although there is surprisingly little year-on-year raw data on learning, one notable example where there *is* longitudinal data is from the Young Lives survey. That survey follows students in India, Vietnam, Peru, and Ethiopia over 15 years and uses learning assessments based on Item-Response Theory (IRT). Using this data and a combination of value-added estimates, instrumental variables, and regression discontinuity methods, Singh (2020) finds that the causal effect of an additional year of primary school in Vietnam is 0.76σ, the largest value among the four countries. This is likely a lower bound for "high performance" on a global scale, since Vietnam—while an excellent performer for its income class—ranks in the second decile of average Harmonized Learning Outcomes (which, as noted above, covers 164 countries from 2000-2017). We can compare these results to an alternative high-benchmark year-on-year comparison: changes analyzed in the United States by Bloom, Hill, Black, and Lipsey (Bloom et al. 2008), building on methods used by Kane (2004). The largest year-on-year learning gains are between grade 1 and 2 and range from 0.97σ in reading to 1.03σ in math. Finally, we can derive approximate year-on-year changes for global high performers. We assume that the appropriate high-performance rescaled HLO benchmark is a score of 325 at the primary level. This score is assumed to be obtained over four years, since most primary international assessments occur in grade 4; average high-performance learning per year is thus 81.25 points. We then assume a within-country standard deviation of 85 points, based on the values for the five highest-performing countries using 2006 PISA microdata. Taking the ratio of these two values yields a year-on-year gain of 0.96σ.

The second approach examines large, system-level gains. Here, we explore what would constitute a large learning gain in systemic terms, as a way to benchmark what high-performing learning progress would look like. One example is to consider cross-country learning gaps in terms of HLO scores used for the World Bank Human Capital Index. A gain of 0.8σ would enable the United Kingdom or Vietnam to catch up to Singaporean learning levels: because the cross-country standard deviation is equivalent to 70 HLO points, so a 0.8σ gain for the United Kingdom (517) or Vietnam (519) translates into nearly closing the gap with Singapore (581). In another example, consider that the black-white achievement gap in the United States in math ranges from 0.99σ to 1.04σ in grades 4 and 8 (Bloom et al. 2008). A gain large enough to nearly close either of these gaps would be highly meaningful in policy terms.

The final approach uses rules of thumb. Cohen (1988) proposed the following standardized effect-size benchmarks: at least 0.20 for "small" effects, 0.50 as "medium" effects, and 0.80 for "large" effects. This framework has been broadly applied across interventions and contexts for decades. However, there is debate about the relevance of these indicators to education interventions, given that almost all interventions in high-, middle-, and low-income countries have much smaller impacts. For high-income countries, the 90th-percentile

effect size is 0.47 (Kraft 2020); for low- and middle-income countries, it is 0.38 (Evans and Yuan 2020). Both of those fall below the traditional Cohen benchmark for even medium effects.

In summary, these various approaches—particularly those focused on high-performance learning trajectories and meaningful systemic improvements—yield high-performance benchmark learning rates ranging from around 0.8σ to 1.0σ. In this paper, we use an artificial benchmark of 0.8σ for learning gains, which is a conservative high-performance benchmark consistent with this range.

5.2 Test-score scaling

Next, we explore sensitivities to score scales, comparing our results based on scores rescaled via a linear transformation of 300 points to the original HLO score scale. This enables us to use a scale that starts at zero, which has useful statistical properties. In Appendix Figure B2, we corroborate this de facto floor with data from EGRA, which shows that an HLO score of 300 corresponds roughly to zero percent reading comprehension.

Appendix Figure B3 compares the L_i^h value using the two score scales. While the scale that we use largely does not affect relative ranks, it does affect the degree of the absolute learning adjustment. Using the original scale (y-axis), the distance between Mexico and Ghana is 0.2; by contrast, under the re-scaled version (x-axis), the distance is closer to 0.5.

Appendix Figure B4 compares results using the two score scales. In Panel A, we see that the main effect of the rescaling is to reduce the micro-LAYS values that are based on participation impacts—for example, conditional cash transfers in Malawi. This is because under the original scale, the maximum learning adjustment discounted a year of school by about half, since the de facto floor of the HLO scale was 300, which produced a learning-adjustment factor, L_i^h, of 0.48 relative to the high-performance benchmark of 625. Under the rescaling, the minimum learning factor converges to zero, and the learning adjustments drop substantially, reducing participation-based LAYS estimates. As an example, the learning adjustment in Kenya shifts from an original L_i^h of 0.73 to 0.48, while countries on the lowest tail of distribution, such as Malawi, shift from a learning-adjustment of 0.57 to 0.18. The re-scaling does not affect the computation of learning-based micro-LAYS, since those values are derived relative to an artificial high-performance benchmark of 0.8σ. However, as an added sensitivity test, we can use the old scale to derive a new corresponding high-performance benchmark of 1.8σ. Panel B plots LAYS using this new benchmark.

These sensitivity tests show that the unscaled scores yield higher values for micro-LAYS based on participation estimates. Also, although unscaled scores largely preserve rank, they understate the degree of learning adjustment, especially for countries on the lower tail of the distribution. For these reasons, in the main results presented in this paper we use micro-LAYS based on re-scaled scores. Since the lowest-performing countries are already far

behind, re-scaling of scores is unlikely to yield major new insights and will not change ranks; however, we think re-scaling is important for capturing the full degree of the learning gap.

5.3 Standard deviations across tests and samples

We also test sensitivity of our results to differences in standard deviations across tests and samples. This is relevant because, for the interventions included in our comparisons, we use effect sizes expressed in standard deviations of each intervention's test and sample as inputs into LAYS. Note that this issue is not inherent to LAYS, which are just a unit of measurement that can be applied to any assessment results. In an ideal world (ideal from a comparability perspective), all interventions and national-level assessments would use the exact same test to measure learning progress, and therefore there would be no issue with LAYS conversion.[7] However, in practice studies and interventions vary widely in the test used, making comparability an issue. LAYS can currently be constructed only with the data as they are, so a separate effort is needed to produce comparable underlying learning assessments.[8] Once that effort bears fruit, it will enhance the comparability of LAYS.

