What Works Hub
for Global Education

What Works Hub for Global Education Working Paper

How to improve education outcomes most efficiently?

Noam Angrist, David K. Evans, Deon Filmer, Rachel Glennerster, Halsey Rogers, Shwetlena Sabarwa

BLAVATNIK SCHOOL OF GOVERNMENT
UNIVERSITY OF OXFORD
UK International Development
Partnership | Progress | Prosperity

How to Improve Education Outcomes Most Efficiently? A Review of the Evidence Using a Unified Metric

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

Abstract

Many low- and middle-income countries lag far behind high-income countries in educational access and student learning. Policymakers must make tough choices about which investments to make to improve education with limited resources. Although hundreds of education interventions have been rigorously evaluated, comparing their impacts is challenging. This paper provides the most recent and comprehensive review of the literature on effective education programs, with a novel emphasis on cost-effectiveness, covering the effectiveness and cost-effectiveness of interventions from over 200 impact evaluations across 52 countries. The analysis uses a unified measure –learning-adjusted years of schooling (LAYS) – that combines access and quality and compares gains to an absolute, cross-country standard. The results identify programs and policies that can be orders of magnitude more cost-effective than business-as-usual approaches, enabling policymakers to improve education outcomes substantially more efficiently.

JEL Codes: H43, H520, I2

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

*Angrist: Blavatnik School of Government, University of Oxford; Youth Impact. Evans: Center for Global Development; Filmer: The World Bank; Glennerster: University of Chicago; Rogers: The World Bank; Sabarwal: The World Bank. The authors are grateful to Abelardo De Anda Casas who provided excellent research assistance. Amina Mendez Acosta also provided valuable research assistance and 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. We also thank Emily Cupito, Shashank Patil, and Sharad Hotha from the University of Chicago Development Innovation Lab, who collaborated with the research team to expand the database of impact evaluations. The views expressed here are those of the authors and should not be attributed to their respective institutions. This work has been supported by the What Works Hub for Global Education (jointly funded by the United Kingdom's Foreign, Commonwealth & Development Office (FCDO) and the Bill & Melinda Gates Foundation) as well as 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. Data availability statement: the data used in this article will be made available.

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] Moreover, by the age of 10, 90 percent of children in low-income countries still cannot read and understand a short, age-appropriate passage (Azevedo et al., 2021). With limited resources, policymakers must make tough choices about what to invest in to improve education outcomes, ranging from constructing schools to improving school management to deploying new educational software. Making these investment decisions requires comparable data on both the benefits and costs of alternative interventions.

However, the impacts of educational interventions are often reported in ways that make these comparisons difficult. For example, 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. Yet the returns to education are often a function of both education quantity and quality (Psacharopoulos & Patrinos, 2004). And policymakers want a combination of the two. Politicians and advocates have called for an increase in the number of "years of quality education," a single concept that incorporates both quality and quantity dimensions (Crawfurd, Evans, Hares, & Moscoviz, 2020; McKeever, 2020). There is evidence that some of the benefits of education, including economic growth, are more closely associated with learning (Angrist, Djankov, Goldberg, & Patrinos, 2021; Hanushek & Woessmann, 2012), whereas others are associated with years of schooling (Baird, McIntosh, & Özler, 2011; De Neve, Fink, Subramanian, Moyo, & Bor, 2015; Duflo, Dupas, & Kremer, 2021). 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.

In this paper, we analyze over 200 educational policies and interventions across 52 countries, identifying the most efficient approaches to improve education outcomes using a unified education measure—Learning-Adjusted Years of Schooling (LAYS)—that combines improvements in both access and quality. By doing so, we make it possible to compare the effectiveness of a broad range of education interventions using a concrete and policy-salient metric.[2] For a subset of interventions for which cost data are available, we include cost-effectiveness analysis and comparisons, which is critical to assessing the most efficient policies to invest in.

We find that while many interventions are not cost-effective, some of the most cost-effective interventions can deliver the equivalent of over three years of high-quality education (i.e., three years of learning in a high-performing country such as Singapore) for as little as $100 per child.

[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.

[2] In previous work, the concept of LAYS has been used to analyze country-level aggregate performance (Filmer, Rogers, Angrist, & Sabarwal, 2020), which we refer to in this study as "macro-LAYS." In this paper, we adapt the LAYS concept to analyze specific intervention and policy treatment effects, which we refer to as "micro-LAYS."

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. Some of the most consistently cost-effective approaches include: interventions to target teaching instruction by learning level rather than grade (e.g., "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, for example, targeted instruction yields up to 3 to 4 additional learning-adjusted years of schooling per $100—a gain equivalent to the entire system-level education gap between India and Argentina.[3] In contrast, other interventions such as 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.

Another striking result from our analysis is that many interventions that increase participation in schooling are often 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; while they have yielded gains in schooling in systems with low-quality education, they have often done so without improving learning, all at a 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. Rather, these results suggest that if the goal is to improve learning outcomes, interventions like cash transfers that succeed in getting children into school should be complemented with policies that improve the learning productivity of schooling.

This work contributes to three major literatures. We contribute to the literature synthesizing results from rigorous impact evaluations in education. Previous reviews of educational interventions in low- and middle-income countries include Glewwe, Hanushek, Humpage, and Ravina (2011), Kremer, Brannen, and Glennerster (2013), Krishnaratne, White, and Carpenter (2013), Snilstveit et al. (2015), Evans and Popova (2016b), Glewwe and Muralidharan (2016), and Ganimian and Murnane (2016). Our study updates the literature with the most recent and comprehensive set of evaluations and provides cost analysis for many more studies than previous work covered. Nearly

[3]This calibration does not imply that interventions 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.

all prior reviews are over a decade old. In the years since these reviews, there have been hundreds of additional impact evaluations in education, necessitating an updated review of the literature.

Moreover, gains in prior reviews are often reported in standard deviations rather than against an absolute benchmark. 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. 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 the same 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.

A concrete example demonstrates the value of LAYS over standard deviation gains, the current standard unit used for comparative analysis in education. For example, we find that deworming in Kenya yields 0.113 years of schooling, but only 0.018 standard deviation gains, while conditional cash transfers (CCTs) in Mexico yield 0.09 years of schooling, and 0.143 standard deviation gains. Thus, standard deviations flip the rank-order among these interventions and make it appear as if CCTs are over five times more effective as an artifact of local variation. While expressing results in terms of standard deviations is typical, this reveals how they can be potentially misleading.

We find that expressing outcomes in terms of LAYS yields substantive insight and new understanding relative to standard deviations. In the same example, deworming in Kenya yields 0.054 LAYS relative to 0.036 LAYS for CCTs in Mexico. This preserves the original ranking of years of schooling gains and reflects education gains in clear, transparent, and absolute terms. LAYS further incorporates the quality dimension, with deworming in Kenya pulling a bit further ahead relative to cash transfers in Mexico, since each year of schooling in Kenya produces more learning than in Mexico. The quality adjustment is even more dramatic when deworming in Kenya is compared to to CCTs in Malawi, where learning levels are far worse. Thus, using LAYS, we see that deworming interventions enhance education slightly more than CCTs (and at substantially lower cost). Had we used standard deviations, we would have drawn the opposite conclusion about which interventions were more effective. A full exposition of these examples is provided in section 2.3. By adjusting for quality and reducing the influence of local variation, using LAYS allows us to say something about how effective education interventions are against an absolute, cross-country standard using a unified education measure.

Another challenge addressed by this paper is that current metrics used in the literature 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. To this end, our analysis takes this literature a step further by using a metric (LAYS) that enables unified comparisons of studies across access and learning in education, increases comparability of results across studies, and provides clear interpretation of the results in concrete policy terms.

The second literature we contribute to 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 (Pliskin, Shepard, & Weinstein, 1980; Torrance, Thomas, & Sackett, 1972; Zeckhauser & Shepard, 1976). While DALYs rely on many assumptions, today they are used widely as the reference standard in cost-effectiveness analysis (Drummond, Sculpher, Claxton, Stoddart, & Torrance, 2015; Murray & Lopez, 1996). In economics, summary measures such as the Multi-dimensional Poverty Index (MPI) (Alkire & Foster, 2011) have enabled researchers to understand poverty as a function of multiple measures, rather than focusing exclusively on income. Our work introduces a summary measure for impact measurement in education.

By setting out the benefits of using LAYS, we hope to encourage more researchers to express their results in common metrics to facilitate comparative analysis. Moreover, by providing a unifying framework with transparent assumptions, we hope to encourage researchers to make greater use of standardized learning assessments, which will in turn facilitate more meaningful comparisons across studies. To this end, introducing a common framework – even with imperfect data in the first instance – allows us to make the best use of available data as well as set into motion a cycle of ever more comparable data and comparisons in education over time. This evolution of metrics and measurement mimics the progression of DALYs and QALYs in the health sector, which started with a framework, assumptions, and a first analysis; over time, the data inputs improved, enhancing the comparability of each underlying study as well as facilitating cross-study comparisons. Even in this first analysis using existing data, our results are broadly 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, and different distributions of performance within samples and across countries. Moreover, for a subsample of studies with identical test items, we find similar results.

Finally, we relate to a literature attempting to inform government intervention through the use of cost-effectiveness and cost-benefit analysis across a broad range of potential government interventions. Much of this literature conducts cost-effectiveness analyses, but in different ways. For example, higher education analyses typically report the cost per enrollment (Dynarski, 2000; Kane, 2004), and early childhood education studies often report a social benefit-cost ratio (Heckman, Moon, Pinto, Savelyev, & Yavitz, 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 investment by governments in low-income children's health and education in the United States has historically had the highest return on investment, with many such policies paying 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.[4]

This work, like other syntheses and summary measures, has limitations. First, while this is the most comprehensive review of the education literature in low- and middle-income countries to date and covers hundreds of studies, available data are still limited, especially on costs, and 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. Our framework is general, allowing use of various learning measurements and units, including standard deviations as well as other metrics. Results are most comparable when comparing studies which use common tests and test items, which we anticipate will become increasingly common, in part motivated by the framework set out in this paper. For now, when including standard deviation units, we use assumptions about the distribution of learning levels in the study area to translate the study findings into LAYS. Third, because both impacts and costs are measured with imprecision (Evans & Popova, 2016a), it would be unwise to focus on small differences in cost-effectiveness. Rather, this analysis aims to inform broad trade-offs in cases where there are large, consistent differences. For example, we consistently see that as a cost-effective tool for improving LAYS, investing in inputs alone (without complementary reforms) ranks 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, such as 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 should consider whether contextual conditions support local adaptation of an intervention (Bates & 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.

[4] It is important to note that while many of the studies we review evaluate individual policies, some of the most effective policies we review combine interventions, consistent with recent evidence suggesting coupling interventions has complementarities (Mbiti et al., 2019).

2 Learning-Adjusted Years of Schooling Framework

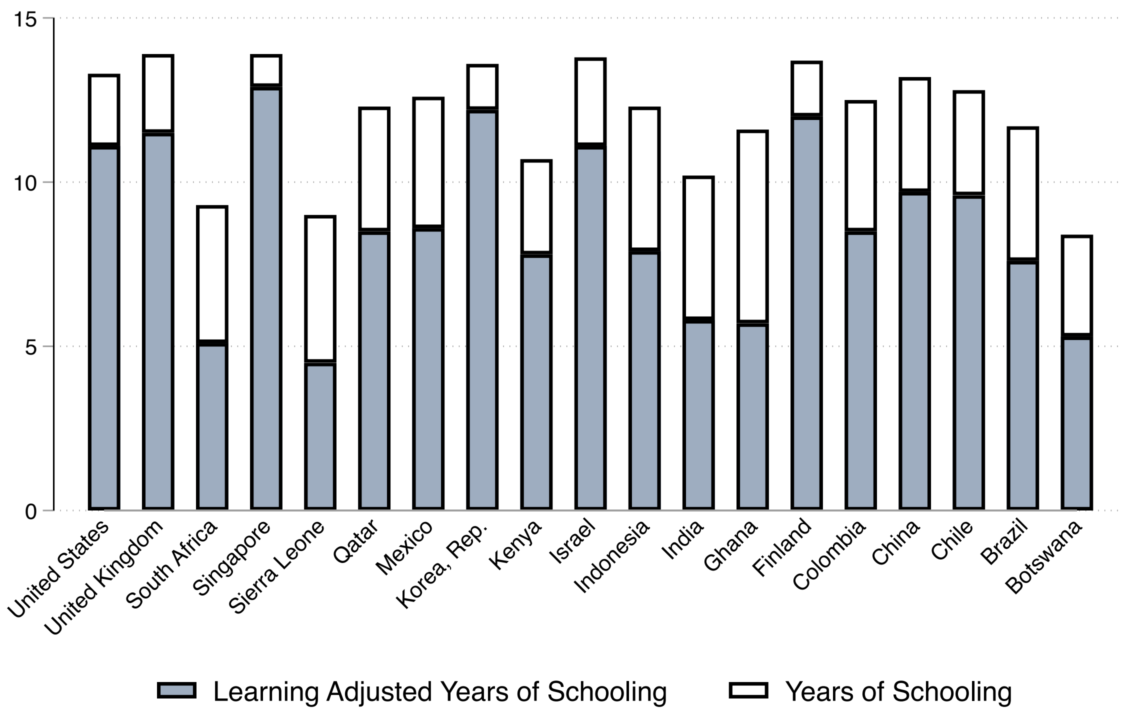
Learning-adjusted years of schooling for a given country—what we call macro-LAYS—are the product of years of schooling and a measure of schooling quality (Filmer et al., 2020). Specifically, they are produced by scaling the country's average schooling by its test-score performance relative to a global high-performance benchmark.[5] 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 provides a unified and user-friendly measure for a variety of education outcomes.

In this section, we show how LAYS can also be used at the micro level to compare specific education interventions and policies. The number of rigorous studies evaluating the effect of interventions on educational outcomes is growing, with hundreds of impact evaluations focused on learning outcomes in low- and middle-income countries (World Bank, 2018). A unified education 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—most notably, that access 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, global, and policy-salient metric. We refer to LAYS gained from an intervention or policy as micro-LAYS.

If impact evaluation studies tested students on common test items or common tests (such as internationally agreed test items from the Trends in International Mathematics and Science Study (TIMMS), the Programme for International Student Assessment (PISA) or the Early Grade Reading Assessment (EGRA), for example), 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.[6] To ensure a coherent unifying approach, the micro-LAYS methodology invokes assumptions similar

[5] The high-performance benchmark used in the World Bank Human Capital Index is an artificial benchmark of high performance of 625 as defined on 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.

[6] 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.



Notes: 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. (2021) learning data and UIS enrollment data.

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

to those used in constructing country-level macro-LAYS estimates. In this section, we outline the approach to producing micro-LAYS for evaluations that report effects on schooling participation, such as attendance or years of school gained, and subsequently for evaluations 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:

$$\text{LAYS}^p = \gamma_i * L_i^h * t$$

where L_i^h is a measure of learning for a cohort of students in country i relative to a high-performing benchmark h, such that $L_i^h = \frac{L_i}{L_h}$. 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 ($t = g$) given this uses available data without making additional projections.

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 deviations per year as a result of an intervention in South Africa, and if students typically learn 0.25 standard deviations 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 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 derive additional day-to-day benefits from schooling throughout their remaining school life expectancy.

