Essays in Development Economics, Human Capital and External Validity

Noam Angrist

Abstract

This thesis comprises three papers under the overarching theme of external validity – a question of how the effectiveness of interventions or policies translate across contexts. I focus on interventions that aim to improve human capital, specifically health and education, in developing countries. To address the topic of external validity, we need comparable interventions and outcomes as well as identification of causal mechanisms that drive intervention effectiveness across contexts. My three papers contribute to each aspect. Two of my papers focus on constructing and comparing outcomes to measure education and human capital globally. In a third paper, I conduct a large-scale randomized controlled trial of a sex education intervention in Botswana in a third of the country. This trial builds on a prior trial showing a similar intervention reduced teenage pregnancy by 28% in Kenya ten years prior. The study in Botswana provides replicable evidence across contexts, showing reductions in pregnancy of up to 40%. The trial also deepens understanding of the underlying mechanism driving behavior by testing novel dimensions, such as the degree to which the messenger delivering the sex education message matters. I find the messenger makes-or-breaks the intervention: while near-peer educators reduce pregnancy, government teachers have a null effect and appear to induce a teenage rebellion response. This contrast is striking, suggesting that an underexplored dimension on which external validity turns – the messenger – might matter more than dimensions which traditionally get more attention, such as time and geography.

1 Introduction

This thesis comprises three papers under the theme of external validity – a question of how the effectiveness of interventions or policies translate across contexts. I focus on external validity for interventions that aim to improve human capital, namely in health and education, in developing countries. To engage with this topic well, we need comparable outcomes, comparable interventions, and identification of causal mechanisms between interventions and outcomes across contexts. My papers contribute to each aspect. Two of my papers focus on globally comparable outcomes to measure human capital. In my main paper, I conduct a large-scale randomized controlled trial of a sex education intervention in Botswana. This trial builds on a prior trial showing a similar intervention reduced risky sex in Kenya ten years prior. My trial both provides replicable evidence across contexts as well as deepens understanding of the underlying mechanism driving behavior by testing novel dimensions, such as the messenger delivering the sex education message.

In my main paper, "The Messenger Matters: Experimental Evidence from a Sex Education Intervention," I run a large-scale randomized controlled trial evaluating whether a sex education intervention shown to reduce risky sex in Kenya in 2003-2004 can work in Botswana in 2014-2015. This is one of the largest sex education trials in the literature, and covers 42,000 students across a third of Botswana. We collect objective data on teenage pregnancy and sexual behavior with near universal follow-up 12-15 months later, enabling reliable detection of risky sexual behavior as well as exploration of detailed mechanisms of behavior change. The Botswana and Kenya studies have near-identical interventions, outcomes and test whether the causal relationship translates across diverse contexts. We find that results translate across countries and over time, with a reduction in pregnancy of 40% in Botswana relative to 28% in Kenya. However, the messenger delivering the sex education message within Botswana makes-or-breaks the intervention. While peer educators reduce pregnancy, government teachers have a null effect and plausibly induce a teenage rebellion response, providing some of the first causal evidence that the sender of information can be first-order for information to result in behavior change. This contrast is striking, suggesting that an underexplored dimension on which external validity turns – the messenger – might matter more than dimensions that traditionally get more attention, such as time and geography.

My results have significant implications for policy. Teachers are often the default messenger for sex education. Teachers are respected figures in the community and a fixture of the education system. However, viable alternatives exist to deliver sex education. For example, many countries have national service programs which deploy young people in schools. AmeriCorps in the United States places young volunteers in over 12,000 schools throughout

the country. In sub-Saharan Africa, over fifteen countries have national service programs with cumulative government investments of over $240 million per year. This presents a model for near-peers to deliver sex education in schools. Our results suggest that while the current default messenger for sex education, teachers, can be effective for transferring information, they might not reduce risky sex. In contrast, near-peers can reduce pregnancy by over 40 percent. Leveraging near-peers to deliver sex education could thus yield large welfare gains. This paper has been drafted to target economics journals.

In my second paper, "Measuring Human Capital," I focus on construction of comparable outcomes of human capital. One of the best possible measures of human capital is learning. However, to date no globally comparable data exists on learning. Instead, for decades, economists have used measures of enrollment in school to measure human capital since it was available and comparable across countries. However, schooling does not necessarily yield learning, making this an imperfect proxy. I bridge this gap, building a database of comparable learning outcomes for 164 countries from 2000-2017. In a first application of the data, I present several stylized facts: (a) while enrollment has increased worldwide, learning has stagnated (b) girls outperform boys on learning – a positive gender gap – in contrast to a negative gender gap observed for schooling (c) learning is associated with growth on a global scale (d) associations with growth are heterogenous (e) human capital accounts for up to a third of cross-country income differences – a middle ground in the recent development accounting literature. These stylized facts demonstrate the potential of the data to reveal new insight into the relationship between human capital and economic development. This paper was co-authored with colleagues at the World Bank, including Chief Economist Pinelopi Goldberg, Simeon Djankov, Director of Development Economics, and Harry Patrinos, Practice Manager. My role in the paper was central and I am first author. The dataset will be public and updated annually, and is a core ingredient in the World Bank's recently launched Human Capital Index. This paper has been drafted to target economics journals.

In my third paper, "A New Micro Measure of Education Interventions: Learning-Adjusted Years of Schooling (LAYS)", I create a unified outcome metric of human capital to express impacts of interventions and policies. Learning-Adjusted Years of Schooling combine the quantity and quality of schooling into a single measure of educational progress. In this paper, I argue this measure is better than either learning or enrollment alone. Whereas a measure of schooling might crowd out a focus on learning, and a measure of learning might incentivize restricting school access to the highest performers, LAYS keeps the focus on the quality of schooling and encourages schooling participation of all children. This measure is another core ingredient in the World Bank's Human Capital Index. This measure can further

be used to express results of interventions and policies on a unified, easy-to-understand and incentive-compatible manner. This paper demonstrates the potential for LAYS to become the education analogy to two standard outcome units in health: Quality-Adjusted Life Years (QALYs) and Disability-Adjusted Life Years (DALYs). This paper builds on a series of collaborative efforts with colleagues Deon Filmer, David Evans, Rachel Glennerster, Harry Patrinos, Halsey Rogers, and Shwetlena Sabarwal to produce global public goods in education, including globally comparable test scores, prior work developing the macro Learning-Adjusted Years of Schooling measure at the country-level and policy applications for the World Bank and the Department for International Development (DFID). This paper has been drafted to target general audience journals such as Science or the Lancet.

Each article has large-scale policy ramifications. Two papers produce global public goods in the form of comparable educations outcomes and have been adopted by the World Bank, DFID and USAID. The "Messenger Matters" paper has potential to inform sex education delivery globally to engage near-peer messengers who might reduce pregnancy by up to 40 percent and thus yield large welfare gains over status quo sex education delivery by teachers.

2 External Validity Literature Review

The ideal parameter of interest for a policy-maker who wants to scale an intervention or policy is an estimate of causal impact that applies in their context. For example, consider a policy-maker in Botswana, who wants to scale-up a remedial education intervention shown in a randomized controlled trial to improve literacy and numeracy in India. This policy problem can be formulated as follows:

$$Y_A = \alpha + \beta_A X_A + \epsilon \tag{1}$$

where β_A is the causal effect that intervention X will have on outcome Y in context A, denoted by the subscript, and indexed by a bundle of characteristics such as geography, time, and implementor. We examine the case where β_A is informed by the causal relationship between the intervention and outcome in a different context B:[1]

[1] There are alternative methods to inform relationship (1) with varying degrees of informativeness. For example, given randomized trials and quasi-experiments do not exist for every policy and intervention, one could consider using information on an observational relationship within the same context A, denoted as follows: $Y = \alpha + \tilde{\beta}_A X_A + \epsilon$ where $\tilde{\beta}$ indicates a correlation rather than a causal relationship β. While seminal work by LaLonde (1986) demonstrated substantial bias between correlations and causations, Pritchett and Sandefur

$$Y_B = \alpha + \beta_B X_B + \epsilon \tag{2}$$

In order to answer this question, we must consider similar interventions (X) and outcomes (Y). We also need to identify the causal relationship β across contexts A and B.

Establishing a credible causal relationship between X and Y, referred to as 'internal validity,' has received significant attention across the social sciences. The concept of internal validity has been characterized and formalized for decades (Rubin 1974; Campbell 1986). Moreover, empiricists have developed and applied a suite of methods to achieve internally valid estimates of causal impact. These methods include propensity score matching, regression discontinuity design, instrumental variables estimation, and randomized controlled trials (Angrist and Pischke 2010). Equipped with a formalization of the problem posed by internal validity and methods to solve for it, a growing literature has emerged identifying if certain programs and policies have had causal impact.