For now, therefore, we rely on standardized effect sizes. Standardized effect sizes are used to account for differences across measurement scales and express those effects in relative terms. This should prove useful when comparing effect sizes in education across various assessments and scales. However, standard deviations will not account for whether a given test is either "too hard" or "too easy," causing floor or ceiling effects. We test for this possibility empirically by comparing standard deviations from tests on nationally representative samples, chosen to ensure that the same underlying population is represented. We focus on primary-level tests for countries that have participated in multiple tests and that have interventions featured in this paper. Appendix Figure B5 compares standard deviations for Tanzania, Malawi, and Indonesia using various assessments: HLO scores derived from EGRA, Progress in International Reading Literacy Study (PIRLS), raw EGRA or Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) tests. We find only small differences of a few points, and as a result, the estimated learning rates per year across assessments are quite similar, as shown in Appendix Figure B6. As a robustness

---

[7] Of course, that ideal test would need to be comprehensive and adaptive, so that it could be equally informative across all parts of the learning distribution. Furthermore, even if all impact evaluations used exactly the same test, it would be necessary to make judgments about the relative value of improvements on one part of the test vs another.

[8] Short of a universal test, a number of efforts currently underway may improve comparability. One approach focuses on a set of common test items that can be inserted into assessments and linked to regional and international assessments, such as TIMSS and PISA. The Jameel Poverty Action Lab has launched an initiative to produce an "item bank" of questions for inclusion in education interventions to produce comparable assessment. The Rosetta Stone initiative of the International Association for the Evaluation of Educational Achievement (IEA), the UNESCO Institute for Statistics (UIS), and other organizations are also working to improve comparability across regional and global assessments by having the same samples of students take both types of assessments.

check, Appendix Figure B6 also shows learning rates per year using raw data from the assessments before they are converted to HLO scores.

In terms of standard deviations across samples, it is well known there is high variation across sample populations. As a result, when students answer five questions correctly in a country with less variation, this will appear as a larger standardized effect size than if they had gotten the same questions correct in a country with more variation. This is a feature of standard deviation comparisons rather than a bug, since standardized effect sizes produce relative comparisons to derive a sense of magnitude. Therefore, while LAYS conceptually have the advantage that they do not need to use standard deviations as inputs, when in practice they do use standardized effect sizes as inputs (as in this paper), comparisons of relative performance and rank orders should be prioritized over precise magnitudes of absolute performance.

5.4 Status-quo learning

When we compute LAYS from learning estimates, we first convert learning into equivalent years of school gained. To do this, we express learning relative to status-quo learning trajectories. For example, if students learned 0.25σ per year in an intervention in South Africa, and students typically learn 0.25σ in a given year in South Africa, impacted students will have learned a year's worth of schooling in South Africa. A few options exist for possible status-quo learning trajectories: for example, we might use the national average learning trajectory, or the learning gains in the control group of the same evaluation from which effect sizes are drawn. Alternatively, we could use an average learning profile across all countries being compared. This choice of status quo will affect our interpretation. If we choose the experimental or quasi-experimental control group, then the resulting figure for equivalent years of schooling gained is relevant to the study sample only. If we choose national average learning trajectories, we can interpret the figure as the number of equivalent years of schooling gained at the national level, with the embedded assumption that the standard deviations are comparable. In the main analysis reported above, we use national-level learning trajectories; in this section, we explore the alternative of using the study's control group to measure status-quo learning.

An advantage of using control-group status-quo learning is that this estimate is drawn from the same sample as the learning gains from the intervention are, so the two are directly comparable. If the study sample is not representative of the nation, however, then when we later apply a national-level learning adjustment to compute LAYS in globally comparable terms, the adjustment will be less reliable. We can test the assumption of representativeness of study samples by examining the degree of variation within a country. If variation within a country is large, this means that any given study sample is likely to diverge from the national average. We test this assumption using a uniform test, EGRA. Two advantages of EGRA data are that (1) EGRAs are included in the World Bank Harmonized Learning Outcomes database and (2) they are often used to assess the impact of interventions and policies. In Appendix Figure B7, we compare variation *within* a country to variation *across* countries as a

benchmark of whether variation within a country is large. We find that for a sample of 39 countries for which we have EGRA HLO scores, the average cross-country standard deviation is 53, whereas the within-country standard deviation is often higher than 53, with a density skewed to the right tail. This finding indicates that within-country variation is often quite large, casting doubt on the assumption that a sample would necessarily represents the nation. As a specific example, control-group status-quo gains in India in our sample of studies range from 0.5σ to 0.76σ, implying high variance from study to study.

In summary, we find that the assumptions for using control-group learning trajectories as our measure of the typical status-quo learning in a country are unlikely to be robust. We instead rely on national-level learning trajectories, which are easily interpretable and can be converted to a global metric. An additional advantage of using national-level learning trajectories is a practical one: greater data availability. Whereas control-group information is often missing from published papers, national learning trajectories can be calculated using HLO scores, which are available for 164 countries.

6. Conclusion

In this paper, we analyze which investments most efficiently improve education outcomes. We expand on previous education reviews and analyze 150 interventions and policies across 46 countries using a new unified education measure: learning-adjusted years of schooling. A central insight from this analysis is that many interventions that increase participation in schooling are less cost-effective than interventions that improve the productivity of schooling— that is, the amount of actual learning gained. Policies that improve the productivity of each year of schooling, such as targeting instruction to a child's learning level or improving pedagogy through structured lessons plans and coaching, can yield large gains in LAYS, narrowing the gap between high- and low-preforming education systems globally. These results should be interpreted with context in mind: challenges should be identified locally, and the global evidence base should then be used to identify possible cost-effective solutions, which should then be carefully adapted to the local context.