Formally, we first express the intervention's learning impacts in terms of equivalent years of school gained. We derive equivalent years of schooling, e, by expressing learning gained relative to learning in the status quo:

$$e = \frac{\beta_i^{\text{test}}}{\delta_i^{\text{test}}}$$

where β is the learning gain produced by the intervention per year in country i; *test* denotes the test used to measure learning; and δ is the status-quo learning rate per year in country i.[7]

The student sample for the status quo learning trajectory, δ, in country i could represent the control group of the same study from which the β estimates are drawn; alternatively, it could be the average national student population in country i, in which case δ becomes the average learning trajectory for the country as a whole. This choice 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 benchmark country h:

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

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 relevant to a given test that is representative at national level n

[7] The conceptual notion of expressing learning gains in terms of equivalent years of schooling builds on the methodology used by Evans and Yuan (2019).

for each country, such that:

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

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 learning outcomes across tests and samples are comparable. This assumption is strong, but not novel: for example, 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. Over time, we expect impacts to be reported using increasingly comparable test items, in part motivated by the framework we put forward in this paper, enhancing the reliability of this assumption. We further explore robustness to this assumption in Section 5. These assumptions simplify our conversion to:

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

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

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

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 ($\beta_{\text{South Africa}} = 0.25$), and that in Singapore, a high-performance benchmark country on international learning assessments, students learn 1σ over the course of a given year ($\delta_{\text{Singapore}} = 1$). 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. Since micro-LAYS is based on participation estimates, t can take on a few values: a single year, such that $t = 1$;[8] 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.[9]

[8]See for instance, Figure B5 in the Appendix.

[9]The value of t, the length of time an intervention's effect is expected to last, might vary by intervention and

While many studies typically report learning results in terms of standard deviations, this framework is more general: it can incorporate results reported in terms of standard deviations as well as results using other learning outcome units. When common tests are used across studies and contexts, or common test items, comparisons are most comparable, although 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 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.

We view this framework as a starting point. By providing a unifying framework with transparent assumptions, we hope to both (a) make the best use of the available data which exists today and (b) encourage researchers to make greater use of standardized learning assessments in the future, which will in turn facilitate more meaningful comparisons across studies. To this end, introducing a common framework can make the best use of available data as well as set into motion a cycle of ever more comparable data and comparisons in education over time. This evolution follows the progression of DALYs and QALYs in the health sector, which started with a framework, assumptions, and a first analysis; and over time data inputs improved, enhancing comparability and facilitating cross-study comparisons. Even in this first analysis, our results are broadly robust to a series of tests and alternative choices, as detailed in the robustness section. In addition, for a set of studies that include identical underlying test items to measure learning, an 'ideal' benchmark comparison, we find similar results.

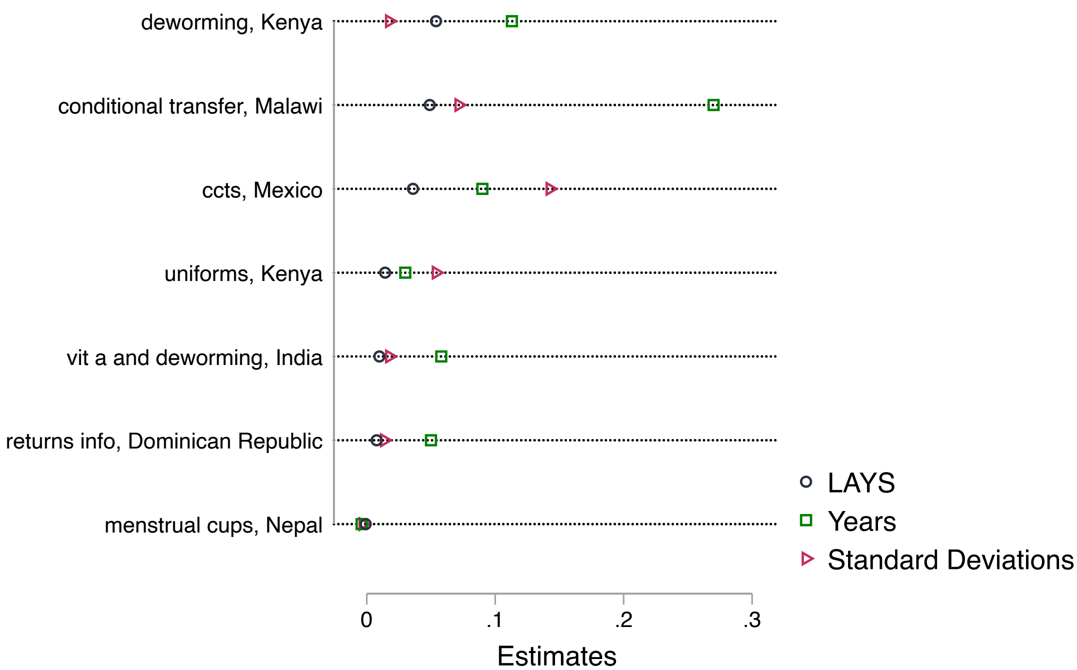
As an illustration of the insights gained by using LAYS, consider the following example. Figure 2 compares LAYS with the raw and standard deviations estimates per year for a subsample of studies that targeted school access and participation. Standardized effect sizes are the most typical measure used in the education literature to benchmark how large effects are relative to local variation, yet local variation can lead to misleading comparisons. As an example, deworming in Kenya yields 0.113 years of schooling, but only .018 standard deviation gains, while conditional cash transfers in Mexico yield 0.09 years of schooling, and 0.143 standard deviation gains. Thus, while both interventions have similar effects on schooling, standard deviations make it appear as if CCTs are five times more effective, but this result is an artifact of local variation.

We find that expressing outcomes in terms of LAYS yields substantive insight and new understanding relative to standard deviations of which interventions most efficiently improve education outcomes. In the same example, deworming in Kenya yields 0.054 LAYS relative to 0.036 LAYS for CCTs in Mexico. This preserves the original ranking of years of schooling gains and reflects education gains in clear, transparent, and absolute terms. Moreover, we see the added value

apply differently to quality improvements versus quantity improvements.

of capturing the quality of education using LAYS. With quality incorporated, deworming in Kenya pulls a bit further ahead relative to cash transfers in Mexico, since each year of schooling in Kenya produces more learning than in Mexico. The impact of quality adjustment is even more dramatic when comparing deworming in Kenya to CCTs in Malawi, where learning levels are far worse. CCTs in Malawi yield substantially more years of schooling (0.270) than deworming in Kenya (0.113). However, when we account for quality, we see deworming interventions yield 0.054 LAYS, while CCTs in Malawi yield 0.049 LAYS. Thus, using LAYS, we see that deworming interventions enhance education outcomes slightly more than CCTs (and at substantially lower cost, as it turns out). Had we used standard deviations, our understanding of which education interventions are more effective would be flipped. By adjusting for quality and reducing the influence of local variation, using LAYS allows us to say something about how effective an education intervention is using an absolute, cross-country standard as well as a unified education measure.

One challenge in assembling 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 whichever LAYS impact 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.



Notes: a subsample of interventions retrieved from J-PAL studies that focus on school access. Interventions are sorted by impacts in terms of LAYS. Estimates are reported per year. Standardized effects are calculated as $\frac{\mu_i}{s.e._i\sqrt{N_i}}$, *where* μ_i *is the raw estimate,* $s.e._i$ *is the standard error, and* N_i *is the number of observations in the raw estimate's regression for each intervention* i.

Figure 2: Comparing LAYS, Standardized Deviations, and Years of Schooling

3 Data and Analysis Framework

We compare impact estimates from over 200 evaluations of education interventions in 52 countries using a unified measure. In our comparison, we highlight findings from a subset of studies that have comparable cost data[10] and that therefore allow us to compare cost-effectiveness of interventions. We examine how many LAYS each policy or intervention delivers; how cost-effective those gains are; and how much of the gap between learning-adjusted years of schooling and actual years of schooling that intervention would close if it were scaled up.[11]

We start with nearly 300 observations aggregated across multiple evaluation databases.[12] We then add studies from the World Bank Strategic Impact Evaluation Fund (SIEF), as well as a large-scale new effort drawing on data collected in partnership with the Global Education Evidence Advisory Panel. This new set of studies draws from multiple rounds of reports as well as a systematic review of education interventions from low- and middle-income countries.[13] In total, we have 363 observations stemming from nearly 230 studies, 96 of which include cost data. Our review nearly doubles the number of cost-effectiveness estimates over earlier analyses, enabling us to draw substantial new insight over prior literature reviews on the most effective and cost-effective education interventions in low- and middle-income countries.

Our inclusion criteria are that studies 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 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 200 impact estimates across 52 low- and middle-income countries.

In this analysis, we make several choices for parameters and data inputs. First, we assume that the intervention's effects (t) last only for the duration of the evaluation since this is a known quantity and requires no further assumptions. In Appendix Figure B5 we explore the alternative assumption

[10] Given that there are substantial difficulties when comparing cost data across contexts and interventions, we believe the field will benefit from efforts to standardize how these costs are reported, for example by consistently using $PPP.

[11] For this last analysis, we assume 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.

[12] These databases include: Evans and Yuan (2019); Ganimian and Murnane (2016); Glewwe et al. (2011); Kremer et al. (2013); Krishnaratne et al. (2013)

[13] The systematic review started with over 13,200 studies. Based on a review of abstracts and titles, this list was narrowed to 725 studies, which were then analyzed and reduced to 325 research papers. Out of these studies, 46 included data on cost. After separating data points by treatment arm, we keep 53 new observations that we incorporate into our analysis. We add these estimates to a prior database that included over 150 studies compiled and used for LAYS analysis. These LAYS analyses have informed multiple Global Education Evidence Advisory Panel (GEEAP) flagship reports (Akyeampong et al., 2023).

that the impacts last just one year ($t = 1$). A second choice that we make is to set the high-quality benchmark learning rate ($\delta_{h,n}$) 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 PISA. In the 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 L_i^h) using Harmonized Learning Outcomes (HLO), which are global measures of learning introduced by Angrist et al. (2021) and used in the World Bank Human Capital Index. Angrist et al. (2021) generate comparable learning measures across 164 countries by linking psychometrically-designed international assessments to regional assessments to construct globally comparable learning outcomes at national levels. 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, these data are used in the World Bank Human Capital Index (HCI), which enables us to produce micro-LAYS that map directly to the macro-LAYS in the HCI.[14]

4 Results

4.1 Aggregate categories of policies and interventions

We first compare results for classes of policies and interventions, rather than focusing on individual studies. To this end, we summarize results by category, 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. These adjustments help classify interventions more precisely based on the primary theory of change underlying them. First, we recategorize technology interventions into either "computer assisted learning" or "additional inputs alone" based on whether they involved adaptive software or were largely a hardware-based intervention. We include a separate category for interventions that leverage the use of mobile phones specifically to deliver tutoring, which was often targeted to students' learning levels. Second,

[14] 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 B6 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.

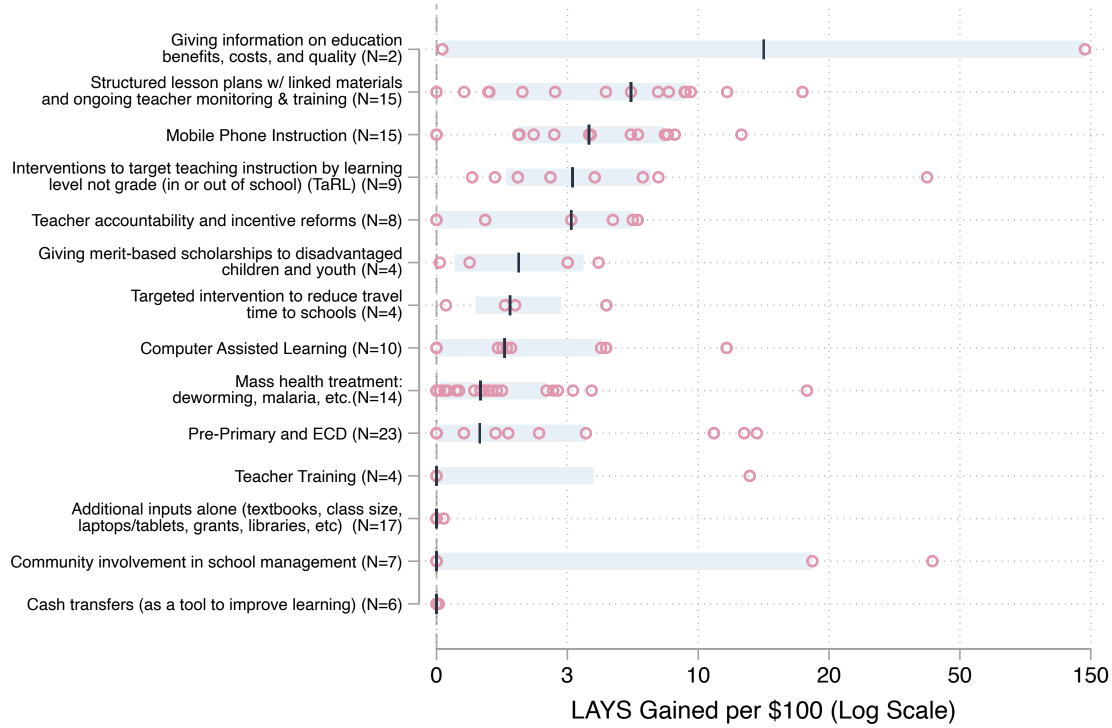
we classify interventions for ECD that involved building or opening of schools or classrooms as "targeted intervention to reduce travel time to schools" alongside other school construction programs. Third, we define teacher training interventions narrowly. Many interventions include training of teachers; 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. "Teacher training" captures only general-skills teacher training programs without other major elements. Fourth, for interventions with multiple components, we selected the central component and used that as a category. Later in the paper, we examine individual studies where we characterize studies 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 3 shows the LAYS gained per \$100. To calculate this, we divide the per-student gains by the per-student costs. Average 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, Lockheed, Ninan, & Tan, 2018). Therefore, cost-effectiveness expressed in terms of LAYS gained per \$100 is a metric consistent with many status-quo spending benchmarks, even at the lower tail of system spending.

The top-performing interventions, ranked by median effect size, are: targeted information campaigns on benefits, costs and quality; improved pedagogy in the form of structured lesson plans with linked materials and monitoring (which includes combination interventions such as Tusome in Kenya), mobile phone instruction such as targeted tutoring, interventions to target teaching instruction by learning level rather than grade (such as "Teaching at the Right Level" interventions and tracking interventions), teacher accountability and incentives (such as camera monitoring of teacher attendance or merit based pay), scholarships for disadvantaged groups, targeted interventions to reduce travel time to school (for example, constructing schools in remote underserved areas), computer assisted learning (such as adaptive learning software), health products (such as anti-malarial or deworming pills), and early childhood development (ECD) broadly defined. The four categories at the bottom of Figure 3—cash transfers, community involvement in school management (such as training for community members), additional inputs alone (such as textbooks, technology hardware, uniforms, school grants, or reducing class size without complementary reforms), and general skills teacher training—have a zero median effect on LAYS.[15]

We also 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

[15]These findings fed into, and are consistent with, the Global Evidence for Education Advisory Panel report (2023). Small differences include that here we combine ECD and interventions focusing on pre-primary education into a single category. We do the same for deworming and other mass health treatment interventions. We also group interventions on accountability and finding pathways to hiring educators into a joint group focused on teacher accountability and incentive reforms.



Notes: Each category of education intervention shows the learning-adjusted years of school (LAYS) per $100 USD gained from a given intervention or policy. Each red triangle represents a cost-effectiveness estimate. The boxplot is ordered from largest to smallest median effects and the shaded boxplot describes the 25th and 75th percentile. The y-axis is reported on a natural log scale. Studies with a negative effect size are set to a value of zero for this figure, given they are by definition not cost effective.

Figure 3: Learning-Adjusted Years of School (LAYS) Gained per $100 by Category

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, or under some conditions, but not in others.