A growing literature has also emerged on extrapolation of causal estimates across contexts A and B, referred to as 'external validity.' Scientific progress and scalable evidence-based policy both turn on the ability to apply and generalize evidence from one study to a new target population of interest. The need for both internally and externally valid evidence has been widely recognized among both theorists and empiricists (Hotz et al. 2005; Heckman and Vytacil 2007; Duflo, Glennerster and Kremer 2007; Pritchett and Sandefur 2013; Allcott 2015; Deaton and Cartwright 2018; Dehejia, Pop-Eleches and Samii 2019).

This recognition is timely. Randomized trials have proliferated and now constitute 30% of published papers in the top 5 journals in development economics (McKenzie 2016). However, causal impact found in one trial does not mean a program always works. Bold et al. (2018) show that when an internally valid estimate for a contract teacher program in Kenya and India was scaled by the government instead of an NGO, it failed to replicate. A recent literature shows significant heterogeneity in effects across contexts including in microcredit (Vivalt 2020; Meager 2019) and energy (Allcott 2015). To this end, the next frontier of research can build on a foundation or internally credible results and focus increasingly on research methods to extrapolate estimates across contexts.

(2013) show in certain cases an observational relationship in the same context can be more informative than a causal estimate from a different context.

The difficulty in extrapolating internally valid results is due to a series of biases which arise when generalizing results from a given sample. These include similar biases which plague internal validity, such as selection bias. In the context of external validity, this is selection bias into context, rather than selection bias into treatment (Heckman and Vytacil 2007; Hotz et al. 2005; Pearl and Barenboim 2014; Allcott 2015). For example, Blair, Iyengar, and Shapiro (2013) examine the distribution of randomized controlled trials across countries and find that such trials are disproportionally conducted in countries with democratic governments. Other challenges include accounting for general equilibrium effects (Acemoglu 2010; Crepon et al 2013), Hawthorne effects (Peters, Langbein and Roberts 2016), and publication bias (Andrews and Kasy 2019).

Thus, despite the proliferation and utility of randomized trials to produce credible causal estimates, the search for "what works" continues. However, it is not unattainable. Banerjee et al. (2017) provide a framework and case study of a remedial education intervention developed in India called "Teaching at the Right Level" which has scaled to over 60 million children across India and Africa: "The issue of context-dependence needs to be addressed through replications, ideally guided by theory. General equilibrium and spillover effects can be addressed by incorporating estimation of these effects into study designs, or by conducting large-scale experiments where the equilibrium plays out. Randomization and piloting bias can be addressed by trying out the programs on a sufficient scale with the government that will eventually implement it, documenting success and failure, and moving from there."

Related solutions have been proposed by numerous authors. Muralidharan and Niehaus (2017) document the median cluster size of a randomized trial as 26 and propose running larger scale trials. Bates and Glennerster (2017) focus on the role of theory-driven experiments, consistent with the Deaton and Cartwright (2018) emphasis on the importance of causal models of human behavior as well as causal relationships between two variables. Another approach involves Bayesian meta-analysis across multiple studies for more generalizable results (Meager 2019). An additional approach includes extrapolation across contexts using explicit weighting of covariates (Hotz et al 2005; Andrews and Oster 2019). In addition to ex-post analysis approaches, an ex-ante design approach includes two-stage randomization, first of context, then of treatment, to enable direct extrapolation of estimates to the broader context as well as quantification of spillovers and general equilibrium effects (Muralidharan and Sundararaman 2015; Davis, Guryan, Hallberg and Ludwig 2017; Crepon et al 2013).

To effectively address external validity, there will also be a need to complement technocratic solutions with disciplinary reporting norms and institutions. Peters, Langbein and Roberts (2016) find in a review of 54 RCTS published in five development economics journals from 2009-2014, that over 96% made generalization claims. However, few examined or discussed potential long-term effects, over 25% did not mention the population sample or potential restrictions to generalizability, and only a fifth described differences between the evaluation and a potential real world set up. If externally valid results are our parameter of interest, it is essential papers report on metrics that approximate the degree of external validity. Various authors have proposed external validity metrics. Pritchett and Sandefur (2013) propose reporting OLS results to approximate the extent of selection into treatment, as well as metrics of error between observational and experimental data. Vivalt (2020) proposes metrics of heterogeneity, such as the variance, and the proportion of variance that can be explained within and across studies (R^2 and I^2). Allcott (2015) suggests including descriptive tables comparing the population and the sample and experimental partners versus scale partners, as well as an F-test on treatment effect heterogeneity within sub-sites as a suggestive metric for likely external validity outside the sample. Dehejia, Pop-Eleches and Samii (2019) propose an external validity function which characterizes how prediction error varies in macro level covariate differences. While a suite of metrics has been proposed in the past few years, the jury is still out on the best external validity metrics, and underreporting on external validity continues. This is a high potential area of further research and institutionalization.

Over the last few decades there has been a movement in development towards bottom-up experimentation to complement broader economic theory. This approach has been described as the search for 'best fit' in addition to the search for 'best practice' (Rodrik, 2008; Rothwell, 2005). A continued focus on both what works and what fits will serve both the scientist and the policy-maker well.

In this thesis, I aim to contribute to the effort to determine what works and fits across contexts. I construct globally comparable education outcomes, a first-order condition to enable comparison across interventions and policies. I further contribute a large-scale replication of a government-backed sex education intervention across a third of Botswana, adapted from an intervention tested ten years prior in Kenya. I incorporate novel experimental dimensions, such as varying both the message and messenger delivering sex education, in order to identify mechanisms to drive effectiveness across contexts.

3 Articles

In this section I include three separate papers, each targeted to an academic journal:

- "The Messenger Matters: Experimental Evidence from a Sex Education Intervention"
- "Measuring Human Capital"
- "A New Micro Measure of Education: Learning-Adjusted Years of Schooling (LAYS)"

References

Andrews, Isaiah, and Emily Oster. 2019. "A simple approximation for evaluating external validity bias." *Economics Letters* 178: 58-62.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and correction for publication bias." *American Economic Review*, *109* (8), 2766-94.

Angrist, Joshua D., and Jörn-Steffen Pischke. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives* 24, no. 2 (2010): 3-30.

Acemoglu, Daron. 2010. "Theory, general equilibrium, and political economy in development economics." *Journal of Economic Perspectives* 24, no. 3 (2010): 17-32.

Allcott, Hunt. 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics* 130, no. 3: 1117-1165

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From proof of concept to scalable policies: Challenges and solutions, with an application." *Journal of Economic Perspectives* 31, no. 4: 73-102.

Bates, Mary Ann, and Rachel Glennerster. 2017. "The generalizability puzzle." *Stanford Social Innovation Review* 15, no. 3: 50-54.

Blair, Graeme, Radha K. Iyengar, and Jacob N. Shapiro. 2013. "Where policy experiments are conducted in economics and political science: The missing autocracies."

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, and Justin Sandefur. 2018. "Experimental evidence on scaling up education reforms in Kenya." *Journal of Public Economics* 168: 1-20.

Campbell, Donald T. 1986. "Relabeling internal and external validity for applied social scientists." *New Directions for Program Evaluation* 1986, no. 31: 67-77.

Cook, Thomas D., Donald Thomas Campbell, and William Shadish. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment." *Quarterly Journal of Economics* 128, no. 2 (2013): 531-580.

Davis, Jonathan, Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. *The economics of scale-up*. No. w23925. National Bureau of Economic Research.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210: 2-21

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2019. "From local to global: External validity in a fertility natural experiment." *Journal of Business & Economic Statistics*: 1-27.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using randomization in development economics research: A toolkit." *Handbook of development economics* 4: 3895-3962.

Heckman, James J., and Edward J. Vytlacil. 2007. "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation." *Handbook of econometrics* 6 (2007): 4779-4874.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125, no. 1-2: 241-270.

LaLonde, Robert J. 1986. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review*: 604-620.

McKenzie, David. 2016. *Have RCTs taken over development economics*? World Bank Development Impact Blog.

Meager, Rachael. 2019. "Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments." *American Economic Journal: Applied Economics* 11, no. 1: 57-91.

Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at scale." *Journal of Economic Perspectives* 31, no. 4: 103-24.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The aggregate effect of school choice: Evidence from a two-stage experiment in India." *The Quarterly Journal of Economics* 130, no. 3: 1011-1066.

Pearl, Judea, and Elias Bareinboim. 2014. "External validity: From do-calculus to transportability across populations." *Statistical Science*: 579-595.

Peters, Jörg, Jörg Langbein, and Gareth Roberts. 2016. "Policy evaluation, randomized controlled trials, and external validity—A systematic review." *Economics Letters* 147: 51-54

Pritchett, Lant, and Justin Sandefur. 2013. "Context matters for size: why external validity claims and development practice do not mix." *Journal of Globalization and Development* 4, no. 2: 161-197.

Rodrik, Dani. 2008. "The new development economics: we shall experiment, but how shall we learn?"