This analysis further builds the foundation for the use of LAYS as a common metric for the economic evaluation of education interventions. Similar unified metrics have played important roles in public health, macroeconomics, and economic welfare analysis, but to date no reference standard exists for education cost-effectiveness analysis, and approaches to comparative analysis have been ad hoc. Using micro-LAYS to express impact sizes achieves three goals: (a) it places attainment and learning outcomes on a unified scale, allowing interventions to be compared directly; (b) it expresses educational outcomes in terms of an easy-to-interpret measure that improves incentives for policymakers to promote both quantity and quality of schooling; and (c) it identifies levers for countries to close gaps between their current performance and the full years of high-quality schooling that they aspire to. Recent research suggests that policymakers may not reap political benefits from learning gains alone (Habyarimana, Opalo, and Schipper 2020), yet an additional year of

schooling can lead to very different levels of learning (World Bank 2018; Singh 2020). Using LAYS as a metric of progress allows a focus on additional years and learning together.

In summary, this paper uses learning-adjusted years of schooling to provide guidance on which policies and interventions are the most efficient investment in education, given the state of evidence and data available today. Although only recently introduced, LAYS has been incorporated into large-scale policy efforts to improve education. It is a component of the World Bank's recently launched Human Capital Index (World Bank 2019), and it has been used by the World Bank and United Kingdom's Foreign, Commonwealth & Development Office (FCDO) to prioritize cost-effective education investments. With this research, our goal is to provide a useful tool for other decisionmakers who are seeking to address learning and access gaps.

References and Notes

Angrist, N., S. Djankov, P. K. Goldberg, H. A. Patrinos, "Measuring Human Capital" (WPS8742, The World Bank, 2019), pp. 1–46.

Azevedo, Joao Pedro, and others. 2019. "Will Every Child Be Able to Read by 2030? Why Eliminating Learning Poverty Will Be Harder Than You Think, and What to Do About It." World Bank Working Paper. Washington, DC: World Bank

ASER Centre, "Annual Status of Education Report" (India, 2017), (available at http://www.asercentre.org/Keywords/p/315.html).

Baird S., C. McIntosh, B. Özler, Cash or Condition? Evidence from a Cash Transfer Experiment. *Q J Econ*. **126**, 1709–1753 (2011).

Bashir, Sajitha, Marlaine Lockheed, Elizabeth Ninan, and Jee-Peng Tan. *Facing forward: Schooling for learning in Africa*. The World Bank, 2018.

Bates, M.A., R. Glennerster, The Generalizability Puzzle. *Stanford Social Innovation Review*. Summer 2017.

Bloom, H. S., C. J. Hill, A. R. Black, M. W. Lipsey, Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions. *Journal of Research on Educational Effectiveness*. **1**, 289–328 (2008).

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

Crawfurd L., D.K. Evans, S. Hares, L. Moscoviz, "12 Years of Quality Education for Every Girl: Five Ways the New UK Government Can Deliver on Its Manifesto Pledge." Center for Global Development (2020).

De Neve, J.-W., G. Fink, S. V. Subramanian, S. Moyo, J. Bor, Length of secondary schooling and risk of HIV infection in Botswana: evidence from a natural experiment. *The Lancet Global Health*. **3**, e470–e477 (2015).

Duflo E., P. Dupas, M. Kremer, The Impact of Free Secondary Education: Experimental Evidence from Ghana. *Massachusetts Institute of Technology Working Paper Cambridge* (2017).

Drummond, M. F., M. J. Sculpher, K. Claxton, G. L. Stoddart, G. W. Torrance, *Methods for the Economic Evaluation of Health Care Programmes* (Oxford University Press, Oxford, New York, Fourth Edition., 2015).

Dynarski, S. "Hope for whom? Financial aid for the middle class and its impact on college attendance." *National Tax Journal* (2000): 629-62

Eble, A., C. Frost, A. Camara, B. Bouy, M. Bah, M. Sivaraman, J. Hsieh et al. "How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in The Gambia." *Journal of Development Economics* (2020): 102539.

Evans, D. K., A. Popova, What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews. *World Bank Res Obs.* **31**, 242–270 (2016a).

Evans, D. K., A. Popova, Cost-Effectiveness Analysis in Development: Accounting for Local Costs and Noisy Impacts. *World Development*. **77**, 262-276 (2016b).

Evans, D. K., F. Yuan, "How Big Are Effect Sizes in International Education Studies?" (CGD Working Paper 545, 2020).

Evans, D. K., F. Yuan, "What We Learn about Girls' Education from Interventions that Do Not Focus on Girls" (WPS8944, The World Bank, 2019), pp. 1–45

Evans, D. K., F. Yuan, "Equivalent Years of Schooling: A Metric to Communicate Learning Gains in Concrete Terms" (WPS8752, The World Bank, 2019), pp. 1–52.

Filmer, D. P., F. H. Rogers, N. Angrist, S. Sabarwal, "Learning-Adjusted Years of Schooling (LAYS) : Defining A New Macro Measure of Education." *Economics of Education Review*. 77 (2020).

Fiszbein A., N. Schady, F.H.G. Ferreira, M. Grosh, N. Keleher, P. Olinto, E. Skoufias, Conditional Cash Transfers : Reducing Present and Future Poverty. World Bank Policy Research Report (2009).

Ganimian, A. J., R. J. Murnane, Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations. *Review of Educational Research* (2016), doi:10.3102/0034654315627499.

Glewwe, P., E. A. Hanushek, S. Humpage, R. Ravina, "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010," in *Education Policy in Developing Countries* (University of Chicago Press, 2013; https://www.universitypressscholarship.com/view/10.7208/chicago/9780226078854.001.0001/upso-9780226078687-chapter-2).

Habyarimana, J., K. Ochieng'Opalo, Y. Schipper. "The Cyclical Electoral Impacts of Programmatic Policies: Evidence From Education Reforms in Tanzania." (2020).

Hanushek E. A., L. Woessmann, Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *J Econ Growth*. 17, 267–321 (2012).

Haushofer J., J. Shapiro, The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya. *Quarterly Journal of Economics*. 131, 1973-2042 (2016).