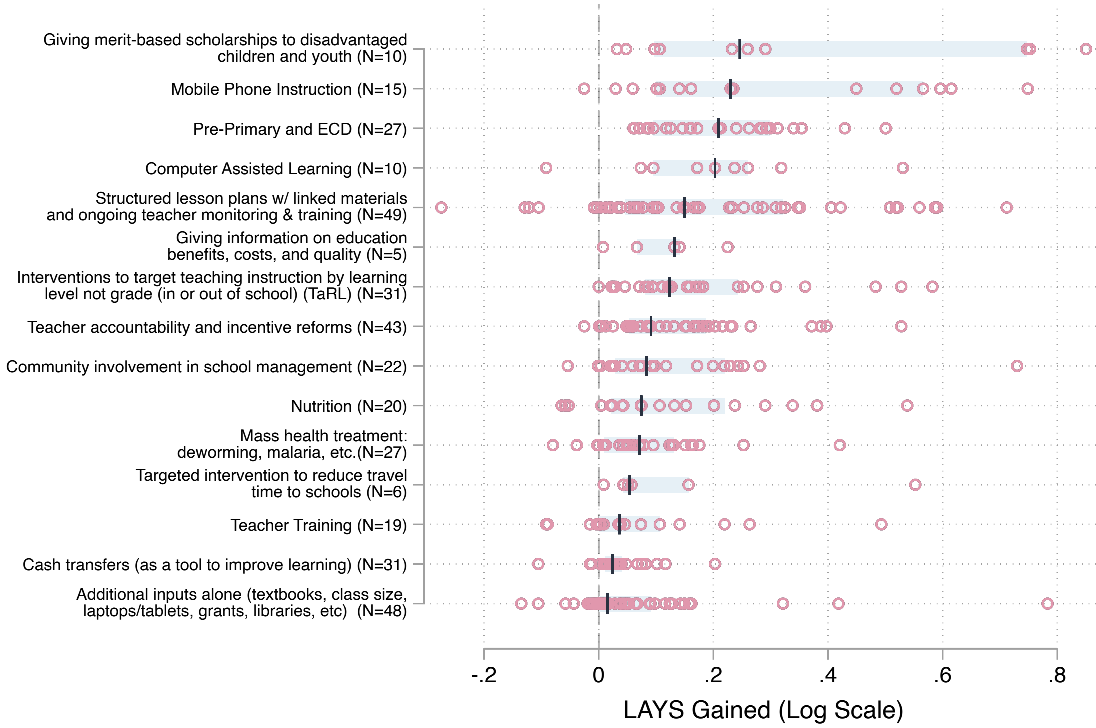
Moreover, context is essential to consider across all categories regardless of variation. For example, early childhood development might be most effective in contexts with strong primary education systems where these early investments translate into preparedness for primary school, thus enabling dynamic complementarities (Johnson & 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.

Some of the categories analyzed 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. Figure 4 illustrates this difference by ranking the interventions on the median LAYS gained per category. This figure 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 3 we see these modest absolute gains translate into over 3 LAYS gains per $100 per child, indicating that these health interventions can be highly cost-effective. Other interventions are highly effective but expensive. Giving merit-based scholarships can yield up to 1 LAYS, but since this policy is relatively expensive, it drops from top-ranked in terms of effectiveness to further down (though still highly ranked) in terms of cost-effectiveness. Finally, Figure 4 also includes a new category: nutrition interventions (such as school feeding), which did not enter the cost-effectiveness analysis in Figure 3 due to a lack of cost data. We observe that school feeding has a positive effect on LAYS, although with high variance, and in future analysis we aim to incorporate more cost data for this category.

We further find that studies with cost-effectiveness data are broadly representative of those with only effectiveness estimates. In Figure B1 we show our full set of studies, highlighting the subset of impact evaluations that include cost-effectiveness data. The most 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.

Overall, 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 targeting instruction to student's learning levels 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.

Our unified analysis reveals some important 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; while they yield gains in schooling in systems with low-quality education, they have done so without improving learning across the studies in our sample, all at relatively high cost. 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



Notes: Each category of education intervention shows the learning-adjusted years of school (LAYS) gained from a given intervention or policy across over 200 interventions in 52 countries. The boxplot describes the 25th and 75th percentile. The boxplot is ordered form largest to smallest median effects. Note the "nutrition" category has no cost-effectiveness data and does not appear in Figure 3.

Figure 4: Learning-Adjusted Years of School (LAYS) Gained by Intervention Category

(Fiszbein et al., 2009; Haushofer & Shapiro, 2016). Rather, these results suggest that if the goal of governments is to improve learning, cash transfers might not be the most efficient tool for this specific purpose. By contrast, some policies can yield on average around 3 additional LAYS per $100. We highlight two categories in particular which are both consistently effective and cost-effective with relatively low variation: (1) interventions to target teaching instruction by learning level rather than grade (e.g., "Teaching at the Right Level" interventions and tracking interventions); and (2) improved pedagogy in the form of structured lesson plans with linked student materials, teacher professional development, and monitoring. These categories of interventions have also been tested under multiple delivery models and are being scaled by multiple governments, demonstrating their relevance beyond the context of a controlled study. 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.

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 B2 shows results for absolute LAYS gained by intervention and country for the studies that include cost-effectiveness data. Some of the top performers are: a combined intervention with improved pedagogy, para-teachers and targeted instruction in The Gambia (4.04 LAYS); the Campaign for Female Education (CAMFED) program in Tanzania – a holistic program that includes scholarships and mentorship for girls along with school materials and training for teachers and parents (1.12 LAYS);[16] phone call tutorials that targeted instruction to students' learning levels in Uganda (1.11); Tusome (the Kiswahili word for "Let's Read") in Kenya—a program that provides structured pedagogy via textbooks, teacher coaching, and teacher training (1.04 LAYS); a comprehensive teacher training, structured curriculum, and coaching intervention in Argentina (0.81 LAYS); and an early literacy program in Uganda (0.80 LAYS). By contrast, about half of all interventions produce no significant effects; those interventions are not included in the figure.

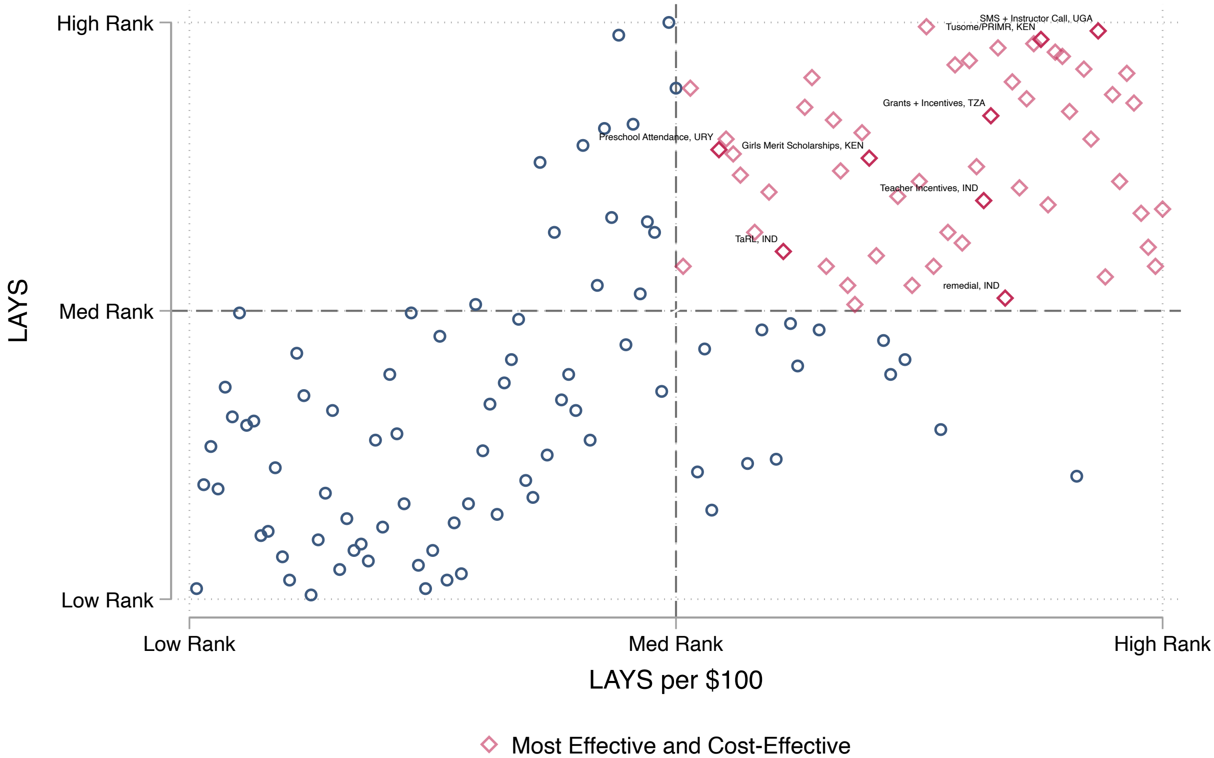
These findings point to a few overall lessons. In this sample of studies, the most effective programs are: multidimensional programs (a combined intervention in The Gambia, Camfed in Tanzania, and Tusome in Kenya); pedagogical instruction that is pitched to students' levels of learning, not based on a rote syllabus or an over-ambitious curriculum;[17] programs that facilitate early childhood development; and programs that are carefully targeted to a local need, such as scholarships for girls.

Figure B3 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. These include, provision of information on the returns to schooling in Madagascar, school links to village councils in Indonesia, tracking and grouping students by their learning level in Kenya, creating community-based preschools in Mozambique, and deworming in Kenya. By contrast, other interventions such as public-private partnerships, scholarship programs, targeted school construction and access, and computer technology-assisted adaptive instruction drop down the list because of their higher cost. However, these programs are still cost-effective in absolute terms.

There are two broader key takeaways from these figures. First, note that relatively few interventions have any positive impact at all. Indeed, over half of interventions reviewed had null effects and are omitted from these figures. Thus, any intervention identified as effective is already near the 50th percentile. Second, 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

[16] CAMFED does not only provide financial support in its program, but rather a holistic package of support.

[17] Related approaches to teaching at the right level include reforming the curriculum so that it focuses on foundational skills and aligns better with students' actual pace of learning. Such reforms have been evaluated in Tanzania with promising effectiveness (Rodriguez-Segura & Mbiti, 2022).



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

Figure 5: Effective and Cost-Effective Interventions

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 5 highlights a set of example interventions that are both effective and cost-effective. Some of the programs that do well on both measures include: targeted scholarships (for girls); instruction targeted to student levels through pedagogical interventions, grouping students, and technology; structured pedagogy interventions; and early childhood development programs. We include a more complete list of interventions that fall in this quadrant in Table B1 in the Appendix.

Overall, this exploration of specific interventions reveals broadly consistent patterns with the aggregate categories in Figures 3 and 4. Rather than delivering precise estimates or identifying specific interventions to invest in, this analysis is most useful for the aggregate patterns that it reinforces, such as the relative efficiency of interventions like targeting instruction to children's level or structured pedagogy over input-only reforms. Results for aggregate categories of policies are often most useful to inform prioritization by governments, with specific interventions being determined based on contextual need and relevance.

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

To explore how specific interventions and policies map onto system-wide gaps, we first calculate how many LAYS a given intervention could contribute toward closing the gap to achieve a full and globally benchmarked quality education 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 an intervention goes to scale. In essence, in this exercise we map micro-LAYS onto macro-LAYS. Figure B4 in the Appendix takes cost-effectiveness into account, showing the system-level gap that a given intervention could close at a cost of \$100 per child. This analysis reveals that policies which improve the productivity of each year of schooling, such as targeting instruction to a child's learning level, can yield up to 3 to 4 additional LAYS per \$100 in India – a gain equivalent to the entire system-level education gap between India and Argentina. This calibration illustrates that shifting the marginal dollar of government expenditure from low-efficiency to high-efficiency educational investments could help countries make much more out of the years of education they offer.

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. Finally, we leverage new data from a cross-country intervention (Angrist et al., 2023) that uses identical test items to measure learning for the same intervention across four countries. Results show that LAYS conversions preserve ranks in line with the 'ideal' scenario that uses identical tests.

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 (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 using rescaled HLO benchmarks, which yields year-on-year gains of 0.96σ.[18]

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, 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 & Yuan, 2022). 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.

[18] We assume a rescaled high performance 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.

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 B6, 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 B7 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.

Rescaling mainly reduces 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 1.6σ. Overall, our findings are not affected by using these alternative definitions of LAYS. We show the correlation between our various rescaled measures, as well as the ranking of interventions they produce, in Table B2; all correlations are above 0.97.

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, re-scaling is likely to be important for capturing the full extent of the learning gap.

5.3 Standard deviations across tests and samples

We test sensitivity of our results to differences in standard deviations across tests and samples. 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 B8 compares, for Tanzania, Malawi, and Indonesia, the standard deviations in HLO score calculated using various assessments: EGRA, the Progress in International Reading Literacy Study (PIRLS), and the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) assessment. 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 B9. As a robustness check, Appendix Figure B9 also shows learning rates per year using raw data from the assessments before they are converted to HLO scores.

5.4 Status-quo learning

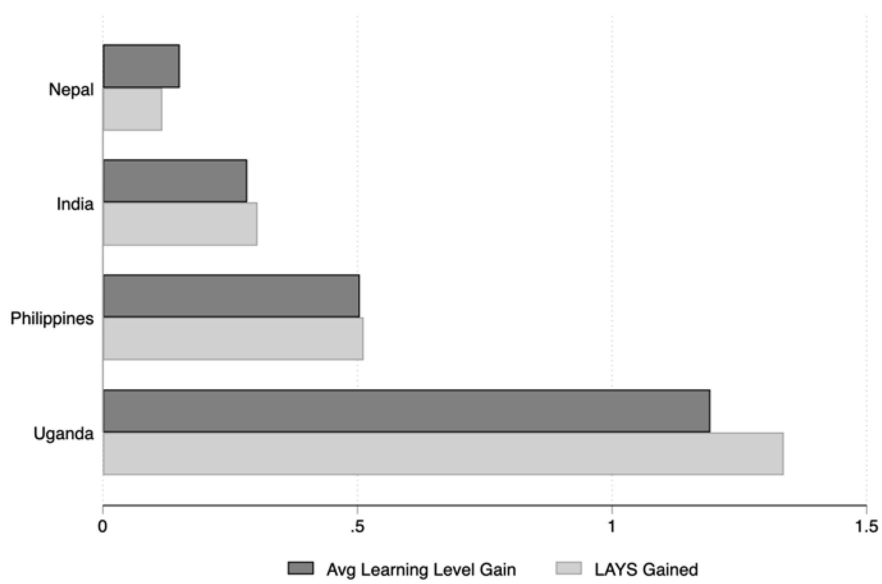
When computing LAYS from learning estimates, we first convert learning into equivalent years of school gained. To do this, we express learning relative to status-quo learning trajectories. For example, if students learned 0.25σ per year in an intervention in South Africa, and students typically learn 0.25σ in a given year in South Africa, impacted students will have gained a year's worth of typical learning in South Africa. A few options exist for possible status-quo trajectories: for example, we might use national averages, or 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 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. In the main analysis reported in the paper, 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 B10, 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 necessarily represents the nation.

In summary, we find that the assumptions for using control-group learning trajectories as our measure of 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 practical: greater data availability. While control-group information is often missing from published papers, national learning trajectories can be calculated using HLO scores, which are available for 164 countries.