Rothwell, Peter M. 2005. "External validity of randomised controlled trials: "to whom do the results of this trial apply?" *The Lancet* 365, no. 9453: 82-93.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and non-randomized studies." *Journal of educational Psychology* 66, no. 5: 688

Vivalt, Eva. 2020. "Forthcoming."How Much Can We Generalize from Impact Evaluations?"." *Journal of the European Economics Association* 5.

The Messenger Matters: Experimental Evidence from a Sex Education Intervention

Noam Angrist*

Abstract

Governments all over the world have institutionalized sex education in schools. However, 1.8 million new HIV infections and 16 million teenage pregnancies still occur every year. One underexplored reason for the varied success of sex education is the *sender* of this information, which often defaults to teachers, in addition to the *type* of information. We conduct a large-scale randomized controlled trial in Botswana varying both the type of information and the sender: near-peers or teachers. We find that information can substantially reduce risky sexual behavior. These results are similar to a trial conducted in Kenya ten years prior, indicating the underlying mechanism is generalizable across highly heterogeneous contexts. However, we find that the sender of information makes or breaks whether information can change sexual behavior. Near-peers in Botswana reduce pregnancy by over 40 percent whereas government teachers have a null effect and potentially increase pregnancy. An analysis of mechanisms reveals that both messengers update student beliefs, but while students appear to be persuaded by near-peers they appear to ignore or rebel against teachers.

JEL codes: D8, D91, I12, I18, I28, J13

Keywords: Sex education, Information, Persuasion, External validity

*Angrist: University of Oxford; e-mail: noam.angrist@bsg.ox.ac.uk. I thank Abigail Adams, Felipe Barrera-Osorio, Paul Collier, Stefan Dercon, Esther Duflo, Rachel Glennerster, Penny Goldberg, Dean Karlan, Julien Labonne, Anandi Mani, Simon Quinn, and Climent Quintana-Domeque. A special thank you to Pascaline Dupas and Tavneet Suri for in-depth contributions and comments on the research study and paper. I thank the research staff at Young 1ove, including Sedilame Bagani, Bianca Bohmer, Danielle Boyda, Frederic Cochinard, Marcal Ribas, Kabo Kagiso and Omphile Maotwanyane as well as the implementation, training and leadership team: Baboloki Kwati, Thato Letsomo, and Moitshepi Matsheng. I thank the staff at J-PAL Africa: Emily Cupito, Ashleigh Morrell, Laura Poswell and Anna Yalouris; Karen Levy from Evidence Action; Gabriel Anabwani, Grace Karugaba, and Mogomotsi Matshaba from Botswana-Baylor Children's Clinical Center of Excellence and Lesego Gabaitiri, Nkobi Pansiri and Gabatshwane Tsayang from the University of Botswana. Finally, I thank government leaders from the Ministry of Education and the Ministry of Health: Grace Muzila, Shenaaz El-Halabi, Martin Keabona, Elizabeth Koko, Onkabetse Mmereki, and Gosata Rabantheng. I thank Claire Cullen and Carlos Garay for careful review of multiple drafts of the paper. The research and implementation was funded by the MAC AIDS Fund, J-PAL Africa and Evidence Action. The author of this paper, Noam Angrist, is also the co-founder and executive director of the NGO, Young 1ove, one of the main partners of this trial, which initially started to run this trial and has since run multiple randomized controlled trials in Botswana.

1 Introduction

Nearly everyone has sex.[2] There are benefits, such as pleasure and procreation. There can also be severe costs, such as contracting HIV and unwanted teen pregnancy. 1.8 million HIV infections and 16 million teen pregnancies occur every year (UNFPA 2013; UNAIDS 2018). To minimize these costs, societies have institutionalized sex education in schools worldwide.[3]

A growing body of evidence in economics and public health exists on the type of information that can reduce high-risk sexual behavior (Gallant and Maticka-Tyndale, 2004; Underhill et al., 2007; Kirby, 2008; Paul-Ebhohimhen, 2008; Dupas, 2011; De Walque, 2014; Duflo, Dupas and Kremer, 2015).[4] Over 80 percent of governments in a recent review have adopted information shown in this research to reduce risky sex in national curricula, such as safe sex strategies over abstinence-only messages (UNESCO 2015).[5] However, despite massive investment and policy adoption, the full potential of sex education remains unrealized, with comprehensive HIV knowledge among adolescents below 30 percent in a sample of 65 countries from the Demographic Health Surveys (DHS) and over one million pregnancies and HIV infections still occurring every year. One potential explanation for the unrealized potential of sex education is the sender of information.

A literature in economics models the conditions under which senders choose to provide information, are able to update recipient beliefs, and effectively persuade recipients to change their behavior (Grossman, 1981; Milgrom, 1981; Crawford and Sobel, 1982; Kamenica and Gentzkow, 2011; Gentzkow and Kamenica, 2016). All three margins appear important for sex education, where information is sensitive and taboo and the relationship between the typical sender (teachers) and receiver (students) is complex. On the one hand, teachers are often authoritative figures in the community and expert information providers. On the other hand, teachers might be uncomfortable discussing sex in public and perceived by students as socially distant and less credible. These factors, among others, generate ambiguous predictions on whether teachers are the ideal sender for sex education.

In this paper we provide evidence from a large-scale, government-backed program that the sender of information can be first-order for sex education to reduce risky sex. We conduct a randomized controlled trial with 42,000 students across a third of Botswana randomizing two senders, near-peers and teachers, to deliver sex education information. Near-peers are

[2] 97 percent of adults report having sex at least once in their lifetime. This statistic is based on data from the Demographic Health Survey (DHS) across 72 countries. The minimum is 80.4 percent and the maximum is 99.9 percent. Figure 1 shows sexual activity across two indicators by country.

[3] A review of 73 countries found nearly all had a law or policy on sex education (UNESCO, 2015; Ketting and Ivanova, 2018).

[4] Lo, Lowe and Bendavid (2016) examine the high-profile case of the President's Emergency Plan for AIDS Relief (PEPFAR) which has spent 80 billion (USD) on the HIV treatment and prevention response, nearly the same amount (PPP) as the Marshall Plan spent to rebuild Europe after World War II. They find that abstinence messaging had no association with a reduction in risky sex.

[5] Over 80 percent of countries reviewed had policy or curriculum mentioning comprehensive safe strategies. (UNESCO 2015). Large-scale consortia have formed to prioritize and advocate for these policies, such as the "Ministerial Commitment for Adolescent and Sexual Reproductive Health" among a group of 20 countries in East and Southern Africa to comprehensive sexuality education.

young and aspirational figures from the capital city. Teachers are guidance and counselling teachers charged with delivering sex education in the status-quo. Table 1 provides a comparison of sender profiles. The sex education message provides information that younger partners are a safer sex option than older partners since they are over five times less likely to have HIV. This message was adapted from a similar message shown to reduce pregnancy by 28 percent in a randomized trial in Kenya (Dupas, 2011). This enabled us to test the relative effectiveness of the sender, conditional on a likely first stage by building on information previously shown to reduce risky sex. We further test whether the information mechanism translates to Botswana by comparing results to a pure control group arm.

This study was conducted at-scale and in collaboration with government. Muralidharan and Niehaus (2017) document the median cluster size of a randomized trial in a subset of journals as 26 and propose running larger-scale trials. Our sample of 343 clusters represents triple the cluster size of the 75 percentile in their review. We further invest in comprehensive data collection with near-universal follow up a year later. We collect data on priors and beliefs as well as objective sexual behavior measures of teenage pregnancy with 99 percent follow-up 12-15 months later. This data enables reliable detection of risky sexual behavior as well as exploration of specific mechanisms of behavior change.

Our results show that providing information on low versus high-risk sexual partners can reduce pregnancy in Botswana by over 40 percent. The consistency in effects with a similar message tested in Kenya ten years prior suggests the information mechanism is generalizable across heterogeneous contexts. However, we find that the sender of information makes or breaks the intervention: while near-peers are effective at reducing risky sex, when teachers deliver the intervention, there is a null effect and a plausible increase in pregnancy. These results hold in rural areas as well as in primary and junior schools. Thus, we find that both the message and, crucially, the messenger, matter.

We analyze mechanisms to understand messenger effects. We explore three margins: compliance, beliefs and preferences. First, compliance. Certain senders may be more likely to deliver the sex education message and do so as intended. For example, teachers, the default messenger for sex education, might be absent, idle or uncomfortable discussing sensitive topics such as sex (Chaudhury et al., 2006; Ross, Dick and Ferguson, 2006; Duflo et al., 2012). Information will only change behavior conditional on being received and sent undistorted. Second, beliefs. The sender of information might be more credible and better update recipient beliefs if they are proximate on a bundle of characteristics such as age and gender. In a heterogeneous world, rational receivers will believe information shared by a similar sender since they will trust it applies to them. Third, preferences. Senders with

similar characteristics and less 'social distance' to receivers might be seen as role models and more likely to influence aspirations and tastes.[6]

We find it is unlikely diverging results by messenger are driven by the compliance or beliefs margins. Teachers largely deliver the information, stay on-message and successfully update student beliefs. This indicates teachers can be effective messengers for their core business: transferring information. Instead, diverging results seem to be driven by a preferences channel. Conditional on implementation, effects between near-peers and teachers diverge more. In other words, the same set of beliefs which reduce risky sex when delivered by near-peers appear to backfire when delivered by teachers.