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, A. Yavitz. "The rate of return to the HighScope Perry Preschool program." *Journal of Public Economics*, (2010): 114128

Hendren, N., B. Sprung-Keyser. "A unified welfare analysis of government policies." *The Quarterly Journal of Economics* 135, no. 3 (2020): 1209-1318.

Johnson, R. C., and C. K. Jackson. "Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending." *American Economic Journal: Economic Policy* 11, no. 4 (2019): 310-49.

Kane, T. J. "College entry by blacks since 1970: The role of college costs, family background, and the returns to education." *Journal of Political Economy* (1994): 878911

Kane, T. J. "The Impact of After-School Programs; Interpreting the Results of Four Recent Evaluations" (William T. Grant Foundation, 2004), (available at https://rhyclearinghouse.acf.hhs.gov/library/2004/impact-after-school-programs-interpreting-results-four-recent-evaluations).

Kraay, A. The World Bank Human Capital Index: A Guide. *World Bank Research Observer*. 34, 1–33 (2019).

Kraft, M. A. "Interpreting effect sizes of education interventions." *Educational Researcher* 49, no. 4 (2020): 241-253.

Kremer, M., C. Brannen, R. Glennerster, The Challenge of Education and Learning in the Developing World. *Science*. 340, 297–300 (2013).

Krishnaratne, S. , H. White, E. Carpenter, "Quality education for all children? What works in education in developing countries | 3ie" (Working Paper 20: 155, New Delhi: International Initiative for Impact Evaluation (3ie), 2013), (available at https://www.3ieimpact.org/evidence-hub/publications/working-papers/quality-education-all-children-what-works-education).

McKeever, V., "Malala Yousafzai completes her Oxford degree, says now it's time for 'Netflix, reading and sleep'." CNBC (2020). https://www.cnbc.com/2020/06/19/malala-yousafzai-completes-her-oxford-degree.html. (Accessed October 13, 2020.)

Murray, C. J. L. , A. D. Lopez, Evidence-Based Health Policy—Lessons from the Global Burden of Disease Study. *Science*. **274**, 740–743 (1996).

Piper, B. , J. Destefano, E. M. Kinyanjui, S. Ong'ele, Scaling up successfully: Lessons from Kenya's Tusome national literacy program. *J Educ Change*. 19, 293–321 (2018).

Pliskin, J. S. , D. S. Shepard, M. C. Weinstein, Utility Functions for Life Years and Health Status. *Operations Research*. **28**, 206–224 (1980).

Romero, M., J. Sandefur, W. A. Sandholtz, Outsourcing Education: Experimental Evidence from Liberia. *American Economic Review* (forthcoming), doi:10.1257/aer.20181478.

Sabates, R., P. Rose, M. Delprato, B. Alcott, "Cost-Effectiveness With Equity: Raising Learning For Marginalised Girls Through Camfed'S Programme In Tanzania," *Policy Paper No. 18/2* (REAL Centre, University of Cambridge, 2018), doi:10.5281/ZENODO.1247315.

Singh A., Learning More with Every Year: School Year Productivity and International Learning Divergence. *Journal of the European Economic Association* (2020), doi:10.1093/jeea/jvz033.

Torrance, G. W., W. H. Thomas, D. L. Sackett, A Utility Maximization Model for Evaluation of Health Care Programs. *Health Serv Res.* **7**, 118–133 (1972).

World Bank, "World Development Report 2018: Learning to Realize Education's Promise" (Washington D.C., 2018).

World Bank, "World Development Report 2019: The Changing Nature of Work" (Washington, DC, 2019).

World Bank. 2020. "The Human Capital Index 2020 Update : Human Capital in the Time of COVID-19." World Bank, Washington, DC

Zeckhauser, R., D. Shepard, Where Now for Saving Lives? *Law and Contemporary Problems*. 40, 5–45 (1976).

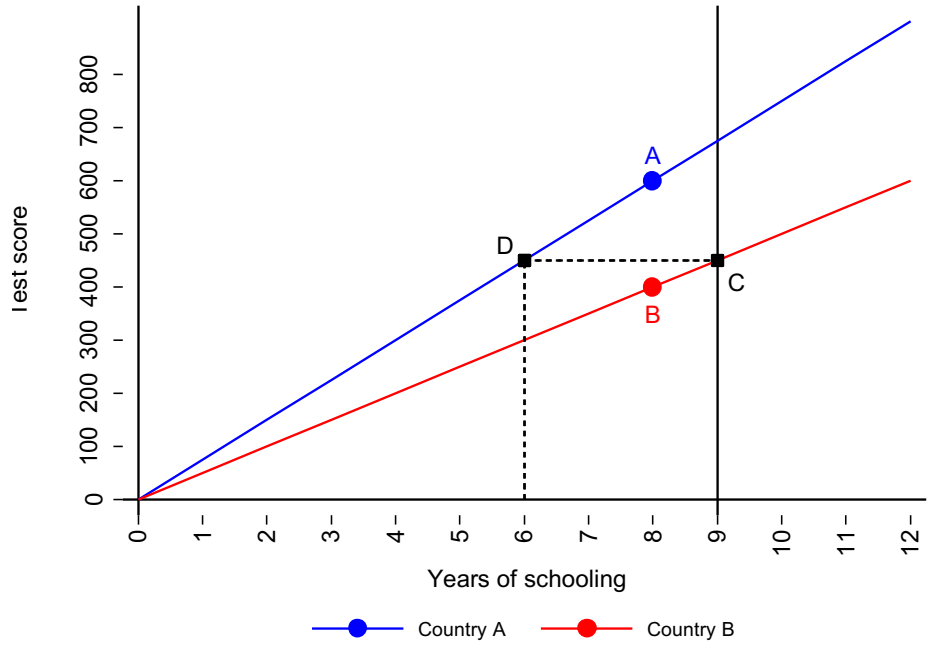
Appendix A. The Assumption of Constant Average Learning Trajectories

The assumptions invoked in the construction of macro-LAYS are explored in depth in Filmer et al. (2020). Here, we highlight one assumption: constant average learning trajectories, or the idea that students learn the same amount with each additional year of schooling. Figure A1 demonstrates the utility of this assumption using a hypothetical example. Assume that we observe Grade 8 test scores of 600 for Country A and 400 for Country B and that individuals in Country B average 9 years of schooling. LAYS allow us to "convert" the 9 years of schooling in Country B into the number of years of schooling in Country A that would have produced the same level of learning. Moving along the average learning profile from Grade 8 allows us to infer what Country B's average score would be if its students were tested in Grade 9. This calculation is represented by the move from point B to point C, or from a test score of 400 to 450. The next step is to go from point C to point D, to find the number of years of schooling that it would take in Country A to produce that level of learning (450) given the average learning profile in Country A. In this example, it takes 6 years, so the resulting LAYS measure in Country B is 6. Both steps of the calculations rely on the linearity assumption, because we do not have data on the actual learning trajectories but rather on learning at one point in time for each country.