5.5 Comparing LAYS with identical underlying test items

An 'ideal' comparison would involve comparing an intervention's effectiveness across contexts using identical underlying test items. While such common assessment is rare in the education sector, a recent study by Angrist et al. (2023) involved a large-scale randomized trial across several countries using the same tests in each context. These assessments were adapted from the ASER numeracy test, which has been widely used in the literature (Banerjee et al., 2017; Banerjee, Cole, Duflo, & Linden, 2007). This assessment captures core competencies, including addition, subtraction, multiplication, and division. Figure 6 compares average levels gained across four countries – India, Nepal, the Philippines, and Uganda – that tested a phone call tutorial intervention that targeted instruction to students' learning levels. First, we estimate the average level gained, coded 1 for each additional proficiency learned. The figure also shows the impact of each intervention measured using LAYS units. We see that the ranking of impacts is preserved. This analysis highlights both the value of using comparable tests to measure the impact of interventions across contexts, as well as robustness of the LAYS metric to the ideal scenario of comparisons using common test items.



Notes: Average level refers to mean proficiencies gained (e.g. addition to division). Source: Angrist et al. (2023).

Figure 6: LAYS vs. Common Test Item Learning Level Gains

6 Conclusion

In this paper, we analyze which investments most efficiently improve education outcomes. Expanding on previous education reviews, we analyze over 200 interventions and policies across 52 countries using a 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, global evidence can then be used to identify possible cost-effective solutions, which should then be carefully adapted to the local context.

This analysis uses 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 multiple 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, & Schipper, 2020), yet an additional year of schooling can lead to very different levels of learning (Singh, 2020; World Bank, 2018). Using LAYS as a metric of progress allows a focus on additional years and learning together. The LAYS framework also helps set in motion a cycle of ever more common approaches to measuring learning in the future.

The LAYS metric has recently 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 is being used by the World Bank, UNICEF, and United Kingdom's Foreign, Commonwealth & Development Office (FCDO) to prioritize cost-effective education investments. These efforts demonstrate the value of the analysis in this paper to provide a useful tool and synthesis for policymakers, researchers, and decisionmakers who are seeking to address persistent gaps in access and learning worldwide.

Finally, we contribute to the literature synthesizing results from impact evaluations in education in general, updating prior reviews with hundreds of new impact evaluations emerging in the past decade. This paper updates the literature with the most recent and comprehensive set of evaluations with a novel focus on cost-effectiveness analysis. Our results provide guidance on policies and interventions which can enable orders of magnitude more efficient investments in education.

References

Akyeampong, K., Andrabi, T., Banerjee, A., Banerji, R., Dynarski, S., Glennerster, R., ... Yoshikawa, H. (2023). *Cost-Effective Approaches to Improve Global Learning - What does recent evidence tell us are "Smart Buys" for improving learning in low- and middle-income countries?* (Tech. Rep.). London, Washington D.C., New York.: FCDO, the World Bank, UNICEF, and USAID.

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of public economics*, *95*(7), 476–487.

Angrist, N., Ainomugisha, M., Bathena, S. P., Bergman, P., Crossley, C., Cullen, C., ... Sullivan, T. (2023). *Building Resilient Education Systems: Evidence from Large-Scale Randomized Trials in Five Countries.*

Angrist, N., Djankov, S., Goldberg, P. K., & Patrinos, H. A. (2021). Measuring human capital using global learning data. *Nature (London)*, *592*(7854), 403–408.

ASER Centre. (2018). *Annual Status of Education Report 2017 'Beyond Basics'* (Tech. Rep.). ASER Centre.

Azevedo, J. P., Goldemberg, D., Montoya, S., Nayar, R., Rogers, H., Saavedra, J., & Stacy, B. W. (2021). Will Every Child Be Able to Read by 2030? Defining Learning Poverty and Mapping the Dimensions of the Challenge.

Baird, S., McIntosh, C., & Özler, B. (2011). Cash or condition?: Evidence from a cash transfer experiment. *The Quarterly journal of economics*, *126*(4), 1709–1753.

Banerjee, A. V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., ... Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *The Journal of economic perspectives*, *31*(4), 73–102.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly journal of economics*, *122*(3), 1235–1264.

Bashir, S., Lockheed, M., Ninan, E., & Tan, J.-P. (2018). *Facing Forward : Schooling for Learning in Africa*. Washington, DC: World Bank.

Bates, M. A., & Glennerster, R. (2017). *The Generalizability Puzzle.*

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions. *Journal of research on educational effectiveness*, *1*(4), 289–328.

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

Crawfurd, L., Evans, D., Hares, S., & Moscoviz, L. (2020). *12 Years of Quality Education for Every Girl: Five Ways the New UK Government Can Deliver on Its Manifesto Pledge.* Retrieved 2022-08-05, from https://www.cgdev.org/blog/12-years-quality-education-every-girl-five-ways-new-uk-government-can-deliver-its-manifesto

De Neve, J.-W., Fink, G., Subramanian, S. V., Moyo, S., & Bor, J. (2015). Length of secondary schooling and risk of HIV infection in Botswana: evidence from a natural experiment. *The Lancet global health*, *3*(8), e470—-e477.

Drummond, M., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes.* (Fourth edi ed.). Oxford, United Kingdom.

Duflo, E., Dupas, P., & Kremer, M. (2021). *The Impact of Free Secondary Education: Experimental Evidence from Ghana.* Cambridge, Mass..

Dynarski, S. (2000). Hope for Whom? Financial Aid for the Middle Class and Its Impact on College Attendance. *National tax journal*, *53*(3), 629–661.

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

Evans, D. K., & Popova, A. (2016b). What really works to improve learning in developing countries?: An analysis of divergent findings in systematic reviews. *The World Bank research observer*, *31*(2), 242–270.

Evans, D. K., & Yuan, F. (2019). Equivalent Years of Schooling : A Metric to Communicate Learning Gains in Concrete Terms. , *8752*.

Evans, D. K., & Yuan, F. (2022). How Big Are Effect Sizes in International Education Studies? *Educational evaluation and policy analysis*, *44*(3), 532–540.

Filmer, D., Rogers, H., Angrist, N., & Sabarwal, S. (2020). Learning-adjusted years of schooling (LAYS): Defining a new macro measure of education. *Economics of education review*, *77*.

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

Ganimian, A. J., & Murnane, R. J. (2016). Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations. *Review of educational research*, *86*(3), 719–755.

Glewwe, P. W., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2011). *School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010.* Cambridge, Mass: National Bureau of Economic Research.

Glewwe, P. W., & Muralidharan, K. (2016). Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications. In *Handbook of the economics of education* (Vol. 5, pp. 653–743).

Habyarimana, J. P., Opalo, K. O., & Schipper, Y. (2020). The Cyclical Electoral Impacts of Programmatic Policies: Evidence from Education Reforms in Tanzania..

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth?: Cognitive skills, economic outcomes, and causation. *Journal of economic growth (Boston, Mass.)*, *17*(4), 267–321.

Haushofer, J., & Shapiro, J. (2016). The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya. *The Quarterly Journal Of Economics*, *131*(4), pp1973—-2042.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of public economics*, *94*(1), 114–128.

Hendren, N., & Sprung-Keyser, B. (2020). A Unified Welfare Analysis of Government Policies. *The Quarterly journal of economics*, *135*(3), 1209–1318.

Johnson, R. C., & Jackson, K. C. (2019). Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending. *American economic journal. Economic policy*, *11*(4), 310–349.

Kane, T. J. (2004). *The Impact of After-School Programs; Interpreting the Results of Four Recent Evaluations.* Los Angeles, CA.

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

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational researcher*, *49*(4), 241–253.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The Challenge of Education and Learning in the Developing World. *Science (American Association for the Advancement of Science)*, *340*(6130), 297–300.

Krishnaratne, S., White, H., & Carpenter, E. (2013). *Quality education for all children? What works in education in developing countries.*

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). INPUTS, INCENTIVES, AND COMPLEMENTARITIES IN EDUCATION: EXPERIMENTAL EVIDENCE FROM TANZANIA. *The Quarterly journal of economics*, *134*(3), 1627–1673.

McKeever, V. (2020, jun). *Oxford degree, says now it's time for 'Netflix, reading and sleep'.* Retrieved from https://www.cnbc.com/2020/06/19/malala-yousafzai-completes-her-oxford-degree.html

Murray, C. J. L., & Lopez, A. D. (1996). Evidence-Based Health Policy-Lessons from the Global Burden of Disease Study. *Science (American Association for the Advancement of Science)*, *274*(5288), 740–743.

Pliskin, J. S., Shepard, D. S., & Weinstein, M. C. (1980). Utility Functions for Life Years and Health Status. *Operations research*, *28*(1), 206–224.

Psacharopoulos, G., & Patrinos, H. A. (2004). Returns to investment in education: a further update. *Education economics*, *12*(2), 111–134.

Rodriguez-Segura, D., & Mbiti, I. (2022). *Back to Basics: Curriculum reform and student learning in Tanzania.* Retrieved from https://doi.org/10.35489/BSG-RISE-WP_2022/099

Singh, A. (2020). Learning More with Every Year: School Year Productivity and International Learning Divergence. *Journal of the European Economic Association*, *18*(4), 1770–1813.

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., . . . Eyers, J. (2015). *Interventions for improving learning outcomes and access to education in low- and middle- income countries: a systematic review.* London.

Torrance, G. W., Thomas, W. H., & Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health services research*, *7*(2), 118–133.

World Bank. (2018). *World Development Report 2018 : Learning to Realize Education's Promise*. Washington, DC: World Bank.

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

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

Zeckhauser, R., & Shepard, D. (1976). Where Now for Saving Lives? *Law and contemporary problems*, *40*(4), 5–45.

Appendices

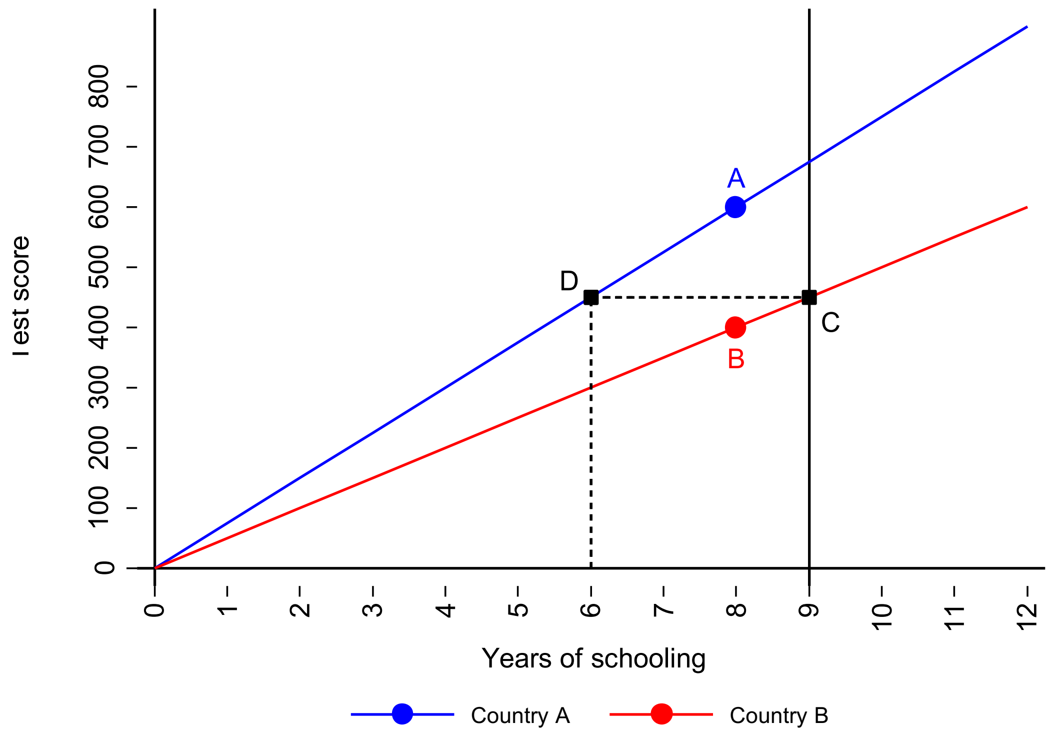
Appendix A Constant average learning trajectories

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.[19] Figure A1 shows the usefulness 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 Centre, 2018). 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.[20] 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).

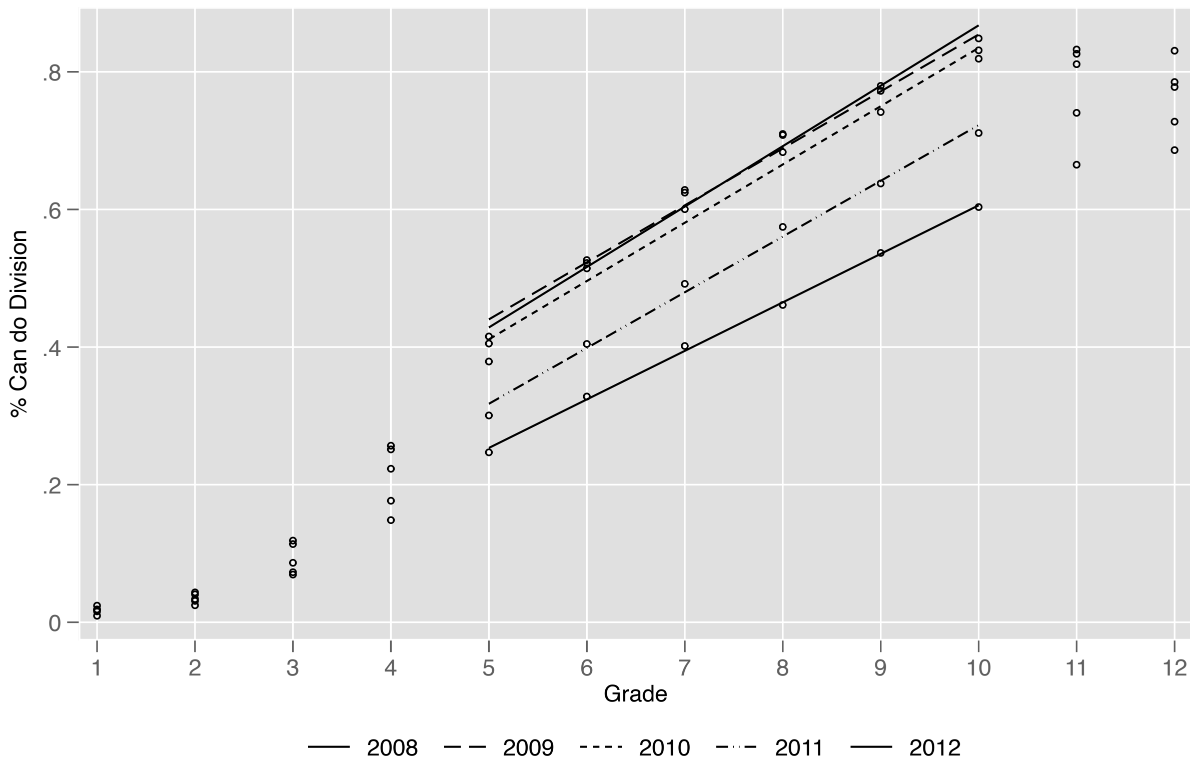
[19] This framework focuses on learning within schools. Clearly, not all learning happens within schools. However, learning outside schools is beyond the scope of this exercise. For a fuller discussion see Filmer et al. (2020).