We explore a simple conceptual framework to explain this behavior. Teenagers can choose to either listen to information, ignore it, or rebel against it. We introduce perceived agency in a teenager's utility function in a variation of the identity model (Akerlof and Kranton, 2000). That is, teenagers want to demonstrate they are in control of their own actions. The importance of the role of agency has been emphasized in the social cognitive theory literature (Bandura, 2001). Teachers, who are in a position of control over students, threaten teenagers' sense of agency when they prescribe a certain behavior. Near-peers, who are not in a position of control and are out-of-network, do not threaten agency. When teachers deliver the message students who take no action on their new beliefs demonstrate partial agency. If they behave in riskier fashion they demonstrate full agency. Thus, even when listening to the prescription has clear economic payoffs, if more utility is derived from expressing agency rebellion could occur. This simple framework illuminates the possibility of teenage rebellion – a phenomenon extensively documented in social psychology where people do not listen or do the opposite of a given prescription even when listening is in their best interest.[7] We provide an economic lens for this behavior and find results consistent with this mechanism.

We explore alternative mechanisms. These include the potential for the intervention to induce teachers to engage in risky behavior. However, we find that pregnancies occur only in schools with female teachers, who cannot impregnate young girls. We also explore descriptive trends in the role of proximity between sender and recipient characteristics. We find that as teachers and near-peers converge on characteristics such as age they are no more effective at reducing risky sex. Rather we find that when the sex education message is delivered by more experienced teachers this leads to higher-risk behaviors among students, supporting the notion that a gradient of authority and control of teachers over students might trigger teenage rebellion.

[6] Female teachers have been shown to change aspirations for low-ability female students (Eble and Hu, 2019) and influence major choice in university for female students (Bettinger and Long, 2005). Kofoed and McGovney (2019) find that both same gender and same race role models influence occupation choice. Alsan, Garrick and Graziani (2019) find that patients are more likely to demand preventative services if randomly assigned same-race doctors.

[7] Milgram (1963); Brehm (1966); Pennebaker and Sanders (1976); Bushman and Stack (1996); Bandura (2001).

We further explore the role of proximity between sender-recipient characteristics with a second randomized trial. We randomly assigned surveyors to different schools in the endline survey, providing an exogenous shock on a bundle of surveyor characteristics, such as gender and age, on student survey reporting behavior. While this embedded trial does not strictly translate to messenger effects in the main trial, it is suggestive and can illustrate possible mechanisms. We examine reporting behavior on sensitive questions (such as 'have you ever had sex') versus non-sensitive questions (such as 'HIV knowledge'), which we interpret as a measure of 'freedom of expression.' We find that surveyor characteristics, such as the age of the surveyor, affect a student's freedom of expression: students are more likely to admit to having sex when asked by younger surveyors. However, while age in absolute terms matters, we do not find that relative proximity matters: older students who are closer in age to surveyors are no more likely to reveal sensitive answers. Thus, messenger characteristics appear to matter on margins beyond beliefs, but for reasons other than relative proximity. This reveals that certain messengers might stifle freedom of expression and potentially threaten agency while others might not.

Our findings relate most closely to five literatures in economics. First, we contribute to the literature on sex education. The importance of the sender of information might help reconcile why despite field experiments in Africa which have found large responses to information (Dupas, 2011; Chinkhumba, Godlonton and Thornton, 2014; Godlonton, Munthali and Thornton, 2016; Kerwin, 2019), risk responses are much smaller at the population-level (Oster, 2012). One reason for this discrepancy might be that at the population-level teachers are almost always the default messenger.

Second, we contribute to a literature on the role of information in changing behavior. Information economics has burgeoned since the seminal work of Stigler (1961). Information has been shown to shape behavior in multiple contexts including education (Jensen, 2010), HIV risk perceptions (Dupas, 2011), HIV status and testing (Thornton, 2008; DePaula et al., 2014), water and sanitation risk (Madajewicz, 2007) and the environment (Jalan and Somanathan, 2008). We contribute to this literature revealing that information can change behavior in multiple directions, with the sender influencing the direction.

Third, we relate to a literature on social learning which finds that the source of information is important for its diffusion. Munshi (2004) and BenYishay and Mobarak (2018) find proximity of in-network peers matters for beliefs about agricultural technology and subsequent adoption. Banerjee et al. (2019) find that 'gossips' are most effective for in-network diffusion of information and in turn increasing rates of vaccination. Our results suggest that the source of information might not only affect the diffusion of information but also influence preferences. This builds on insights from studies finding different senders of information can affect demand for preventative health care services (Alsan, Garrick and

Graziani 2019), major choice in university (Bettinger and Long 2005) and occupational choice (Kofoed and McGovney 2019).

Fourth, we relate to a literature on persuasion. Economists typically model persuasion as the provision of information (Nelson, 1974; Crawford and Sobel, 1982). Kahneman and Tversky (1982) and Mullainathan, Schwartzstein, and Shleifer (2008) extend this to emphasize the role of non-informative associations, such as advertisements of 'silky shampoo' where 'silk' evokes a positive association. While much attention has been given to the role and type of information in the persuasion process, more limited empirical attention has been given to the sender. Our results suggest a persuasion rate of 98.2 percent when information is delivered by near-peers, higher than nearly all persuasion rates in a review by DellaVigna and Gentzkow (2010).[8] This illustrates the first-order role of the sender of information in persuasion. In addition, the persuasion literature largely focuses on advertising (Stigler, 1961), politics (Downs, 1957), and law (Dewatripont and Tirole, 1999). Yet, schools play a central role in persuasion by shaping beliefs and preferences among adolescents. Consider campaigns by Coca-Cola to install vending machines in schools to build their future customer base. We contribute to the persuasion literature in the context of beliefs and preferences formed in school.

Fifth, we contribute to the literature on external validity. Studies have found significant treatment effect heterogeneity across contexts including microcredit (Vivalt, 2015; Meager, 2019), fertility and labor supply (Dehejia, Pop-Eleches and Samii, 2019), and the returns to agriculture (Rosenzweig and Udry, 2019). A growing literature explores the degree to which estimates can be extrapolated across contexts.[9] In our case, the effectiveness of a similar sex education message across both Botswana and Kenya, with up to 40 percent and 28 percent reductions in pregnancy, respectively, is striking since the trials took place across highly diverse contexts. The Kenya trial took place from 2003-2005, in a low-income context and when HIV medication availability was limited. The Botswana trial took place from 2014-2015 in an upper-middle income setting with widely available HIV medication. It is not clear ex-ante if the effects of the program will replicate. For example, the cost of HIV risks posed by older partners might be higher in Botswana since average HIV rates are five times higher than Kenya. On the other hand, the cost of HIV might be lower since today in

[8] We calculate the persuasion rate in line with DellaVigna and Gentzkow (2010): $f = 100 * \frac{y_T - y_C}{e_T - e_C} * \frac{1}{1-y_0}$ for a binary outcome, treatment group T, and control group C. e_i is the share of group i receiving the message, y_i is the share of group i adopting the behavior of interest, and y_0 is the share that would adopt the behavior if there were no message (approximated by $y_c$where y_0 is not observed). The most conservative estimate of $y_T - y_C$ for near-pears is estimated by a reduction in pregnancy incidence of .214 percent (Table 7). $e_T - e_C$ can be estimated by the percentage of students who learn older partners have the highest rate of HIV of 44.2 percent (Table 4), and 1-y_0 can be estimated as the share of the population that falls pregnant in the control group in a given year of roughly half a percent (control mean in Table 7). This produces a persuasion rate of 98.2.

[9] This literature includes definitions and formal approaches (Rothwell, 2005; Holtz et al, 2005; Duflo, Glennerster and Kremer, 2007), the importance of understanding mechanisms (Bates and Glennerster, 2017; Deaton and Cartwright, 2018), extrapolation across contexts (Pritchett and Sandefur, 2013; Allcott, 2015; Dehejia, Pop-Eleches and Samii, 2019), and empirical practice (Peters, Langbein and Roberts, 2016).

Botswana there is free and widely available anti-retroviral medication and HIV is no longer lethal. These differences, among others, produce an ambiguous prediction on whether results will replicate. Table 2 compares Kenya and Botswana, as well as a related study in Cameroon, along multiple indicators at the national level and for each study sample.