How realistic is this assumption? Filmer et al. (2020) explore this question with a series of empirical tests on whether learning trajectories are constant on a locally defined interval. Figure A2 showcases one example using data from India's Annual Status of Education Report (ASER), which administers the test consisting of the same questions to students from ages 5 to 16, covering Grades 1 to 12 (ASER 2017). The ASER data enable us to assess the rate of learning with a stable, comparable metric across grades and over time. To allow us to map out the specific trajectory for learning in school, we restrict our sample to school-going children.[9] In the case of a mathematical skill, division, Figure A2 shows that students learn along an "S-shaped" learning trajectory, but with a locally linear interval from Grades 5 to 10. Other, more complex skills than division are likely to have a linear learning trajectory across an even wider interval because they cannot be mastered so quickly. The other empirical tests in Filmer et al. (2020) also yield results consistent with the linearity assumption (at least over a significant local interval).

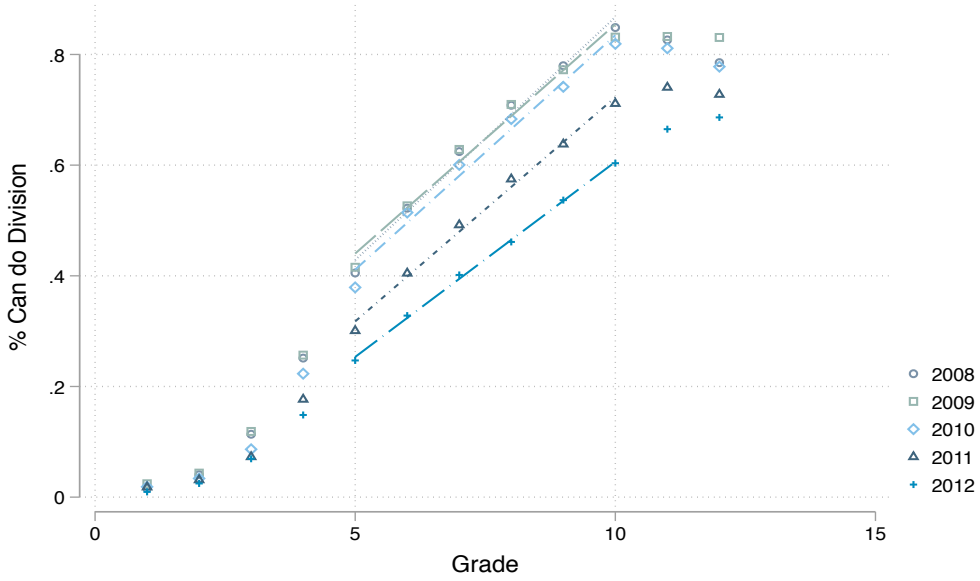
[9] This comparison is conducted across different cohorts of students at different grades.

Figure A1. Constant Learning Trajectories



Note: We map hypothetical learning trajectories in countries A and B to demonstrate the utility of the assumption of constant learning trajectories.

Figure A2. Learning Trajectories in India

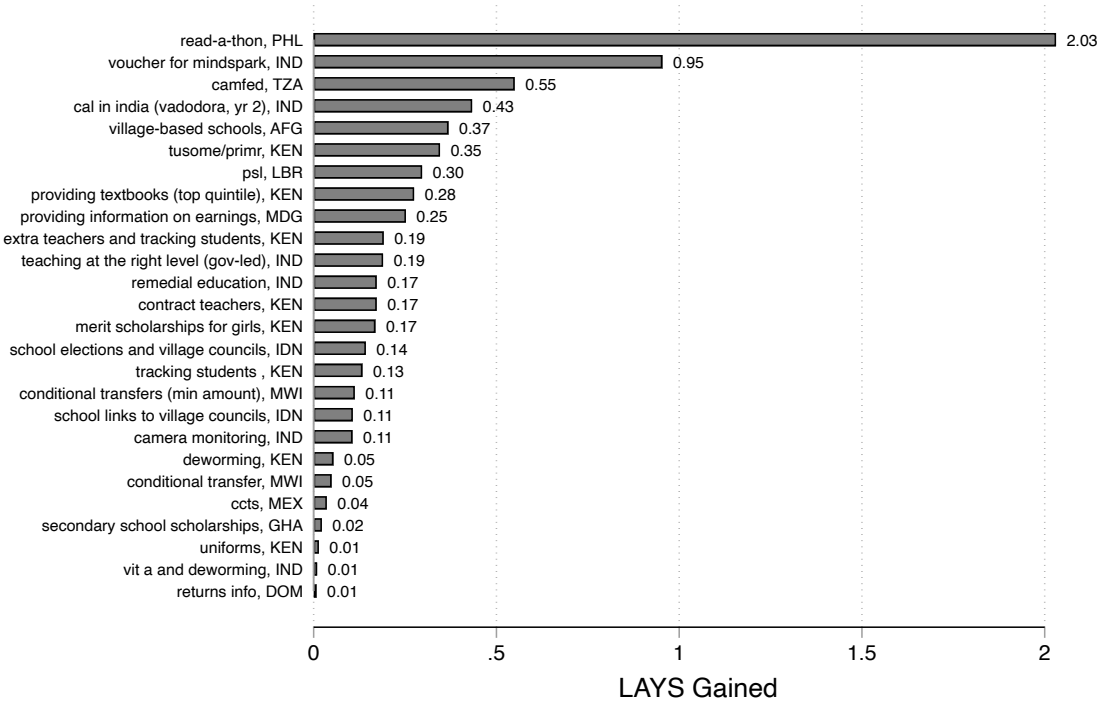


Note: We derive learning trajectories using empirical data from a national survey conducted in households in India for students aged 5 to 16 in grades 1 through 12. We include only students at the household who are in school.