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



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

Figure A1: Constant Learning Trajectories

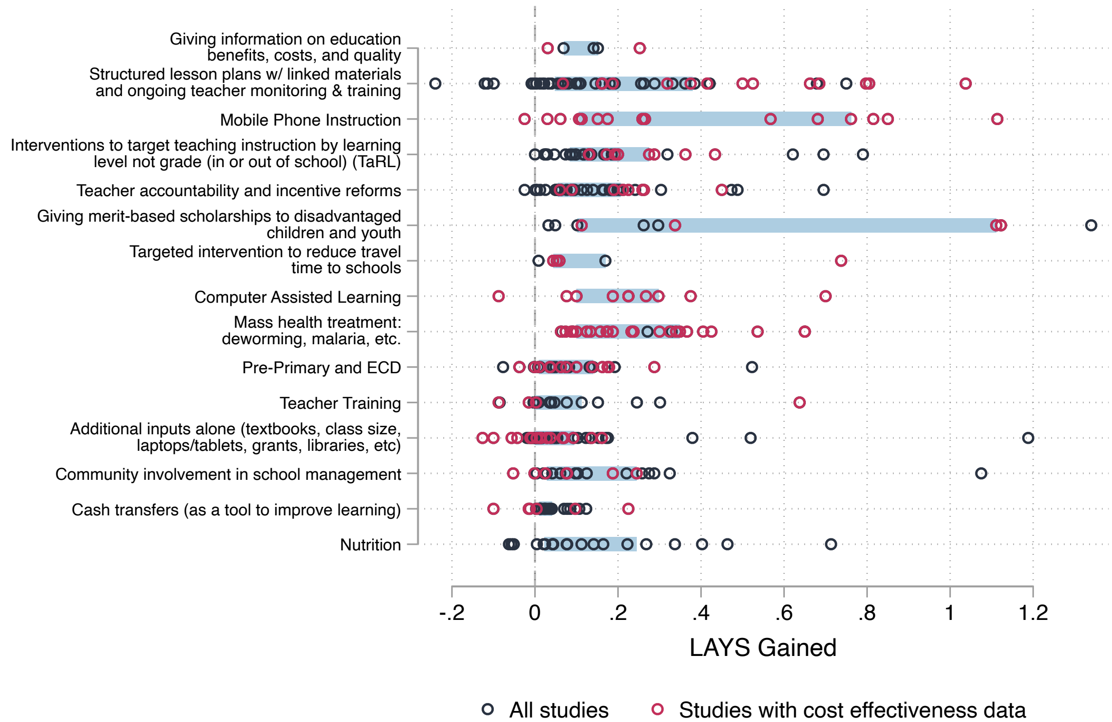


Notes: We derive learning trajectories using empirical data from a national survey conducted in households in India for students aged 5 to 16 in grades 1 through 12. We include only students at the household who are in school.

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

Figure A2: Learning Trajectories in India

Appendix B Additional Figures and Tables



Notes: Each category of education intervention shows the learning-adjusted years of school (LAYS) gained from a given intervention or policy across over 200 interventions in 52 countries. The boxplot describes the 25th and 75th percentile. The boxplot is ordered in the same order as Figure 3 to provide a direct analogy, with the exception of the "nutrition" category, which has no cost-effectiveness data and therefore does not appear in Figure 3.

Figure B1: Learning-Adjusted Years of School (LAYS) Gained by Intervention Category

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

Figure B2: Learning-Adjusted Years of Schooling (LAYS) by Intervention

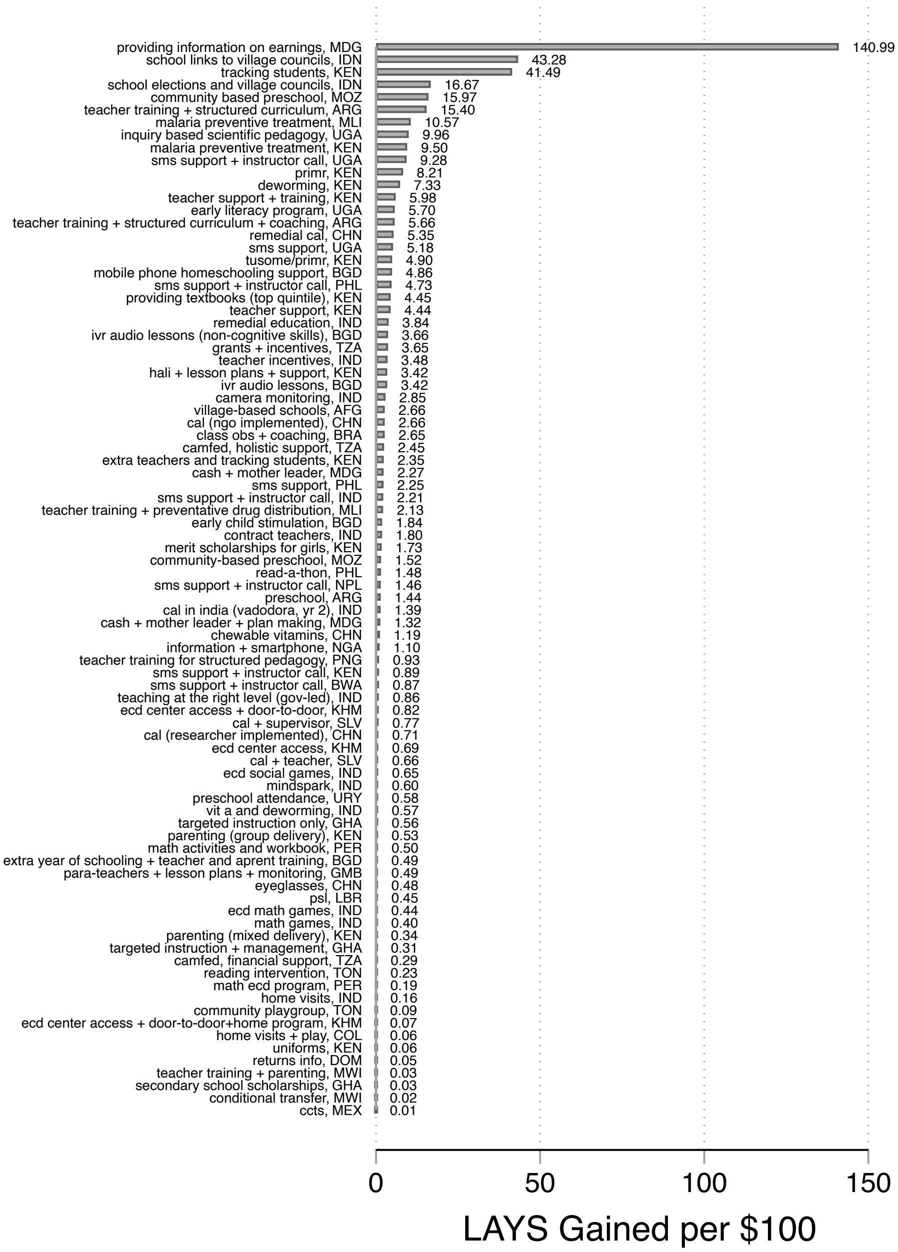
Table B1: Interventions with Above Median Rank in Cost Effectiveness and Effectiveness

Intervention	LAYS per $100 USD	LAYS
providing information on earnings, MDG	141	.252
school links to village councils, IDN	43.3	.188
tracking students, KEN	41.5	.2
school elections and village councils, IDN	16.7	.245
community based preschool, MOZ	16	.536
teacher training + structured curriculum, ARG	15.4	.685
malaria preventive treatment, MLI	10.6	.287
inquiry based scientific pedagogy, UGA	9.96	.637
malaria preventive treatment, KEN	9.5	.179
sms support + instructor call, UGA	9.28	1.11
primr, KEN	8.21	.375
remedial cal, CHN	8.18	.7
teacher support + training, KEN	5.98	.5
early literacy program, UGA	5.7	.798
teacher training + structured curriculum + coaching, ARG	5.66	.805
sms support, UGA	5.18	.259
tusome/primr, KEN	4.9	1.04
mobile phone homeschooling support, BGD	4.86	.85
sms support + instructor call, PHL	4.73	.567
providing textbooks (top quintile), KEN	4.45	.275
teacher support, KEN	4.44	.663
remedial education, IND	3.84	.172
ivr audio lessons (non-cognitive skills), BGD	3.66	.815
grants + incentives, TZA	3.65	.45
teacher incentives, IND	3.48	.262
hali + lesson plans + support, KEN	3.42	.319
ivr audio lessons, BGD	3.42	.761
camera monitoring, IND	2.85	.213
village-based schools, AFG	2.66	.738
cal (ngo implemented), CHN	2.66	.225
remedial cal, CHN	2.52	.188
camfed, holistic support, TZA	2.45	1.12
extra teachers and tracking students, KEN	2.35	.287
cash + mother leader, MDG	2.27	.175
sms support + instructor call, IND	2.21	.265
contract teachers, IND	1.8	.19
merit scholarships for girls, KEN	1.73	.338
community-based preschool, MOZ	1.52	.405
read-a-thon, PHL	1.48	.162
sms support + instructor call, NPL	1.46	.175
preschool, ARG	1.44	.3
cal in india (vadodora, yr 2), IND	1.39	.434
cash + mother leader + plan making, MDG	1.32	.188
information + smartphone, NGA	1.1	.681
teacher training for structured pedagogy, PNG	.927	.525
teaching at the right level (gov-led), IND	.863	.192
cal + supervisor, SLV	.768	.268
cal (researcher implemented), CHN	.71	.225
cal + teacher, SLV	.656	.298
ecd social games, IND	.65	.34
mindspark, IND	.599	.375
preschool attendance, URY	.585	.349
parenting (group delivery), KEN	.531	.65
math activities and workbook, PER	.498	.188

Table B2: Pairwise Correlations Between LAYS and Unscaled LAYS

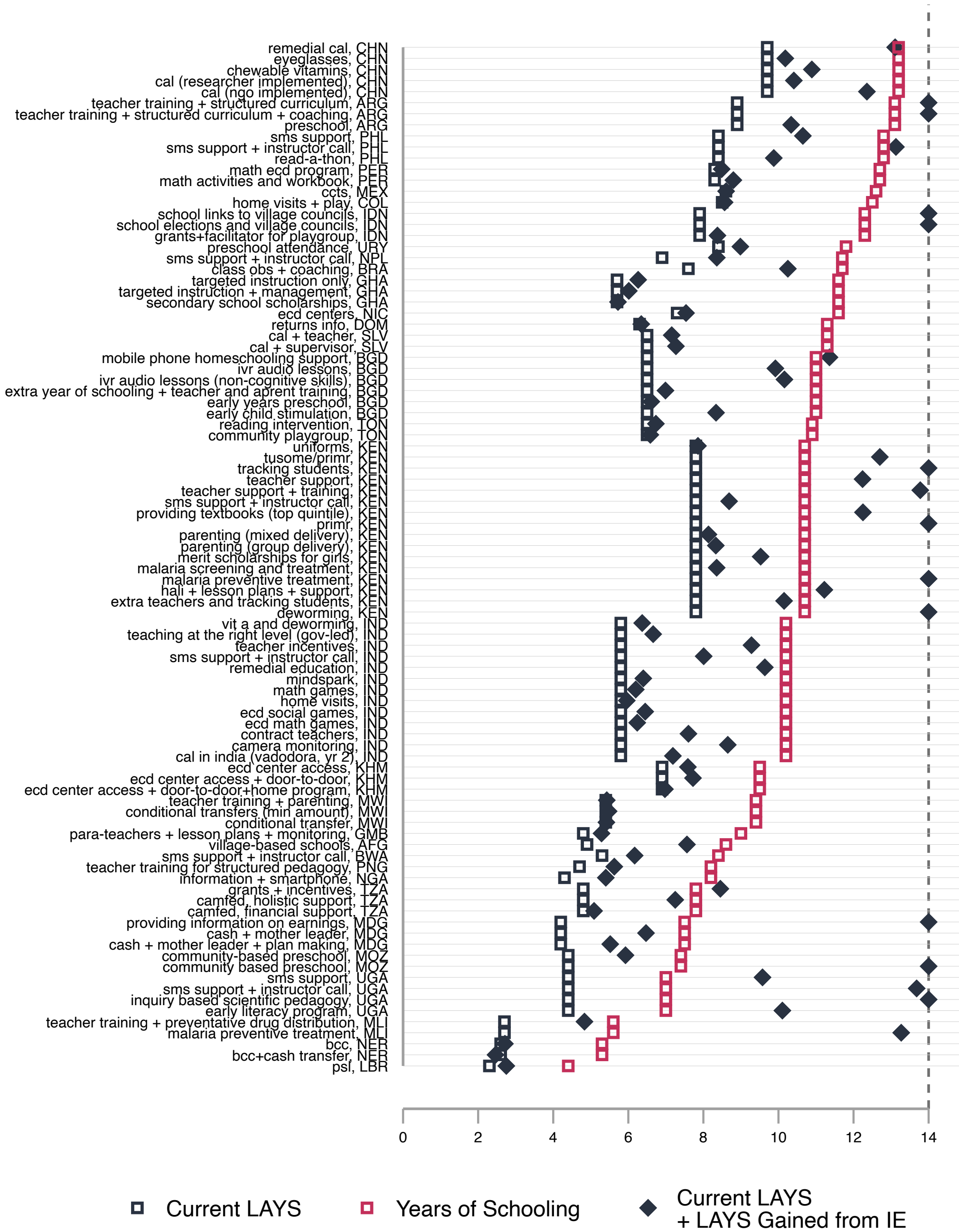
Panel A: LAYS			
LAYS	1.000		
LAYS Unscaled (Learning conversion rates)	1.000	1.000	
LAYS Unscaled (1.6σ Performance Benchmark)	0.978	0.978	1.000

Panel B: Ranks			
LAYS	1.000		
LAYS Unscaled (Learning conversion rates)	1.000	1.000	
LAYS Unscaled (1.6σ Performance Benchmark)	0.988	0.988	1.000



Notes: We omit interventions with a null effect.

Figure B3: Learning-Adjusted Years of Schooling (LAYS) Gained per $100, by Intervention



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 exercise, 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.