Our results suggest that despite significant scope for heterogeneity across these very different contexts, the information mechanism generalizes across countries and time from Kenya to Botswana fifteen years later. These results reinforce that information does matter and can reduce risky sex, despite numerous other factors driving risky sex, such as credit constraints. However, our finding that the sender determines the direction and magnitude of effects within the exact same geography, Botswana, suggests that external validity might turn less on geography and associated covariates, which often significant attention in the literature, and more on the implementor or sender. Of note, while our study finds large effects of the messenger, one of the first studies to experimentally isolate the relative role of the sender in the context of sex education found the sender was of limited importance in Cameroon (Dupas, Huillery, and Seban, 2018). In the discussion section we attempt to reconcile these findings. One possibility is variation in messenger profiles. Teachers in Cameroon are substantially more likely to be non-permanent contractors than teachers in Botswana (90 percent relative to 10 percent) and correspondingly younger and less formally trained. Thus, in practice teachers in Cameroon might be more similar to external consultants or near-peers, which might lead to less difference among messengers in Cameroon than we observe in the Botswana trial. Another possibility is varying levels of trust. In Botswana trust in science is lower than Cameroon while trust in neighbors and NGOs is higher. This suggests that factual information alone might not be enough to persuade in Botswana and that particular sources of information are more trusted and persuasive.

Our results have significant implications for policy. Teachers are often the default messenger for sex education. Teachers are respected figures in the community and a fixture of the education system. However, complementary government schemes exist which could be leveraged to deliver sex education. For example, many countries have national service programs which deploy young people in schools.[10] AmeriCorps in the United States places young volunteers in over 12,000 schools throughout the country. In sub-Saharan Africa, over fifteen countries have national service programs with cumulative government investments of over $240 million per year (Bodley-Bond and Cronin, 2013). This presents a scalable model for near-peers to deliver sex education in schools. Our results suggest that while the current default messenger for sex education, teachers, can be effective for transferring information and might be best placed to provide basic sex education facts, such as what HIV is, they

[10] In addition to national youth service programs, plausible near-peer messengers could include students in similar or higher grades in the same school.

might not persuade students to reduce risky sex. In contrast, near-peers can reduce pregnancy by over 40 percent. Leveraging near-peers as complements in delivering sex education could yield large welfare gains.

The rest of this paper is structured as follows. Section 2 describes the context and experimental design. Section 3 describes the data. Section 4 outlines the empirical strategy and results followed by a discussion of mechanisms in Section 5. Section 6 concludes.

2 Context and Experimental Design

The study took place from 2014 to 2015 in Botswana in a third of the country. The study is a three-armed randomized trial conducted in government schools. 347 schools were included in the initial sampling frame across three grade levels: grade 6 in primary schools and grade 8 and grade 9 in junior secondary schools. This encompasses the universe of all schools in each region according to administrative records from the Ministry of Education. Schools not listed on official government lists were not included in the randomization.

Of the 347 schools in the original sample, 115 schools were selected for sex education messaging by near-peers, 116 schools by teachers, and 116 schools selected for the control group. Of the 347 schools in the initial sample, 343 schools were included in the final sample. Four schools were excluded: two were excluded as they did not have grade 6 students, another was excluded as it was a private school, and a final school was not able to be found. The final sample includes 273 primary schools and 70 junior schools, with roughly 42,000 youth enrolled. The randomization was conducted based on the 343 schools in this final sample, stratified by region and school level (primary and junior). Figure 24 in Appendix A shows a map of all schools in the study.

Table 3 presents summary statistics and balance tests across survey data as well as administrative data on test scores, dropout, and pregnancy. We find overall balance, with no significant baseline difference between treatment groups in any of the administrative data variables. We note two survey indicators with a statistically significant difference. It is possible these are driven by surveyor fixed effects, since for near-peers the baseline was conducted prior to the intervention by the implementor, selected to be charismatic and trustworthy. To account for differences and enhance statistical power we control for a vector of baseline covariates.[11]

Figure 2 provides a study timeline. The baseline survey commenced in August 2014, followed by the implementation of extending through November 2014. The endline survey

[11] These variables were determined in a pre-analysis plan to be of interest. They include: male and female enrollment, gender composition, average age, baseline dropouts, self-reported rates of pregnancy, the percent of girls in the class reported to be pregnant, and the percent of girls in the class reported to have older boyfriends.

commenced 12 months later in August 2015 and was complete by the end of December 2015. Follow-up occurred within schools at the next grade level assuming standard progression. All 343 schools and grade levels in the final sample at baseline were surveyed 12 months later except students in grade 9 in one school in the teacher arm.

The study occurred in four regions, representing over a third of the country, and included both rural and urban settings. One region includes the capital and is highly urbanized. The other three regions are largely rural. The intervention was conducted in late primary school and lower secondary schools with students ranging in age from 12-16. This age range covers adolescents around the age of sexual debut. While exact data on age of sexual debut is inconsistent, the sum of the evidence suggests the average first sexual encounter occurs between ages 15-19. While only 4 percent of youth in Botswana younger than 15 have had sex, over 60 percent of youth 19 or older are sexually active (Statistics Botswana, 2013; Ministry of Basic Education, 2016). During this phase of adolescence, sex education can influence a large margin of students since sexual relationships are just starting to be formed.

Moreover, most students in this age range are easy to reach since they are in school. The primary net enrollment rate in Botswana is 90 percent (UIS, 2014). The transition rate from primary to junior school is 97 percent (Statistics Botswana, 2012). However, there is a sharp drop-off between lower and upper secondary school of over 30 percent of students (Statistics Botswana, 2012). Thus, school-based interventions that intervene before this large drop-off can reach most students before negative outcomes occur.

The status-quo sex education curriculum in schools covers comprehensive sex education messaging, including basic information on HIV, contraception, and related risks. Each school has a dedicated Guidance and Counselling teacher who is expected to deliver this curriculum over the course of the year. In practice, while every school has dedicated teachers and time to deliver sex education, over 80 percent of students in Botswana have never heard of the curriculum (Ministry of Basic Education, 2016). Instead teachers often don't deliver the curriculum or if they do deliver sex education the typical message teachers promote is abstinence. Figure 25 in the Appendix A shows an example of a typical sign at a school. This presents an opportunity to activate delivery of proven sex education messaging in Botswana. To this end, we examine the effect of a sex education message shown to work in a prior context and explore differential effects when delivered by near-peers and teachers.

2.1 Sex Education Message

A study in Kenya found that revealing that older partners are significantly more likely to have HIV than younger partners led to a large reduction in teen pregnancy of 28 percent (Dupas, 2011). We deliver a similar sex education message in Botswana. Similar to Kenya, HIV rates increase dramatically with age in Botswana. 3.6 percent of 15 to 19 year-old boys

are infected with HIV relative to 43.8 percent in 40 to 44 year old men (Statistics Botswana, 2013). Figure 3 shows the HIV rate by age and gender in Botswana. Between ages 10 to 24, girls have a 3x increase in HIV prevalence. This illustrates two important facts. First, HIV rates increase substantially with age. Second, young girls are likely getting infected by older partners, since young girls' infection rate increases exponentially while the infection rate of their male near-peers remains flat.

Despite the dire risks older partners pose, intergenerational relationships are common. According to a study by PSI, 12 to 25 percent of young women in sub-Saharan Africa had partners ten or more years their senior (Luke and Kurz, 2002). The pervasiveness of relationships with older partners is driven by a few factors. In some instances, coercion or economic circumstances are the driving cause. In other instances girls opt for older men, expressing agency over entering into the relationship (Olson, 2010). Older partners present tangible, immediate benefits in the form of gifts, while the costs – such as the risk of unprotected sex, teen pregnancy or contracting HIV - are unknown, underestimated, long-term and not visible. Our baseline in this study revealed that over 90 percent of youth could not identify that forty-year-old men, 45 percent of whom have HIV, were the highest HIV risk group. Girls perceive these older men as low risk describing them as "mature and financially secure" and often enter these relationships unaware of the severe risks.

In this context, revealing the relative HIV risk of older versus younger male sexual partners may make the previously underestimated cost of older partners known and salient. This knowledge of relative risks by age further makes clear that dating age-mates is a relatively safer sex option. With the risk of older partner now acute and an attractive safer sex option presented, girls might shift from older partners to young partners. Among similar-age partners, the risk of engaging in unprotected sex, falling pregnant or contracting HIV is significantly lower.[12]

This intervention is in sharp contrast to the dominant alternative message: abstain. The typical abstinence message emphasizes extreme risk avoidance ("don't have sex until marriage"). This message fails to provide a realistic safe sex option to youth who will not abstain. If sexual behavior is more likely to change on the intensive margin (type of sex) rather than the extensive margin (sex or no sex) providing safe sex alternatives, such as revealing younger partners are lower risk than older partners, will be more effective. The results of the Kenya intervention provide support for this mechanism. These results relate to findings from high income settings which show that messages about safe sex alternatives are more effective at reducing risky sex than abstinence messages.