Source: ASER India data from 2008 to 2012 as analyzed by Filmer et al. (2020).

Appendix B. Additional Figures

Figure B1. Expressing LAYS gained per year ($t = 1$)



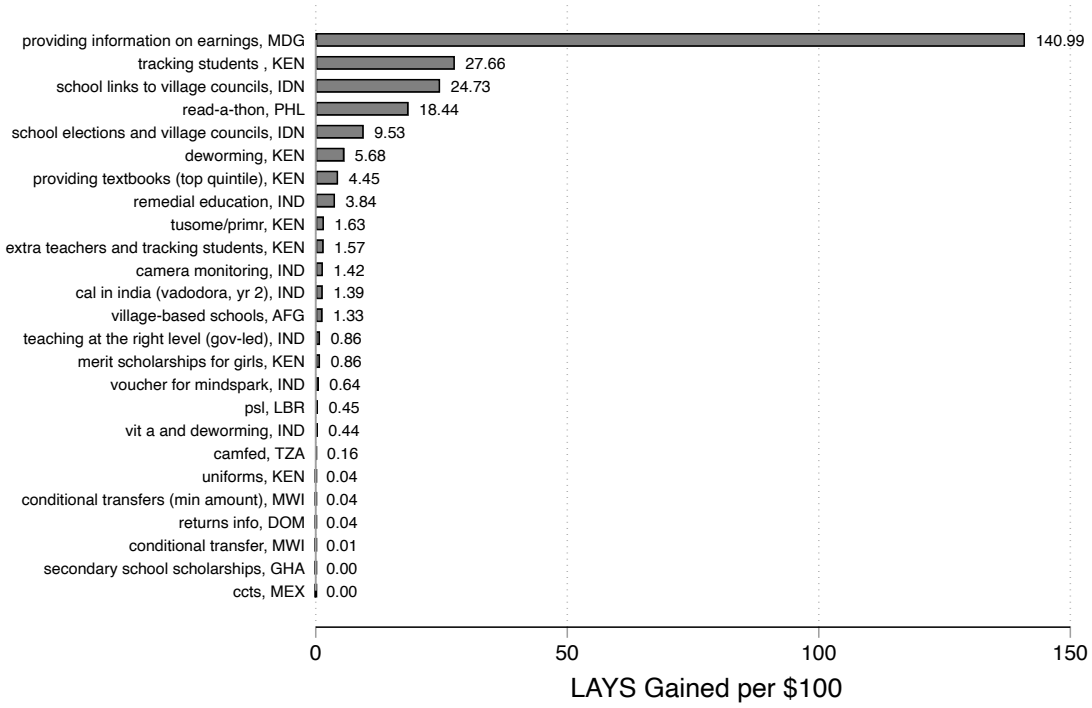
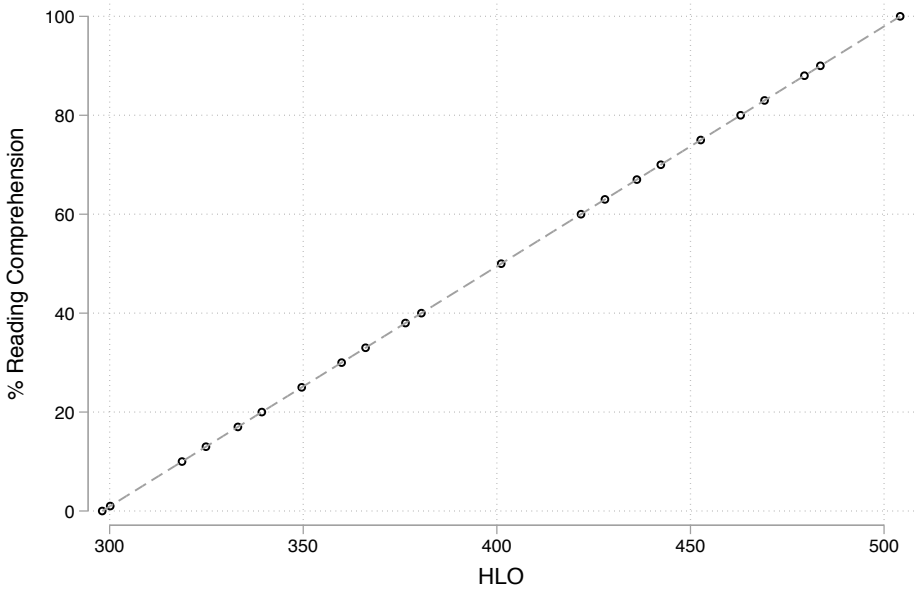
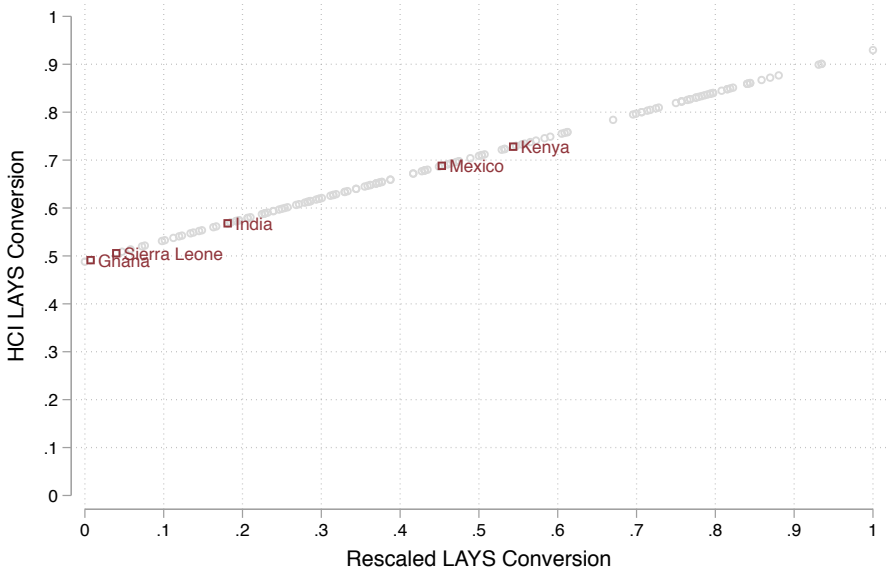