Figure B4: LAYS Gained per $100 per Intervention, Calibrated to Country-Level LAYS Gaps

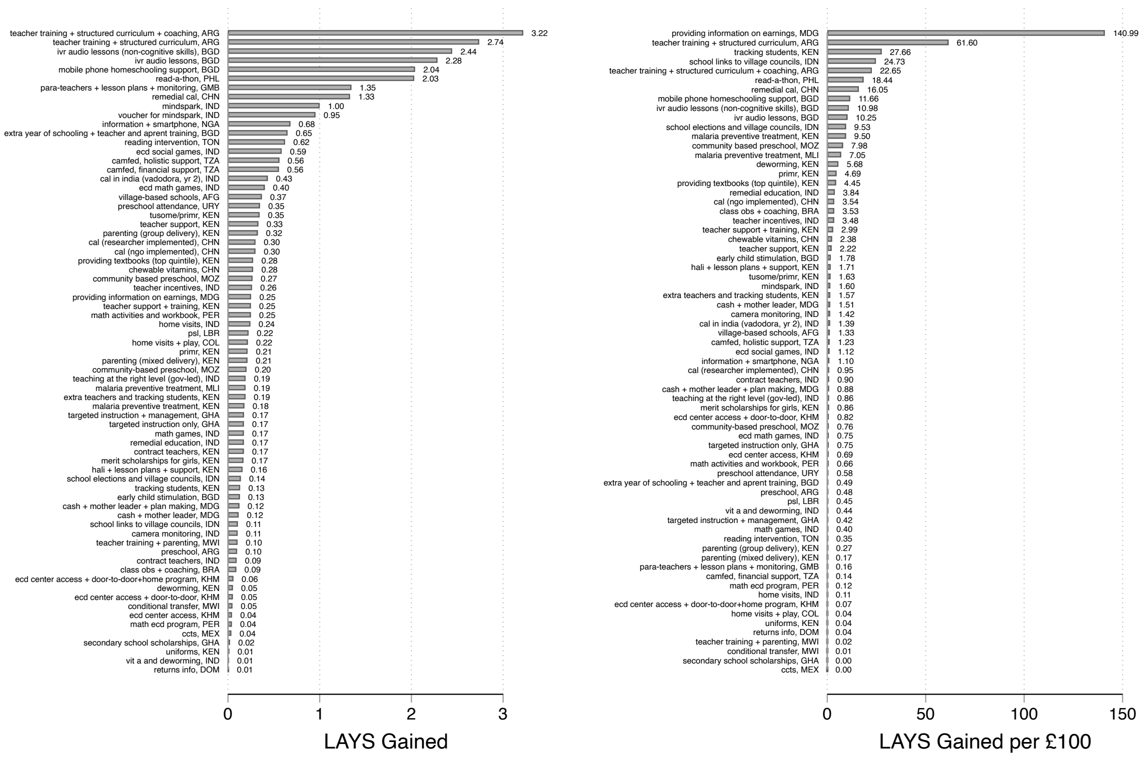
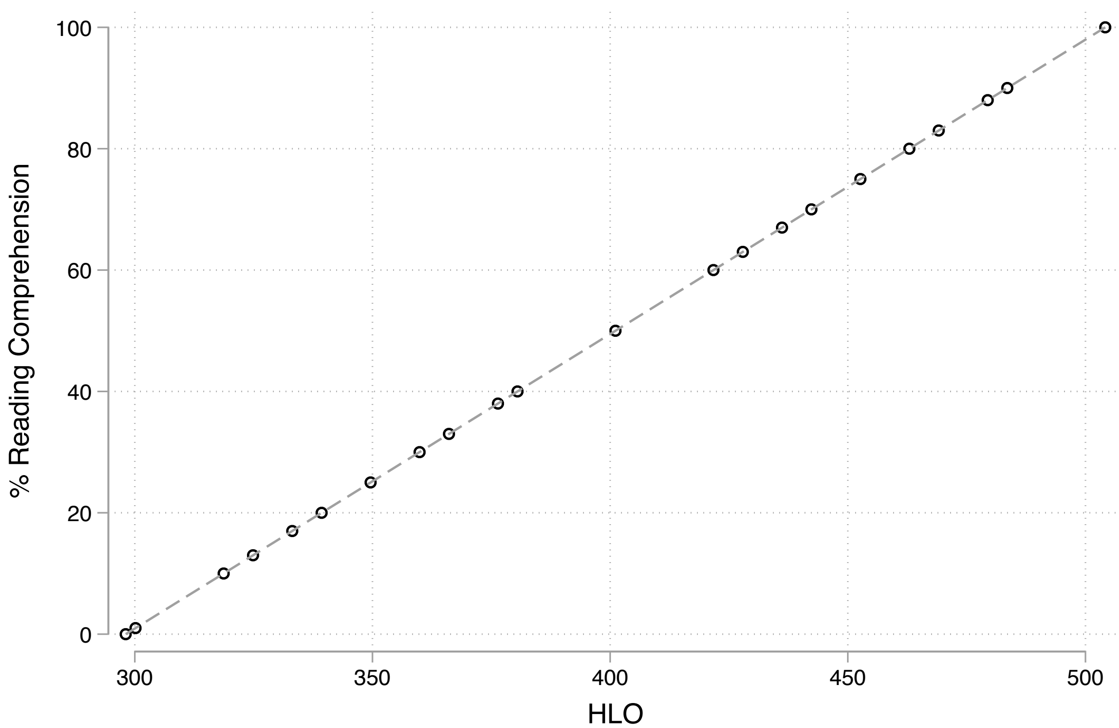
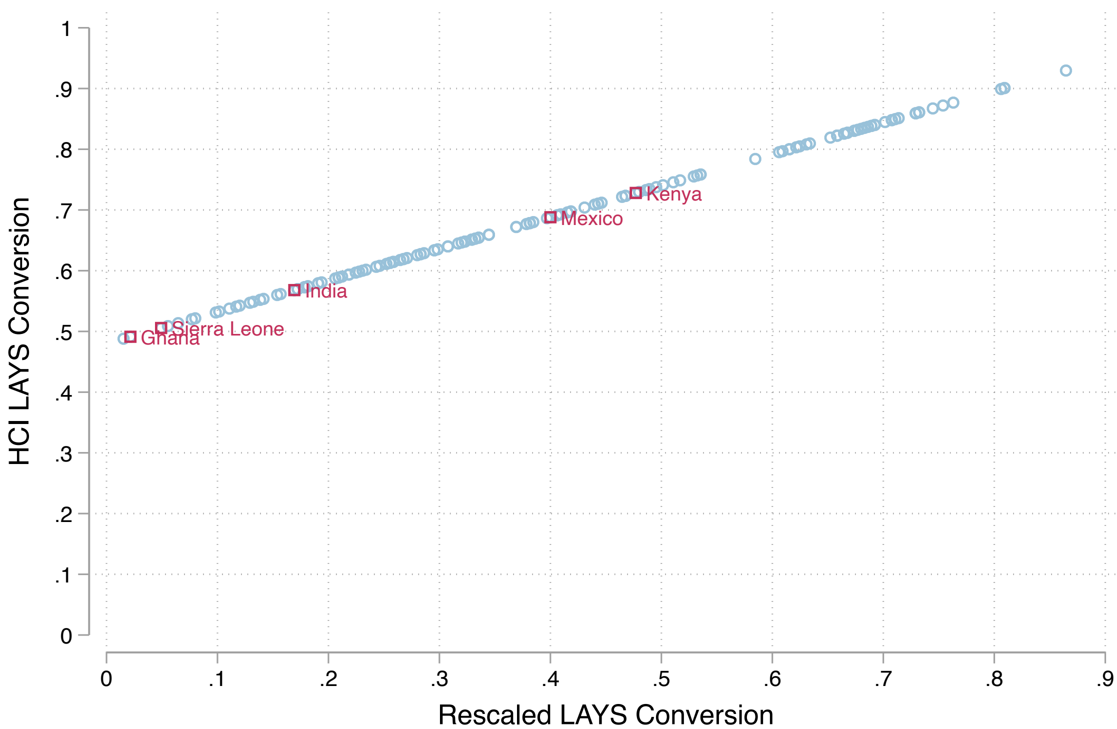


Figure B5: Expressing LAYS gained per year (t = 1)



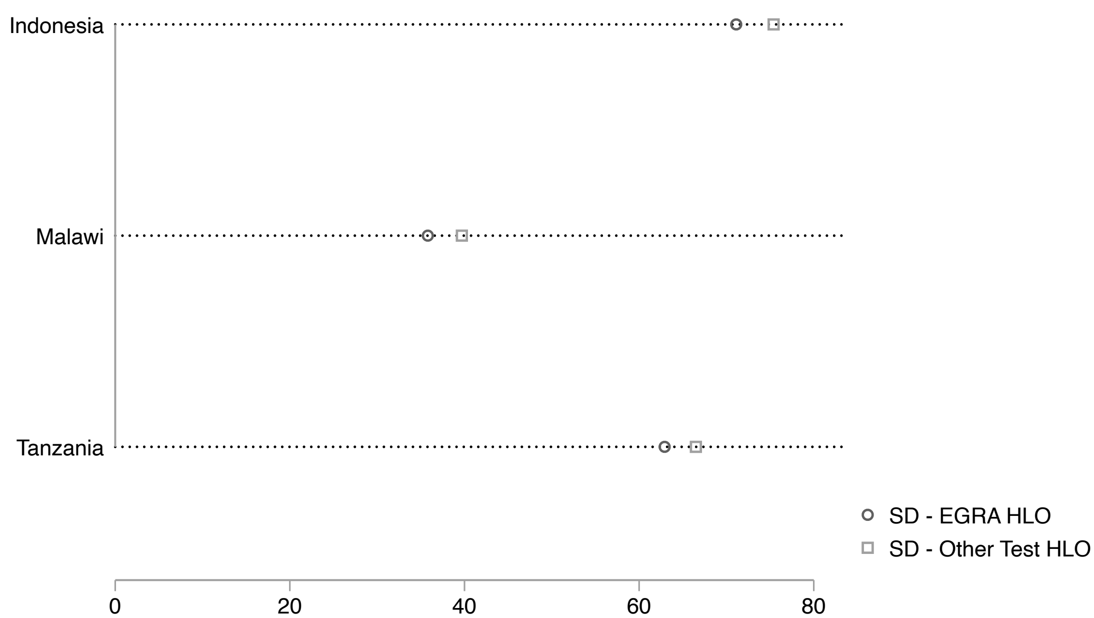
Notes: We analyze EGRA data across 39 countries and match raw score on reading comprehension modules with the Harmonized Learning Outcome (HLO) scores used for the World Bank Human Capital Index.

Figure B6: EGRA raw reading comprehension relative to HLO score



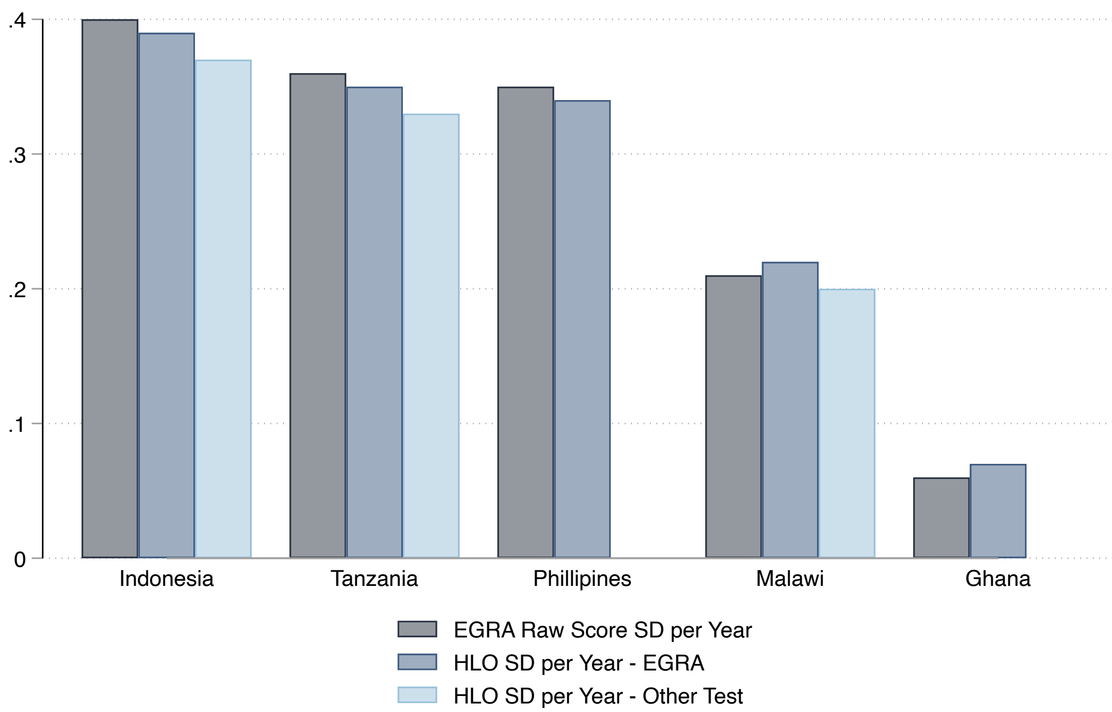
Notes: We rescale LAYS conversion rates. Initial conversion rates are based on scores which often floor around 300 due to underling test scores scales. Since the LAYS conversion rate is calculated out of 625, this produces a floor conversion rate of .48. However, when learning levels are very low this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

Figure B7: Learning Adjustment Rates



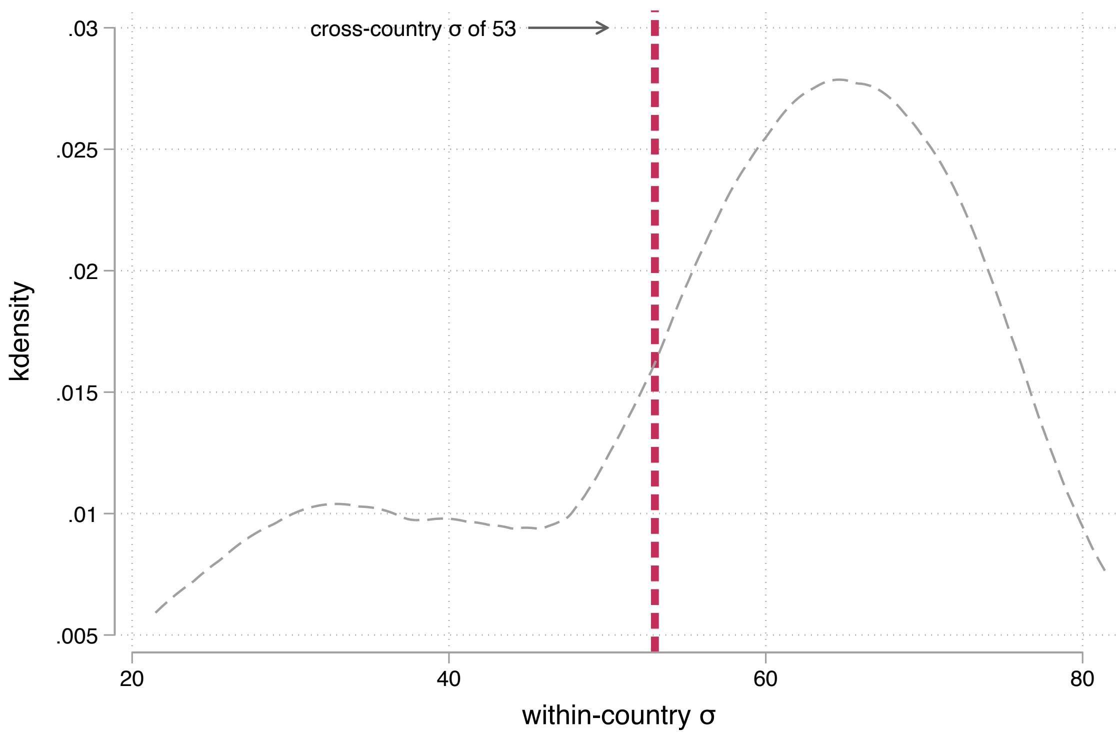
Notes: For Indonesia, the "other test" is PIRLS 2011; for Tanzania and Malawi it is SACMEQ 2007.

Figure B8: SD Comparisons, 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 B9: Learning Per Year (in SD), by Source Test



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.

Figure B10: Within- vs. Cross-Country Variation in Test Scores

Appendix C Full List of Citations

Abeberese, A. B., Kumler, T. J., & Linden, L. L. 2014. Improving Reading Skills by Encouraging Children to Read in School: A Randomized Evaluation of the Sa Aklat Sisikat Reading Program in the Philippines. Journal of Human Resources, 49(3),611-633

Adrogue, C. & Orlicki, M. E. 2013. Do in-school feeding programs have an impact on academic performance? The case of public schools in Argentina. Education Policy Analysis Archives, 21, 2-20.

Adukia, A. 2016. Sanitation and education. American Economic Journal: Applied Economics (forthcoming).

Afridi, F., Barooah, B. & Somanathan, R. 2014. The mixture as before? Student responses to the changing content of school meals in India. IZA Discussion Paper.