[12] In addition to the much lower rates of HIV among younger relative to older partners, a series of factors culminate in lower likelihood of unprotected sex and pregnancy among similar-aged partners. This includes increased negotiation power to use condoms or delay sexual debut. In addition, there is often mutual interest among male and female partners in having protected sex or delaying sex since young boys do not have the resources to relocate village or take care of a child. To this end, pregnancy likely reflects a lower bound of delayed sexual debut, since protected or delayed sex decreases both pregnancy and HIV, while HIV might decrease even if unprotected sex occurs among similar-aged partners since HIV rates are an order of magnitude lower.

A curriculum was adapted from Dupas (2011) with a near-identical sex education message adapted to the Botswana context. The message was delivered through a one-hour class in government schools and comprised five segments: ice-breakers to generate comfort talking about sex, a puzzle activity where youth guess in groups which male age group is most likely to have HIV, information on the true HIV graph disaggregated by age with statistics from the Botswana AIDS Impact Survey (a color printout of Figure 3 for men only), a 14-minute video and corresponding comic book produced by UNICEF showcasing a young girl overcoming an older partner in a rural setting, and a group discussion on the risks of intergenerational sex along with actionable lessons learned to avoid riskier older partners. The adaptations made in the Botswana curriculum include: country-specific HIV rates and the use of 10-year age bins rather than 5-year bins, a puzzle guessing the HIV rates prior to the reveal of the true rates, and ice-breakers. These activities facilitated understanding and comfort in the context of sex education delivery in Botswana. This near-identical curriculum and message from Kenya to Botswana enables us to uniquely assess external validity of the intervention across contexts.

2.2 Messengers: Near-peers and Teachers

We randomize near-peers and teachers to deliver a sex education message. The control group consists of teachers delivering a status-quo curriculum. This design enables us to capture the relative effect of the messenger when delivering the sex education message.

In the near-peer arm, young and aspirational figures were selected from a competitive application pool to deliver the sex education message. Near-peers were trained for five days on the intervention by a local NGO, Young 1ove, and two professors from the Department of Primary Education at the University of Botswana. Once at the school, one near-pear was allocated to each classroom ranging from 31-40 students on average. All classrooms in a school were reached simultaneously during study hour. All schools randomized to receive treatment were reached within 40 working days. Random monitoring visits were conducted by Young 1ove, the Ministry of Education, and the Botswana Baylor Children's Clinical Center of Excellence, the largest HIV clinic in the country.

In the teacher arm, one teacher was nominated per school to receive training. The criteria for selection was to ensure the teacher was a guidance and counselling teacher. This ensured delivery of sex education was in their job description to increase the likelihood of delivery. It further enables comparison with the control group where guidance and counseling teachers are expected to deliver the status-quo curriculum. Teachers received trainings hosted by the Ministry of Education in each of the four regions in the sample. An official of the Ministry of Education opened each training. The training was delivered by the same head facilitator from Young 1ove and professors from the University of Botswana who trained the near-peer

arm on the same five segments to provide a near-identical intervention. While it was not possible to use video with all teachers, since working projectors were not available across all schools, teachers were given an animated comic with the same story line as the video to deliver the same message. Teachers were compensated for attending the training. All teachers were requested to sign commitment cards indicating when they planned to implement the intervention within the term. This served as a commitment device. All teachers were sent reminder text messages to implement the intervention and were instructed to give students written reflections to be turned into the regional Ministry of Education authorities for verification of implementation. Random monitoring visits were conducted by the same monitoring team as the near-peer arm.

The teacher intervention was built with a scaffolding of support to ensure teachers implemented the program. This was designed to enable a powered and credible analysis of effects of the messenger conditional on successful implementation. Implementation fidelity was high. Attendance at training can be considered a proxy for the upper bound of implementation, commitment card scheduling a middle bound, and written student responses submitted back to the regional authority a lower bound. At the school-level, 90.53 percent of schools randomly selected to receive the sex education message by teachers attended training, 73.59 percent scheduled implementation, and 30.24 percent of schools sent written student responses to the ministry regional offices. All are imperfect measures but provide some confidence that we can detect effects of teacher implementation.

Each messenger represents a bundle of characteristics. Figures 4 and 5 provide visual comparisons including key characteristics, such as age and gender. Table 1 includes additional covariates such as teaching experience. Near-peer educators are substantially younger than teachers. Gender composition is similar: 88 percent of near-peers are female relative to 80 percent female for teachers. In terms of teaching experience, near-peers have none relative to teachers who have an average of 20 years of teaching experience. Along this bundle of characteristics for which we have data, near-peers messengers appear significantly more proximate to students who receive the sex education message than teachers.

It is not clear ex-ante if teachers or near-peers will be more effective. Near-peers might be less effective at updating beliefs due to lack of experience and authority. Therefore, recipient behavior might not change. On the other hand, near-peers might be more reliable implementors as well as proximate and credible messengers. This might result in greater belief updating and corresponding sexual behavior change. Alternatively, teachers and near-peers might be similarly effective at changing beliefs but differentially effective at changing behavior. In Section 5 we explore mechanisms that might explain messenger effects consistent with our results.

3 Data

3.1 Baseline

We collected baseline data using surveys as well as administrative records obtained from the Ministry of Education and the Botswana Education Council. Baseline surveying in schools took place from August to November 2014 in all 343 schools with 42,195 youth in the final sample. A short, anonymous one-page survey was conducted where students were asked to answer 20 questions covering demographic information, knowledge, attitudes and self-reported sexual behavior. All surveys were completed prior to the intervention.

In addition to the baseline survey, surveys were conducted immediately after the intervention to capture immediate shifts in beliefs in both treatment arms. All near-peer arm schools participated in this survey. A third of schools in the teacher arm submitted student responses.

Additional data was collected on teacher demographics and implementation. This includes data on teacher attendance at training (106 teachers attended[13]) as well as a record of signing of commitment cards to implement. The commitment cards also included scheduling specific implementation times. In addition, data was collected on teacher demographics, education levels and years teaching for 88 of the 114 guidance teachers.[14]

Administrative data were obtained from the Ministry of Education on enrollment, dropouts, and dropout rationale (including pregnancy, truancy, performances, illness and others) in 2012 for all 343 schools in our final sample.[15] Administrative data were obtained from the Botswana Education Council on test score performance from the 2014 Primary School Leaving Examination (PSLE) and the Junior Certificate Examination (JCE), both standardized national examinations conducted annually. This data contains absolute test score performance, percent of students achieving grade A, B, or C and demographic information on age and gender composition at the school level.

The baseline data serves three purposes: checks for balance to verify the integrity of the randomization; descriptive information on the profiles of messengers and schools where the intervention was delivered; and immediate shocks to beliefs.

3.2 Endline

[13] All 114 schools were formally invited by the Ministry of Education to send a guidance teacher for training. NGO staff followed up with all schools via letters and individual phone calls to increase likelihood of attendance, in addition to providing compensation for training. 106 teachers showed up for training across all 4 districts. This likely marks a high bound for attendance at any typical teacher training. Absence is common due to sick leave, long travel distances (schools are hundreds of kilometers from regional training headquarters), exams and other priorities, and lost letters or inability to contact the school.

[14] All 106 teachers who attended the training filled out a contact form with their mobile number and were thereafter contacted for these details via a mobile survey. Some numbers were broken and some teachers did not respond. Follows ups were made with the school to retrieve individual teacher contact information. There was an imperfect response rate as some mobiles did not work and some teachers were unresponsive. Some teachers had also been transferred to another school.

[15] This data was collected using written forms sent to all school heads which are in turn completed and returned via post to the Ministry of Education. This data is compiled at the school level.

An in-school survey similar to the baseline was conducted in all classes and schools. The endline school data collection commenced in mid-August 2015 and was completed by the end of December 2015.[16] Surveying took place within the same schools surveyed at baseline at the next grade level assuming standard progression. Grade 6 students progressed to grade 7 in primary schools, and grade 8 and 9 students in junior school progressed to grade 9 and 10, respectively. In total, 41,465 students were surveyed at endline.

The same set of surveyors conducted the endline survey across all three arms and the average difference between the survey date and the start of endline ranged from 13 to 13.5 days across all arms and was statistically equivalent. Adherence to instructions and the percent of responses missing were also statistically equivalent across all three arms. The survey was a short, anonymous, two-page survey where students were asked to answer 34 questions covering student demographics, knowledge, attitudes and self-reported sexual behavior. Surveys were split into "female" surveys and "male" surveys to tailor gender-specific questions.[17]

In addition, a 'roll call' method was used to collect data on girls missing from school using baseline registers of students present at baseline.[18] If a girl was missing, classmates were asked to identify the reason, including whether she had fallen pregnant. In rural communities it is common knowledge if a girl has dropped out due to pregnancy, a visible and biological outcome. Legally, all girls who fall pregnant are required to withdraw from school for a full year suggesting that this measure captures nearly all pregnancies.[19]

We confirmed the roll call data by also asking school administrators to identify any girls missing from school due to pregnancy.[20] School administrator surveys were triangulated real-time with the roll call data. The lead enumerator at each school was instructed to check if both surveys aligned. A maximum of the roll call and administrator data was taken to identify pregnancies. The survey also asked demographic information on the school administrator's age, credentials, and years teaching, as well as overall enrollment numbers.