Figure B2. EGRA raw reading comprehension relative to HLO score



Notes: We analyze EGRA data across 39 countries and match raw score on reading comprehension modules with the Harmonized Learning Outcome (HLO) scores used for the World Bank Human Capital Index.

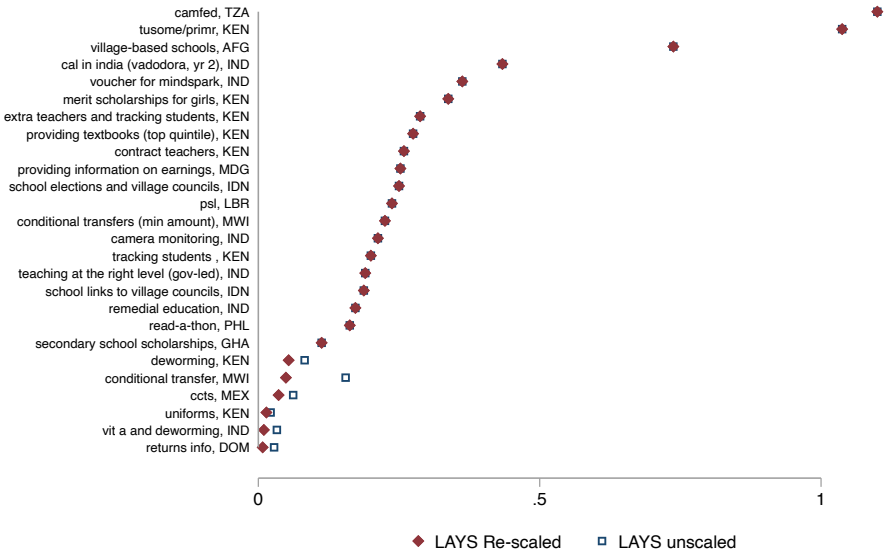
Figure B3. Learning Adjustment Rates



Notes: We rescale LAYS conversion rates. Initial conversion rates are based on scores which often floor around 300 due to underling test scores scales. Since the LAYS conversion rate is calculated out of 625, this produces a floor conversion rate of .48. However, when learning levels are very low this conversion will under-adjust learning. We rescale LAYS exchange rates to range from 0 to 1.

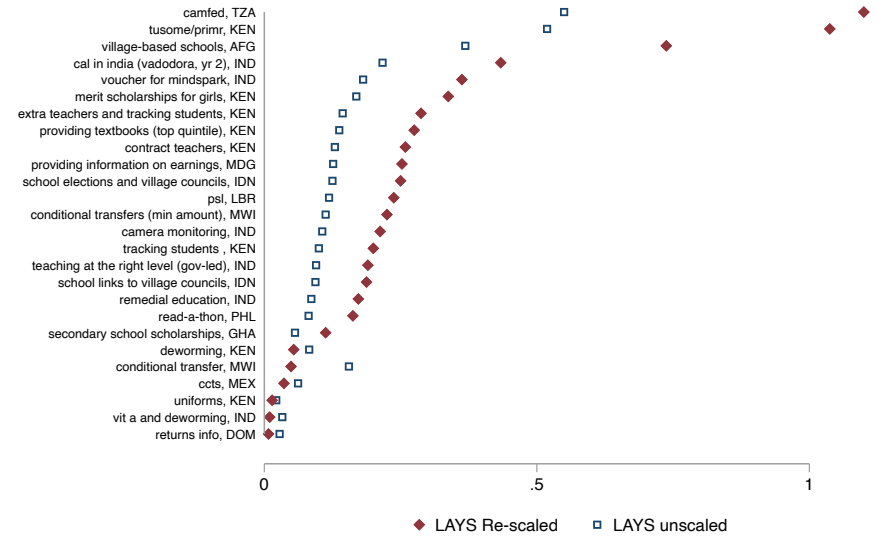
Figure B4. Comparing LAYS using scaled and unscaled test scores

Panel A



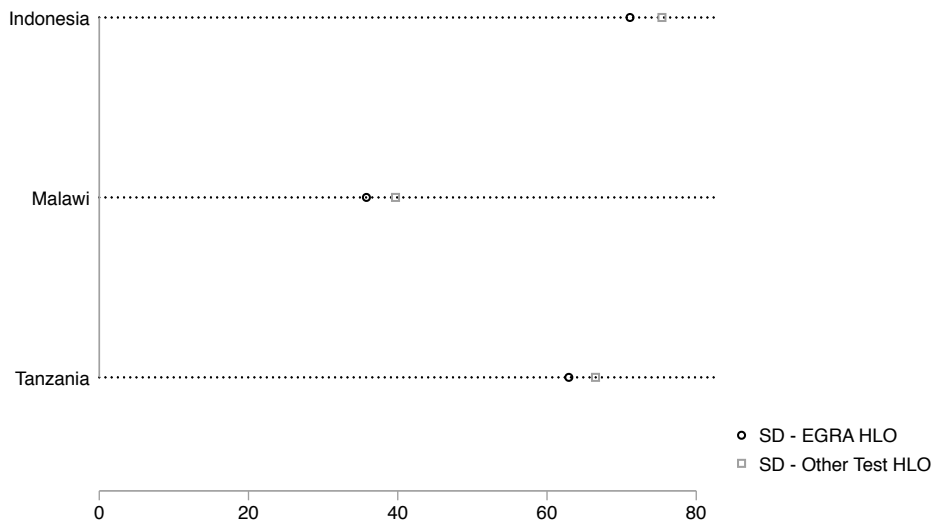
Notes: We rescale LAYS conversion rates for participation-based estimates to adjust for a quality factor that accounts for the new HLO scale. We do not adjust learning-based estimates in this figure since they are not adjusted by a learning factor. Rather, they are based on an assumption of high-benchmark learning trajectories of .8σ.