Akresh, R., De Walque, D., & Kazianga, H. 2013. Cash Transfers and Child Schooling: Evidence from a Randomized Evaluation of the Role of Conditionality World Bank Policy Research Working Paper No. 6340. Washington, DC: The World Bank.

Akresh, R., de Walque D., & Kazianga H.. 2016. Evidence from a randomized evaluation of the household welfare impacts of conditonal and unconditional cash transfers given to mothers or fathers. World Bank Policy Research Working Paper 7730.

Alderman, H., Gilligan D.O., & Lehrer K. 2012. "The impact of food for education programs on school participation in northern Uganda." Economic Development and Cultural Change 61(1):187-218.

Alviar, C., Davis, B., Handa, S., Hurrell, A., Hussein, A., Musembi, D., Ochieng, S., Pearson, R., Pellerano, L., Visram, A., Ward, P. and Kenya, C. T. O. V. C. E. T., 2012. The impact of Kenya's cash transfer for orphans and vulnerable children on human capital. Journal of Development Effectiveness, 4, 38-49. 533

Amarante, V., Ferrando, M. and Vigorito, A., 2013. Teenage school attendance and cash transfers: An impact evaluation of panes. Economia. Vol 14, No 1(Fall 2013), pp.61-96.

Ambler,K., Aycinena D., & Yang D. 2015. "Channeling Remittances to Education: A Field Experiment among Migrants from El Salvador." American Economic Journal: Applied Economics, 7(2): 207-32.

Attanasio, O., Fitzsimons, E., Grantham-McGregor, S., Meghir, C., & Rubio-Codina, M. 2012. Early Childhood Stimulation, Micronutrient Supplementation and Child Development: A Randomised Control Trial. London, UK: Center for the Evaluation of Development Policies (EdePo) at Institute for Fiscal Studies (IFS).

Aturupane, H., P. Glewwe, T. Keeleghan, R. Ravina, U. Sonnadara, and S. Wisniewski. 2013. "An Impact Evaluation of Sri Lanka's Policies to Improve the Performance of Schools and Primary School Students through its School Improvement and School Report Card Programs", World Bank.

Baird, S, McIntosh C., and Özler B. 2011. Cash or Condition? Evidence from a Cash Transfer

Experiment. The Quarterly Journal of Economics, 126, 1709–1753.

Baird, S., Hicks J.H., Kremer M., & Miguel E. 2016, Worms at work: Long-run impacts of a child health investment. The Quarterly Journal of Economics, 131(4), 1637-1680.

Banerjee, A, Duflo, E and Gallego, F. 2016. Removing barriers to higher education in Chile: evaluation of peer effects and scholarships for test preparation, 3ie Impact Evaluation Report 36. New Delhi: International Initiative for Impact Evaluation (3ie)

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., Walton, M. 2016. Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India. Unpublished manuscript. Massachusetts Institute of Technology (MIT), Cambridge, MA

Banerjee, A., Cole S., Duflo E., Linden L. 2007. Remedying Education: Evidence from Two Randomized Experiments in India. The Quarterly Journal of Economics, 122(3), 1235-1264.

Banerji, R., Berry, J., & Shotland, M. 2013. The Impact of Mother Literacy and Participation Programs on Child Learning: Evidence from a Randomized Evaluation in India. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (J-PAL).

Barr, A., Mugisha, F., Serneels, P., & Zeitlin, A. 2012. Information and collective action in the community monitoring of schools: Field and lab experimental evidence from Uganda. Unpublished manuscript, University of Nottingham.

Barrera-Osorio, F., & Filmer, D. 2015. Incentivizing Schooling for Learning: Evidence on the Impact of Alternative Targeting Approaches. Journal of Human Resources.

Barrera- Osorio, Felipe, and Leigh L. Linden. 2009. The use and misuse of computers in education: Evidence from a randomized experiment in Colombia. The World Bank, Policy Research Working Paper Series

Barrera-Osorio, F., & Raju, D. 2014. Teacher Performance Pay: Experimental Evidence from Pakistan. Washington, DC: The World Bank.

Barrera-Osorio, F., Blakeslee, D. S., Hoover, M., Linden, L., Raju, D., & Rya, S. P. 2017. Delivering Educaiton to the Underserved through a Public-Private Partnership Program in Pakistan. NBER working paper 23870. Cambridge, MA: National Bureau of Economic Research (NBER).

Barrera-Osorio, F., Galbert De, P., Habyarimana, J. and Sabarwal, S., 2015. The impact of public-private partnerships on private school performance: Evidence from a randomized controlled trial in Uganda. Working paper.

Bassi, Marina, Ana Reynoso and Costas Meghir. 2016. Education Quality and Teaching Practices. NBER Working Paper Np. 22719.

Beasley, E and Huillery, Élise. 2016. Willing but Unable: Short-Term Experimental Evidence on Parent Empowerment and School Quality .The World Bank Economic Review

Beg, S., Fitzpatrick, A., Lucas, A., Tsinigo, E., & Atimone, H. 2020. Strengthening teacher accountability to reach all students (STARS). IPA Ghana.

Bellei, C. 2009. Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile. Economics of Education Review 28 (5): 629-40

Benhassine, Najy, Florencia Devoto, Esther Duflo, Pascaline Dupas, and Victor Pouliquen. 2015. “Turning a Shove into a Nudge? A ”Labeled Cash Transfer” for Education.”American Economic Journal: Economic Policy 7(3): 86-125.

Benjamin Piper, Wendi Ralaingita, Linda Akach & Simon King. 2016. Improving procedural and conceptual mathematics outcomes: evidence from a randomised controlled trial in Kenya, Journal of Development Effectiveness, 8:3, 404-422.

Berlinski, S., & Busso, M. 2017.Pedagogical Change in Mathematics Teaching: Evidence from a Randomized Control Trial. Economics Letters, V156, 172-175

Berlinski, Samuel, Matias Busso, Taryn Dinkelman and Claudia Martinez A. 2016. Reducing parental-school information gaps and improving education outcomes: Evidence from high frequency text messaging in Chile. Working paper.

Berlinski, Samuel, Sebastian Galiani, and Marco Manacorda. 2008. “Giving Children a Better Start: Preschool Attendance and School-Age Profiles.” Journal of Public Economics 92(5): 1416–40. https://doi.org/10.1016/j.jpubeco.2007.10.007.

Berlinski, Samuel, Sebastian Galiani, and Paul Gertler. 2009. “The Effect of Pre-Primary Education on Primary School Performance.” Journal of Public Economics 93(1): 219–34. https://doi.org/10.1016/j.jpubeco.2008.09.002.

Berkes, Jan Lukas; Bouguen, Adrien; Filmer, Deon P.; Fukao, Tsuyoshi. 2019. Improving Preschool Provision and Encouraging Demand : Heterogeneous Impacts of a Large-Scale Program (English). Policy Research working paper; no. WPS 9070; Impact Evaluation series. Washington, D.C. : World Bank Group.

Beuermann, D. W., Cristia, J. P., Cruz-Aguayo, Y., Cueto, S., & Malamud, O. 2015.One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru. American Economic Journal: Applied Economics, 7(2), 53-80.

Bernal, R., Attanasio, O., Peña, X., & Vera-Hernández, M. 2019. The effects of the transition from home-based childcare to childcare centers on children’s health and development in Colombia. Early childhood research quarterly, 47, 418-431.

Blimpo, M. P. 2014. Team incentives for education in developing countries: A randomized field experiment in Benin. American Economic Journal: Applied Economics, 6(4), 90-109. doi:10.1257/app.6.4.90

Blimpo, M., Evans, D. and Lahire, N., 2015. Parental human capital and effective school management: Evidence from the Gambia. Washington D.C.: World Bank.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng’ang’a, and Justin Sandefur. 2013. March. Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education. The Center for Global Development, Working Paper no. 321.

Bobonis, G. J., Miguel, E., & Puri-Sharma, C. 2006. Anemia and School Participation. Journal of Human Resources, 41, 692-721.

Borkum, E., He, F., & Linden, L. L. 2013. School libraries and language skills in Indian primary schools: A randomized evaluation of the Akshara Library program. Unpublished manuscript. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA.

Brooker, Simon, Hellen Inyega, Benson Estambale, Kiambo Njagi, Elizabeth Juma, Caroline Jones, Catherine Goodman, Matthew Jukes. 2013. Impact of malaria control and enhanced literacy instruction on educational outcomes among Kenyan school children: a multi-sectoral, prospective, randomized evaluation. Draft Grantee Final Report. 3ie.

Bruns, B., Costa, L., & Cunha, N. 2018. Through the looking glass: Can classroom observation and coaching improve teacher performance in Brazil? Economics and Education Review, 64(2018), 214-250.

Buhl-Wiggers, Julie, Jason T. Kerwin, Jeffrey A. Smith and Rebecca Thornton. 2017.The impact of teacher effectiveness on student learning in Africa. RISE Conference Paper.

Burde, D., & Linden, L. L. 2013. Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. American Economic Journal: Applied Economics, 5(3), 27–40. doi:10.1257/app.5.3.27

Chaudhury, N., Friedman, J. and Onishi, J., 2013. Philippines conditional cash transfer program impact evaluation 2012. Report.

Chen, X., Shi, Y., Yi, H., Zhang, D. M., Chu, J., Loyalka, P. and Rozelle, S. 2013. The impact of a senior high school tuition relief program on poor junior high school students in rural China. China and World Economy, V21(3), 80-97.

Cheung, M. and Berlin, M. P., 2014. The impact of a food for education program on schooling in Cambodia. Asia & the Pacific Policy Studies, 2, 44–57.

Chinen, M., Bos, J.M., 2016. Final Report for the Impact Evaluation of the Save the Children Early Childhood Stimulation Program in Bangladesh. Washington, DC: American Institutes for Research.

Chitolina, L., Foguel, M. N., & Menezes-Filho, N. 2013. The Impact of the Expansion of the Bolsa Família Programme on the Time Allocation of Youths and Their Parents. Rev. Bras. Econ. vol.70 no.2 Rio de Janeiro Apr./June 2016

Chyi, H. and Zhou, B., 2014. The effects of tuition reforms on school enrollment in rural china. Economics of Education Review, 38, 104-23.

Contreras, D., & Rau, T. 2012. Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile. Economic development and cultural change, 61(1), 219-246.

Correa, J. A., Parro, F. and Reyes, L., 2014. The effects of vouchers on school results: Evidence from Chile's targeted voucher program. Journal of Human Capital, 8, 352-398.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. 2012."Technology and child development: Evidence from the One Laptop per Child program" (IZA Discussion Paper No. 6401, Forschungsinstitut zur Zukunft der Arbeit GmbH [IZA], Bonn, Germany)."

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik x', and Venkatesh Sundararaman. 2013. School Inputs, Household Substitution, and Test Scores. American

Economic Journal: Applied Economics, 5(2), 29–57.

De Brauw, A., Gilligan, D. O., Hoddinott, J. and Roy, S., 2014. The impact of Bolsa Família on schooling. IFPRI

De Janvry, A., Finan, F. and Sadoulet, E., 2012. Local electoral incentives and decentralized program performance. Review of Economics and Statistics, 94, 672-685.

De Melo, G., Machado, A. and Miranda, A., 2014. The impact of a one laptop per child program on learning: Evidence from Uruguay. Bonn Germany: The Institute for the Study of Labor (IZA).

De Ree, J., Muralidharan, K., Pradhan, M., & Rogers, F. H. 2016. Double for nothing? Experimental evidence on the impact of an unconditional teacher salary increase on student performance in Indonesia. Quarterly Journal of Economics (forthcoming).

Diagne, A., Lô, M. M., Sokhna, O. and Diallo, F. L., 2014. Evaluation of the impact of school canteen programs on internal efficiency of schools, cognitive acquisitions and learning capacities of students in rural primary schools in Senegal. Nairobi: PEP working paper.

Dillon, Moira.R, Harina Kannan, Joshua T. Dean, Elizabeth S. Spelke, Esther Duflo, 2017. Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. Science 357, 47-55.

Dinkelman, T., and Martínez, A. C. (2013). Investing in Schooling In Chile: The Role of Information about Financial Aid for Higher Education. Review of Economics and Statistics, 96, 244-257.

Dizon-Ross, R. 2016. Parents' perceptions and children's education: Experimental evidence from Malawi. Unpublished manuscript. Massachusetts Institute of Technology (MIT), Cambridge, MA.

Duflo, E., Berry, J., Mukerji, S., & Shotland, M. 2015. A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India. (Impact Evaluation Report No. 22). International Initiative for Impact Evaluation (3ie), New Delhi, India.

Duflo, E., Hanna, R., & Ryan, S. P. 2012. Incentives Work: Getting Teachers to Come to School. American Economic Review, 102(4), 1241–1278. doi:10.1257/aer.102.4.1241

Duflo, E., Dupas, P., Kremer, M. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." American Economic Review 101 (5): 1739 -1774.

Duflo, E., Dupas P., and Kremer M. 2015. "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools." Journal of Public Economics, 123, March 2015, 92 - 110.

Duflo, Esther, Pascaline Dupas and Michael Kremer. 2017. The impact of free secondary education: Experimental evidence from Ghana. Working paper

Ebenezer, R., Gunawardena, K., Kumarendran, B., Pathmeswaran, A., Jukes, M. C. H., Drake, L. J. and Silva, N. D., 2013. Cluster-randomised trial of the impact of school-based deworming and iron supplementation on the cognitive abilities of schoolchildren in Sri

Lanka's plantation sector. Tropical Medicine and International Health, 18, 942-951.

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.

Edmonds Eric, V. and Schady, N., 2012. Poverty alleviation and child labor. American Economic Journal-Economic Policy, 4, 100-124.

Edmonds, E. V. and Shrestha, M., 2014. You get what you pay for: Schooling incentives and child labor. Journal of Development Economics, 111, 196-211.

Evans, D., Hausladen, S., Kosec, K. and Reese, N., 2014. Community-based conditional cash transfers in Tanzania: Results from a randomized trial. Report.

Eyal, K. and Woolard, I., 2014. Cash transfers and teen education: Evidence from South Africa. Report.

Fang, H., Eggleston, K. N., Rizzo, J. A., Rozelle, S., & Zeckhauser, R. J. 2012. The returns to education in China: evidence from the 1986 compulsory education law (No. w18189). National Bureau of Economic Research.

Ferré, C. and Sharif, I., 2014. Can conditional cash transfers improve education and nutrition outcomes for poor children in Bangladesh? Evidence from a pilot project. Evidence from a pilot project. Working Paper.

Filmer, D., & Schady, N. 2014. The Medium-Term Effects of Scholarships in a Low-Income Country. Journal of Human Resources, 49(3), 663-694.

Fiorella Benedetti, Pablo Ibarrarán, and Patrick J. McEwan. 2016.Do Education and Health Conditions Matter in a Large Cash Transfer? Evidence from a Honduran Experiment. Economic Development and Cultural Change 64, no. 4 (July 2016): 759-793.

Gajigo, S. 2016. Closing the education gender gap: Estimating the impact of girls' scholarship program in the Gambia. Education Economics. 24(2).

Galiani, S., & McEwan, P. 2013. The Heterogeneous Impact of Conditional Cash Transfers. Journal of Public Economics, 103, 85-96.

Gallego, F. A., Näslund-Hadley E., Alfonso M. 2019. Changing Pedagogy to Improve Skills in Preschools: Experimental Evidence from Peru. The World Bank Economic Review.

Garlick, R., 2013. How price sensitive is primary and secondary school enrollment? Evidence from nationwide tuition fee reforms in South Africa. Unpublished working paper.