[16] 338 of 343 schools were surveyed by September 17th, with the remaining 5 schools delayed by examinations and scheduling challenges. All schools were reached by December besides one grade level in one school in the teacher treatment arm.

[17] For example, question 27 in the male survey reads "Have you ever made anyone pregnant?" whereas in the female version it reads "Have you ever been pregnant?"

[18] Students not on the baseline roster but present in the classroom were noted as "new comers." In our sample, there were roughly 3,772 students who were on the endline but not baseline roster, 1,918 students because they had transferred in from a different school, and 1,854 students because they were either absent or not on the roster at baseline. A copy of baseline rosters from every class in the school were made available to a lead enumerator at the school to cross-check whether new-comers were transfers within classrooms in the school or from another school. Students transferred from a classroom within a school were noted as present in their classroom and removed from their former classroom so as not to falsely be captured as a dropout or absentee.

[19] This provision is detailed in Section 34 of the Botswana Education Act.

[20] At primary schools, this was conducted with the guidance and counseling teacher. At junior schools, this was conducted with one of the Head of Departments (HOD). Guidance teachers and HODs were chosen as the administrators with the most accurate knowledge of missing girls. School heads are removed from day to day activity, and individual teachers are only knowledgeable about their individual classrooms. To this end, guidance teachers and HODs are arguably most likely to have a complete picture of which girls are missing and why. HODs were standardized as the administrator at junior schools so as to avoid potential reporting bias in treatment arms where guidance and counselling teachers were trained or responsible for supporting scheduling of implementation. Primary schools, however, did not have HODs, so guidance teachers were selected as a second-best standardized school administrator.

We verified pregnancies identified at the school by conducting at-home visits. This occurred from September through December 2015.[21] Out of school surveys were conducted with both girls and their caregivers.[22] Girls received a written survey and an oral interview. This survey consisted of 35 questions nearly identical to the in-school endline survey. The written and oral survey were identical and the written survey was always taken first. Caregivers were interviewed orally. The caregiver survey contained 27 questions and contained similar questions on knowledge, attitudes and reported sexual behaviors about the girl. The surveys were conducted simultaneously with a physical structure or significant distance separating girls and their caregivers for privacy. If girls and caregivers were not able to be interviewed at the same time the girl was interviewed first.[23] At the household, enumerators also noted visible signs of pregnancies, such as a noticeably large stomachs or a baby. In total, 588 girls were identified for at-home visits with a follow-up success rate of 99 percent of girls identified at the school as most likely to be pregnant were found. These follow-up visits required enumerators to survey the entire country, including meetings with local chiefs as well as to conduct visits across the border in Zimbabwe and South Africa. Appendix B describes verification of our main measure of pregnancy using this data. Follow-up data from at-home visits further enabled us to pool written survey data with survey data collected in schools. This ensures that analysis using survey data captures the full sample rather than only capturing the sample of girls in school which could be affected by treatment.

In summary, our measure of pregnancy is computed using a maximum of roll-call and school administrator data at the school. Over 90 percent of pregnancies identified at the school were confirmed during at-home visits. In addition to providing a measure of objective sexual behavior this measure has been used in the literature to identify pregnancy (Dupas, 2011; Duflo, Dupas and Kremer, 2015; Dupas, Huillery, and Seban, 2018) and provides a direct analogy.

4 Results

4.1 Beliefs

[21] A random sub-sample of girls who had transferred to schools outside of the study sample were followed up to in case 'transfer' was a euphemism used by school classmates and administrators to cover up dropouts due to pregnancies.

[22] Of the 539 at-home visits, 504 included both a girl and caregiver out-of-school survey or the girl survey only when found in school, 22 included a caregiver survey only, 6 only include a girl survey only, and 7 girls were confirmed to be in school.

[23] A few exceptions where the caregiver was interviewed first occurred only if the caregiver had no contact with the girl (no phone number or no knowledge of her whereabouts) to ensure the caregiver interview did not bias the subsequent girl interview.

We first examine effects on student beliefs. We exploit random assignment to identify causal effects and recover the effect of the sex education message when delivered by near-peers and teachers. We run an ordinary least squares regression using the following specification:

$$beliefs_{ij} = \alpha + \beta_1 peer_j + \beta_2 teacher_j + \delta_s + \varepsilon_{ij} \quad (1)$$

where $beliefs_{ij}$ refers to a belief outcome, such as knowledge that older partners are more likely to have HIV or be riskier in general, for individual i in school j. The terms *near-peer* and *teacher* denote a dummy variable which takes on a value of 1 when a school is assigned to each arm and 0 otherwise. δ_s is a dummy for strata fixed effects. Standard errors are clustered at the school level. Our primary analysis focuses on effects for girls to relate beliefs to pregnancy and sexual behavior outcomes later on. A pre-analysis plan was constructed prior to analysis and completed on the American Economic Association (AEA) registry.[24]

We split the analysis into two stages: beliefs immediately after the intervention and beliefs 12 months later. Table 4 summarizes the results. We see a large and immediate shift in beliefs right after the intervention in the near-peer arm in columns 1 and 2 with a 44.2 and 33.0 percentage point increase in HIV and general risk knowledge. We see a large although smaller shock to beliefs in the teacher arm of 17.4 and 12.0 percentage points. We observe that 12 months later about a quarter of all belief outcomes are retained among both near-peers and teachers in columns 3 and 4. Figures 6 through 9 depict results graphically.

The beliefs most likely to affect sexual behavior by endline are closest to the period shortly after the intervention since it takes six to nine months for observable pregnancy to occur. We supplement our analysis of immediate belief shocks with analysis of beliefs 12 months later for two reasons. First, if trends a year later exhibit similar trends to those immediately after the intervention, this increases our confidence that these trends are monotonically decreasing and therefore hold in the following months after the intervention. Second, persistence of beliefs increases the likelihood of long-run impact.

Table 4 also indicates the degree of belief updating for teachers most likely to implement the program. This is captured using data from student written responses sent by teachers to the regional ministry headquarters upon completion of the intervention. Over a third of all randomly selected teachers handed in this evidence of implementation. This is likely a lower bound of teacher implementation, since teachers might implement but forget to send evidence of implementation or mail might get lost in transit.[25]

[24] The trial registry was completed in December 2014.

[25] We further conduct quality checks to ensure student reflections capture implementation and are not filled with fake data. We code open response answers from students reflections along four dimensions: wording, context, handwriting, and ink color. We treat uniqueness among these categories as a proxy for the likelihood that these student reflections capture true implementation rather than fake data being filled out by the teacher. On average, we find that just .1 percent appear to be fake responses. The open response format was designed to incentivize implementation since filling out hundreds of open responses by teachers is time-consuming and high-effort,

For the sub-set of teachers that hand in evidence of implementation, beliefs update similarly to the near-peer arm. In Table 4, columns 1 and 2 indicate that teachers increase HIV and risk knowledge right after the intervention by 53.0 and 44.0 percentage points. Similar trends hold a year later, with 7.7 and 6.2 percentage point increases in the teacher arm relative to 8.7 and 8.4 in the near-peer arm. This suggests that conditional on implementation teachers can update beliefs as effectively as near-peers. Figures 10 through 13 depict these results graphically showing convergence between near-peer and teachers for the subset that implement. These shifts in beliefs are substantial. Immediately after the intervention, beliefs update over four-fold relative to the control mean average; a year after the intervention beliefs update by almost 90 percent.

The subsample of teachers who implement is potentially selected. Table 5 explores possible selection along baseline school characteristics, including test scores, dropouts and pregnancy, as well as teacher characteristics such as age, gender, experience, qualification and attitude towards the intervention. We find no selection for the sub-sample of schools where teachers implement along any measure. While this might appear surprising at first glance, it is consistent with the context and design of the intervention. The intervention is only one hour. Thus, the marginal cost of delivering it is very low. Moreover, schools already have time in the school scheduled to deliver sex education, again reducing barriers to entry and cost to delivery. Finally, all teachers randomly selected for training were guidance and counselling teachers and are thus already selected to engage with sex education at the outset. Most selection is likely between guidance and counselling teachers and other types of teachers, rather than within guidance and counselling teachers. With this context in mind, it is not entirely surprising that implementation was simply sporadic rather than selected. The lack of selection suggests that our teacher results are plausibly reflective of guidance and counselling teacher effectiveness in general beyond the sub-sample who implement.