Panel B



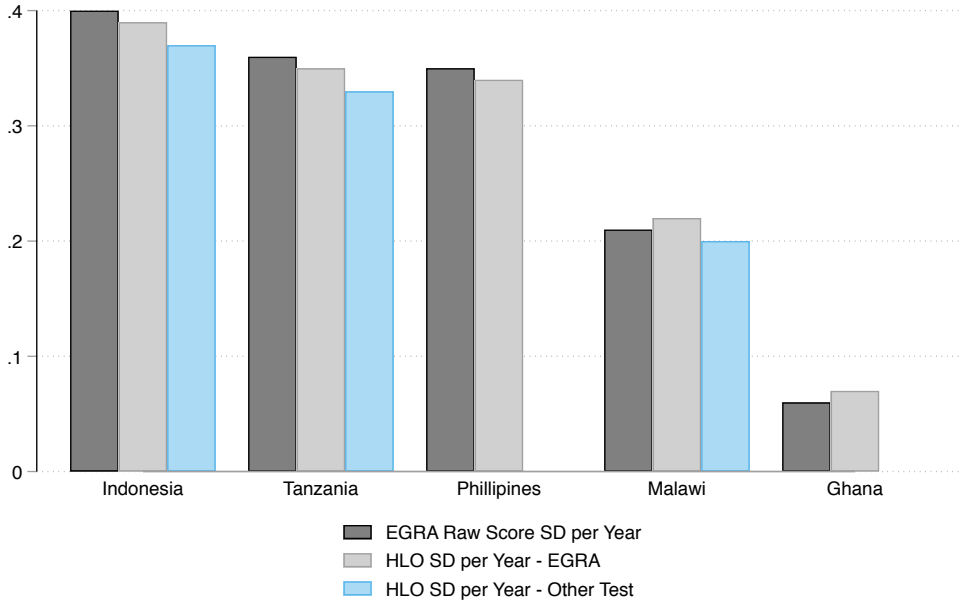
Notes: We rescale LAYS conversion rates and apply this rescaling to participation-based estimates to adjust for a quality-adjustment factor that accounts for the new HLO scale. In this figure, as a robustness test we adjust learning-based estimates by deriving a new high-benchmark learning trajectory based on a new scale, which yields 1.8σ.

Figure B5. SD Comparisons, by Source Test



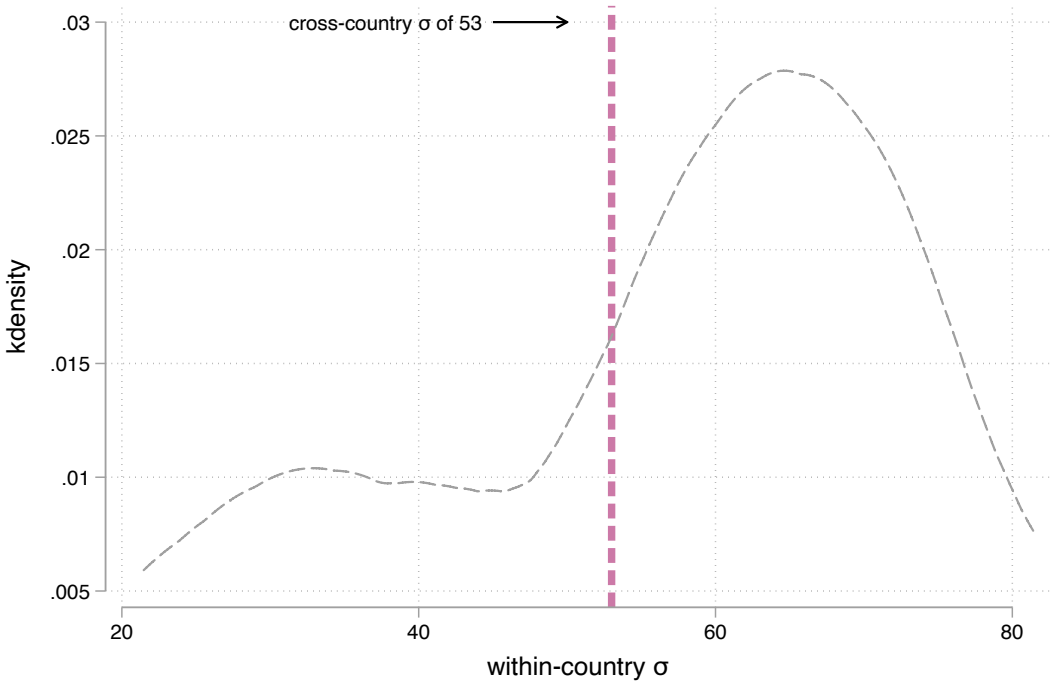
Notes: For Indonesia, the "other test" is PIRLS 2011; for Tanzania and Malawi it is SACMEQ 2007.

Figure B6. Learning Per Year (in SD), by Source Test



Notes: For Indonesia, the "other test" is PIRLS 2011; for Tanzania and Malawi it is SACMEQ 2007. We assume all scores were obtained in Grade 4 as a placeholder for primary school scores.

Figure B7. Within- vs. Cross-Country Variation in Test Scores



Notes: We use micro EGRA data across 39 countries and include country-year observations. The x-axis represents within-country variation. The vertical line represents the cross-country standard deviation: 53 for the cross-country variation of this EGRA as a benchmark. Variation is often greater within country than across countries, with most-within-country SDs falling to the right of the vertical line.