Garn, J. V., Greene, L. E., Dreibelbis, R., Saboori, S., Rheingans, R. D. and Freeman, M. C., 2013. A cluster-randomized trial assessing the impact of school water, sanitation and hygiene improvements on pupil enrolment and gender parity in enrolment. Journal of Water, Sanitation and Hygiene for Development, 3 (4), 592-601.

Gertler, P. J., Patrinos, H. A., & Rubio-Codina, M. 2012. Empowering parents to improve education: Evidence from rural Mexico. Journal of Development Economics, 99(1), 68-79.

Glewwe, P. and Kassouf, A. L., 2012. The impact of the Bolsa Escola/Familia conditional cash transfer program on enrollment, dropout rates and grade promotion in Brazil. Journal of Development Economics, 97, 505-517.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. Many Children Left Behind? Textbooks and Test Scores in Kenya. American Economic Journal: Applied Economics, 1(1), 112–135.

Glewwe, P. and Maiga, E. W. H. 2011. The impacts of school management reforms in Madagascar: Do the impacts vary by teacher type? Journal of Development Effectiveness, 3, 435-469.

Glewwe, P., Nauman Ilias and Michael Kremer. 2010. Teacher incentives. American Economic Journal: Applied Economics, 2(3), 205-227.

Gutiérrez, E. and Rodrigo, R., 2014. Closing the achievement gap in mathematics: Evidence from a remedial program in Mexico City. Latin American Economic Review, 23:14, 1-30.

Hasan, A., Jung, H., Kinnell, A. Maika, A., Nakajima, N., & Pradhan, M. 2019. Contrasting experiences: Understanding the longer-term impact of improving access to preschool education in rural Indonesia. Washington, D.C.: World Bank.

Halliday, K. E., Okello, G., Turner, E. L., Njagi, K. and Mcharo, C., 2014. Impact of intermittent screening and treatment for malaria among school children in Kenya: A cluster randomised trial. PLOS Medicine, 11, 1-16.

Hidalgo, D., Onofa, M., Oosterbeek, H. and Ponce, J., 2013. Can provision of free school uniforms harm attendance? Evidence from Ecuador. Journal of Development Economics, 103, 43–51.

Hirshleifer, S. (2015). Incentives for effort or outputs? A field experiment to improve student performance Unpublished manuscript. San Diego, CA: University of California at San Diego.

Hoekstra, Mark, Perre Mouganie, Yaojing Wang. 2016. Peer quality and the academic benefits to attending better schools. NBER Working Paper 22337.

Hojman, A., & Boo, F. L. 2019. Cost-effective public daycare in a low-income economy benefits children and mothers. IDB Working Paper Series, 1036. Washington D.C.: Inter-American Development Bank.

Huan, Wang, James Chu, Prashant Loyalka, Xin Tao, Yaojiang Shi, Qinghe Qu, Chu Yang and Scott Rozelle. 2014. “Can School Counseling Reduce School Dropout in Developing Countries?” REAP Working Paper #275

Hulett, J. L., Weiss, R. E., Bwibo, N. O., Galal, O. M., Drorbaugh, N. and Neumann, C. G., 2014. Animal source foods have a positive impact on the primary school test scores of Kenyan schoolchildren in a cluster-randomised, controlled feeding intervention trial. British Journal of Nutrition, 111, 875-86.

Humpage, S., 2013. Essays on child development in developing countries. Dissertation Submitted to the Faculty of University of Minnesota in Partial fulfilment of the Requirements for the Degree of Doctor of Philosophy.

Imbrogno, J., 2014. Essays on the economics of education. Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy Economics from Carnegie Mellon University.

Islam, Asad. 2019. Parent-Teacher meetings and student outcomes: Evidence from a developing country. Working Paper.

Jackson, C. K. 2013. Can Higher-Achieving Peers Explain the Benefits to Attending Selective Schools? Evidence from Trinidad and Tobago. Journal of Public Economics, 108, 63-77.

Jayaraman, R. and Simroth, D., 2015. The impact of school lunches on primary school enrollment: Evidence from India's midday meal scheme. Scandinavian Journal of Economics, 117, 1176–1203.

Jere R. Behrman, Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools," Journal of Political Economy 123, no. 2 (April 2015): 325-364.

Jensen, R. 2010. The (Perceived) Returns to Education and the Demands for Schooling. Quarterly Journal of Economics 125(2): 515-548.

Karlan, D., & Linden, L. L. 2013. Loose Knots: Strong Versus Weak Commitments to Save for Education in Uganda NBER Working Paper No. 19863. Cambridge, MA: National Bureau of Economic Research (NBER).

Kazianga, H., Levy, D., & Linden, L. L. 2013. The Effects of "Girl-Friendly" Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. American Economic Journal: Applied Economics, 5(3), 41-62.

Kerwin, J. T. and Thornton, R., 2015. Making the grade: Understanding what works for teaching literacy in rural Uganda.

Kleiman - Weiner, M., Luo, R., Zhang, L., Shi, Y., Medina, A., & Rozelle, S. (2013). Eggs versus chewable vitamins: Which intervention can increase nutrition a nd test scores in 2 rural China? China Economic Review, 24, 165 - 176. doi: 10.1016/j.chieco.2012.12.005

Kremer, M., Glewwe P., Moulin S., Zitzewitz E. 2004. Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya. Journal of Development Economics. 74(1), 251-268.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. Incentives to Learn. The Review of Economics and Statistics, 91(3) 437-456.

Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. 2015. Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. Economics of Education Review, 47, 34-48. doi:10.1016/j.econedurev.2015.03.005

Lai, F., Zhang, L., Hu, X., Qu, Q., Shi, Y., Qiao, Y., ... & Rozelle, S. 2013. Computer assisted learning as extracurricular tutor? Evidence from a randomised experiment in rural boarding schools in Shaanxi. Journal of Development Effectiveness, 5(2), 208-231.

Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M. and Rozelle, S., 2015, Teaching the language of wider communication, minority students and overall educational performance: Evidence from a randomized experiment in Qinghai province, china. Economic Development and Cultural Change 2015 63:4, 753-776

Lakshminarayana, R., A. Eble, P. Bhakta, C. Frost, P. Boone, D. Elbourne and V. Mann. 2013.

"The Support to Rural India's Public Education System (STRIPES) Trial: A Cluster Randomised Controlled Trial of Supplementary Teaching, Learning Material and Material Support." PLOS ONE 8, no. 7.

Li, T., Han, L., Zhang, L. and Rozelle, S., 2014. Encouraging classroom peer interactions: Evidence from Chinese migrant schools. Journal of Public Economics, 111, 29-45.

Loyalka, P., Liu, C., Song, Y., Yi, H., Huang, X., Wei, J.,...Rozelle, S. 2013. Can information and counseling help students from poor rural areas go to high school? Evidence from China (Working Paper No. 241). Stanford, CA: Rural Education Action Project.

Loyalka, P., Sylvia, S., Liu, C., Chu, J. and Shi, Y., 2015. Pay by design: Teacher performance pay design and the distribution of student achievement. Working Paper.

Loyalka, Prashant, Anna Popova, Guirong Li, Chengfang Liu and Henry Shi. 2017. Unpacking Teacher Professional Development. Working Paper

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. 2014. Improving early grade literacy in East Africa: Experimental evidence from Kenya and Uganda. Journal of Policy Analysis and Management, 33, 950–976.

Martinez, Sebastian, Sophie Naudeau, and Vitor Pereira. 2017. Preschool and Childhood Development Under Extreme Poverty: Evidence from a Randomized Experiment in Rural Mozambique. Policy Research Working Paper 8290. World Bank.

Matthew C. H. Jukes, Elizabeth L. Turner, Margaret M. Dubeck, Katherine E. Halliday, Hellen N. Inyega, Sharon Wolf, Stephanie Simmons Zuilkowski & Simon J. Brooker. 2016. Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2016.1221487

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. 2019. Inputs, incentives, and complementarities in primary education: Experimental evidence from Tanzania. Quarterly Journal of Economics, 134(3): 1627-1673.

Melissa Adelman, Peter Holland, and Tillmann Heidelk. 2017. "Increasing Access by Waiving Tuition: Evidence from Haiti," Comparative Education Review 61(4): 804-831.

Miguel, E. & Kremer, M. 2003. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Econometrica 72(1):159-217

Mo, D, Bai, Y, Boswell, M and Rozelle, S, 2016, Evaluating the effectiveness of computers as tutors in China, 3ie Impact Evaluation Report 41. New Delhi: International Initiative for Impact Evaluation (3ie)

Mo, D., Zhang, L., Lui, R., Qu, Q., Huang, W., Wang, J.,...Rozelle, S. 2013. Integrating computer assisted learning into a regular curriculum: E vidence from a randomized experiment in rural schools in Shaanxi (Working Paper 248). Stanford, CA: Rural Education Action Project.

Mo, D., Zhang, L., Wang, J., Huang, W., Shi, Y., Boswell, M., and Rozelle, S. 2015. Persistence of learning gains from computer assisted learning: Experimental evidence from China. Journal of Computer Assisted Learning, 31: 562–581. doi: 10.1111/jcal.12106.

Morabito, Christian, Dirk Van de gaer, Jose Luis Figueroa and Michel Vandenbrouck. 2017. Effects of high versus low quality preschool: A longitudinal study in Mauritius. CSAE Conference 2017.

Muralidharan, K., & Prakash, N. 2013. Cycling to School: Increasing Secondary School Enrollment for Girls in India NBER Working Paper No. 19305. Cambridge, MA: National Bureau of Economic Research (NBER).

Muralidharan, K., & Sundararaman, V. 2013. The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India NBER Working Paper No. 19441. Cambridge, MA: National Bureau of Economic Research (NBER).

Muralidharan, K., Singh, A., & Ganimian, A. J. 2019. Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. American Economic Review, 109 (4): 1426-60.

Nguyen, T. 2008. Information, Role Models and Perceived Returns to Education Experimental Evidence from Madagascar. Working Paper. https://www.povertyactionlab.org/sites/default/files/documents/Nguyen%202008.pdf

Olken, Benjamin A., Junko Onishi, and Susan Wong. 2014. "Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia." American Economic Journal: Applied Economics 6(4): 1-34.

Oster, Emily, and Rebecca Thornton. 2011. "Menstruation, Sanitary Products, and School Attendance: Evidence from a Randomized Evaluation." American Economic Journal: Applied Economics, 3(1): 91-100.

Özler, Berk, Lia C.H. Fernald, Patricia Kariger, Christin McConnell, Michelle Neuman and Eduardo Fraga. 2018. Combining Preschool Teacher Training with Parenting Education. A Cluster-Randomized Controlled Trial. Journal of Development Economics. 133:448-467

Park A., Zhao M., & Glewwe P. 2016. A better vision for development: Eyeglasses and academic performance in rural primary schools in China Journal of Development Economics Volume 122, September 2016, Pages 170–182

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

Piper, B. and Mugenda, A. 2014. The primary math and reading (PRIMR) initiative: Endline impact evaluation. Research Triangle Park, NC: RTI International.

Piper, Benjamin and Medina Korda. 2011. EGRA Plus: Liberia, Program Evaluation Report. Produced for review by the United States Agency for International Development, Prepared by RTI International and the Liberian Education Trust.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha. 2014. "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia." American Economic Journal: Applied Economics, 6(2): 105-26.

Premand, Patrick, Oumar Barry, and Marc Smitz, 2016. Cash transfers, Value-Added of Behavioral Change Accompanying Measures and Early Child Development in Niger. Short-Term Impact Evaluation Policy Report: Washington, DC: World Bank.

Pop-Eleches, C., & Urquiola, M. 2013. Going to a Better School: Effects and Behavioral Responses. American Economic Review, 103, 1289-1324.

RTI International, 2015. The primary math and reading (PRIMR) initiative: DFID/Kenya rural expansion programme. Bungoma and Machakos endline study. RTI International.

Robertson, L., Mushati, P., Eaton, J. W., Dumba, L., Mavise, G., Makoni, J., Schumacher, C., Crea, T., Monasch, R., Sherr, L., Garnett, G. P., Nyamukapa, C. and Gregson, S., 2013. Effects of unconditional and conditional cash transfers on child health and development in Zimbabwe: A cluster-randomised trial. Lancet, 381, 1283-1292.

Romero, M., J. Sandefur, W. A. Sandholtz. 2020. Outsourcing Education: Experimental Evidence from Liberia. American Economic Review 110(2), 364-400.

Ricardo Sabates, Pauline Rose, Benjamin Alcott & Marcos Delprato. 2021. Assessing cost-effectiveness with equity of a programme targeting marginalised girls in secondary schools in Tanzania, Journal of Development Effectiveness, 13:1, 28-46, DOI: 10.1080/19439342.2020.1844782

Sabarwal, S., Evans, D., & Marshak, A. 2014. The Permanent Textbook H ypothesis: School Inputs and Student Outcomes in Sierra Leone. Washington, DC: The World Bank.

Santibanez, L., Abreu-Lastra, R. and O'donoghue J., L., 2014. School based management effects: Resources or governance change? Evidence from Mexico. Economics of Education Review, 39, 97-109.

Sean, S., Renfu, L., Linxiu, Z., Yaojiang, S., Alexis, M. and Scott, R., 2013. Do you get what you pay for with school-based health programs? Evidence from a child nutrition experiment in rural china. Economics of Education Review, 37, 1-12.

Seid, Yared. 2017. The impact of learning in mother tongue first: Evidence from a natural experiment in Ethiopia. RISE conference paper.

Sharma, U. 2014. Can computers increase human capital in developing countries? An evaluation of Nepal's one laptop per child program. In 2014 Annual Meeting, July 27-29, 2014, Minneapolis, Minnesota (No. 169846). Agricultural and Applied Economics Association, 2014 2014 Minneapolis, Minnesota USA.

Schultz, T. P. 2000. Impact of PROGRESA on school attendance rates in the sampled population. Report submitted to PROGRESA. International Food Policy Research Institute, Washington, D.C.

Shultz, T. P. 2004. School subsidies for the poor: Evaluating the Mexican Progresa poverty program. Journal of Development Economics, 74(1), 199–250 [Special Issue June 2004].

Spier, E., Kamto, K, Molotsky, A., Rahman, A. Hossain, N., Nahar, Z. & Khondker, H. 2020. Bangladesh early years preschool program impact evaluation: Endline report for the World Bank Strategic Impact Evaluation Fund. Washington, D.C.: AIR.

Spratt, Jennifer, Simon King, and Jennae Bulat. 2013. June. Independent Evaluation of the Effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) Program in Mali. RTI International. Endline Report.

Visaria, Sujata, Rajeev Dehejia, Melody M. Chao and Anirban Mukhopadhyay. 2016. Unintended consequences of rewards for student attendance: Results from a field experiment in Indian classrooms. NBER Working Paper 22528

Wong, H. L., Luo, R., Zhang, L., & Rozelle, S. 2013. The impact of vouchers on preschool attendance and elementary school readiness: A randomized controlled trial in rural China. Economics of Education Review, 35, 53-65.

Wong, H. L., Shi, Y. J., Luo, R. F., Zhang, L. X. and Rozelle, S., 2014. Improving the health and education of elementary schoolchildren in rural china: Iron supplementation versus nutritional training for parents. Journal of Development Studies, 50, 502-519.

Yamauchi, F., 2014. An alternative estimate of school-based management impacts on students' achievements: Evidence from the Philippines.Journal of Development Effectiveness, 6(2), 97-110.

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Arbour, M. C., Barata, M. C., Weiland, C., Gómez, C., Moreno, L., Rolla, A., & D'Sa, N. 2015. Experimental Impacts of a Teacher Professional Development Program in Chile On Preschool Classroom Quality and Child Outcomes. Journal of Developmental Psychology.