We next explore the degree to which beliefs update as intended. It is possible that messengers might update knowledge of older partner risks but not make clear that similar-aged partners are relatively safer, thus partially deviating from the sex education message. It is also possible teachers revert back to delivering an abstinence message akin to the status quo. For example, sending the message "older partners are riskier, so abstain." We capture whether the message is implemented as intended through a linear index based on the degree to which beliefs update on the intensive margin (revealing *relative* rather than *absolute* risks) for both HIV knowledge and risk perceptions. Table 6 summarizes effects on beliefs after 12 months. These estimates provide a lower bound for effects immediately after the intervention. We also examine effects in junior school and rural areas. In the near-peer arm, there is an increase in beliefs of relative risks of about 10 percentage points a year later. For the

potentially even more so than implementing the one-hour intervention. Our data indicates this was true in practice and that this measure reliably captures implementation by teachers.

teacher arm, the effect is smaller but significant with up to 5 percentage point increases in knowledge of relative risks. Figure 14 summarizes results graphically for HIV knowledge. We see beliefs update in a similar direction across arms, with a partial transfer in the teacher arm.

We next explore belief updating for the teachers that implement. Figure 15 shows convergence of belief updating for this sub-sample similar to earlier results. We observe that both near-peers and teachers who implement are able to update beliefs to a similar degree and achieve belief updating as intended. Both near-peers and teachers produce a relative risk belief shock of roughly 8 to 12 percentage points a year later. This holds across all samples including in rural areas and in junior school.

In summary, beliefs update in a similar direction for both teachers and near-peers. Both near-peers and teachers induce a large shock to risk perceptions of older versus younger partners in line with the design of the sex education message. This shock to beliefs persists up to a year later. Moreover, while teachers are on average less likely to implement than near-peers, conditional on implementation, beliefs update in the same direction and near-identical magnitude. This indicates near-peers and teachers can be similarly effective at updating beliefs.

4.2 Risky Sexual Behavior

Intent-to-treat estimates. — We next examine effects on risky sexual behavior. We capture the effect of the sex education message when delivered by near-peers and the relative effect when delivered by teachers. We estimate intent-to-treat effects as follows:

$$Y_{ij} = \alpha + \beta_1 SexEd_j + \beta_2 SexEd_j \times teacher_j + \gamma \boldsymbol{X}_j + \delta_s + \varepsilon_{ij} \tag{2}$$

where Y_{ij} is a sexual behavior outcome for individual i in school j. $SexEd_j$ refers to the sex education message and takes on the value 1 in both treatment arms and 0 otherwise. X_s denotes a vector of school-level baseline control variables to enhance statistical power and precision. These variables were determined in a pre-analysis plan to be of interest. They include: male and female enrollment, gender composition, average age, baseline dropouts, self-reported rates of pregnancy, the percent of girls in the class reported to be pregnant, and the percent of girls in the class reported to have older boyfriends. We also examine effects in junior school and rural areas. Since we stratified within these dimensions we can recover causal effects for each sample. We estimate the regression using both ordinary least squares (OLS) regression.

Table 7 shows results for our main pregnancy outcome. Pregnancy incidence goes down by .214 percentage points for all girls in the near-peer arm. This represents a 43 percent

reduction in pregnancy incidence relative to the control. Effect sizes are larger for junior schools and in rural areas, with reductions in pregnancy of up to 64 percent. These results indicate that a safer sex education message can be significantly more effective than status-quo abstinence messaging. This supports the theory that sexual behavior is more likely to change along the intensive margin. Our results in Botswana are consistent in direction and size with results from Kenya ten years prior (Dupas, 2011), indicating that this mechanism translates across highly diverse contexts.

We find, however, that the messenger is first-order in determining effect size and direction within Botswana. Teachers increase pregnancy, more than offsetting the near-peer reduction. This result is surprising given that beliefs update in a similar direction across both near-peer and teacher arms. Instead of producing effects that go in the same direction, belief updating seems to drive results in *opposite* directions across messengers. While students reduce risky sex when they learn from near-peers that same-age partners are safer, it seems there is a backlash to these same beliefs when the message is delivered by teachers, with students behaving in a slightly riskier fashion.

Figures 16 through 19 depict the distribution of school-level pregnancy incidence by treatment arm. We see that the near-peer arm generates a reduction in pregnancy with a spike in the zero pregnancies tail. For junior schools, we see that teachers shift the distribution towards higher pregnancy incidence. To facilitate intuitive visualization, we graph the log of pregnancy incidence. This provides a visual of effects on pregnancy conditional on any pregnancy in the school.

We further examine distributional effects in Table 8 and Figure 20 with quantile treatment effects. We observe that effects shift in corresponding directions with average treatment effects, with teacher and near-peer effects diverging at all points on the distribution. We further observe that effects on pregnancy are largest at the upper end of the distribution, suggesting that baseline risk-seeking behaviors might interact with the information and sender of information.

We explore the degree to which belief updating mediates effects. In Table 9, we include school-level beliefs of whether older partners are riskier as an interaction term. Figure 21 shows these beliefs immediately after the intervention (a detailed version of Figure 8). Near-peers shift the distribution of beliefs almost completely towards the upper tail, while teachers have high variance. We examine whether teachers who update beliefs are more effective in reducing risky sex. In Table 9, we see that the belief that older partners are riskier is correlated with reductions in pregnancy in general. We further observe that if beliefs do not update, the sex education information does not reduce risky sex. Thus, belief updating seems crucial for success. However, when teachers update student beliefs, this information backfires, with risky sex increasing, corroborating earlier findings.

Instrumental variables estimates. — We next explore the extent to which results are driven by degree of implementation. We use three core measures to capture the degree of implementation among teachers: attendance at training, commitment card signing, and returning of students' responses post intervention. All measures are imperfect but plausibly capture upper and lower bounds of implementation. 90.53 percent of schools randomly selected to receive the sex education message by teachers attended training, 73.59 percent scheduled implementation, and 30.24 percent of schools returned written student responses. For near-peers, there is full implementation. These estimates capture the local average treatment effect and are plausibly relevant to all teachers, given we observe limited selection of teachers who implement in Table 5.

We estimate the effects of implementation and outcomes Y_{ij} as follows:

$$Y_{ij} = \alpha + \beta_1 SexEd_j + \beta_2 SexEd_j \times I_j + \gamma \boldsymbol{X}_j + \delta_s + \varepsilon_{ij} \tag{3}$$

where I is one of our measures of implementation. Since implementation in the teacher arm may be endogenous to teacher motivation and related variables we instrument for I using random assignment to receive the sex education message by teachers.

Table 10 shows instrumental variables (IV) estimates along three measures of implementation: attendance, commitment card signing, and student responses. Our results demonstrate that the more likely teachers were to implement, the larger and more statistically significant the effects. The relative effect on pregnancy of teachers implementing the program increases from .253 significant at the 10 percent level to .426 significant at the 5 percent level. This pattern is consistent for junior and rural girls. This indicates that teacher effects are unlikely to be a null driven by weak or non-implementation, but rather small increases in pregnancy driven by the intervention. Of note, while larger effects might be a mechanical function of a smaller first stage, increased statistical significance points to an underlying mechanism driving these messenger effects.

Table 11 shows corresponding results on beliefs by degree of implementation to explore consistency in the dose-response relationship. We examine beliefs both immediately after the intervention and 12 months later. We see that the more likely teachers are to implement, the more beliefs converge with near-pears. For all randomly assigned teachers, beliefs update 12 months later 4.6 percentage points less for students who received the sex education message from teachers relative to near-peers, however teachers who attend training and commit to implement converge to near-peers, with only 3.9 and 2.9 percentage points less belief updating. The teachers most likely to implement, evidenced by student responses, achieve full convergence 12 months later, with no difference between near-peers and are in fact more effective at updating beliefs immediately after the intervention. Thus, we see a dose-response relationship where beliefs converge between near-peers and teachers the more

likely teachers are to implement. This contrasts to the relationship with risky sexual behavior, where the more likely teachers are to implement, and therefore the more beliefs update, the more pregnancy outcomes diverge. These findings corroborate intent-to-treat estimates showing messengers seem to drive effects in opposite directions.

5 Discussion and Mechanisms

When near-peers deliver the sex education message, there is significant belief updating, and a large corresponding reduction in risky sex as measured by pregnancy between 40 to 65 percent. These results indicate that a sex education message that reveals safer sex options can be significantly more effective than status-quo abstinence messaging. This supports the theory that sexual behavior is more likely to change along the intensive margin. Moreover, our results in Botswana are similar to Kenya ten years prior, indicating that this phenomenon is remarkably robust across time and geography.

However, when teachers deliver the sex education message, while beliefs update in the same direction, we see a null effect on pregnancy and a plausible increase – a possibility that is important to explore given the large ramifications of such a finding since teacher-led sexual health programs are being implemented globally. This indicates that while near-peers and teachers are similarly effective agents for information transmission, they differ in their effectiveness in changing sexual behavior. To this end, both the message and, crucially, the messenger matter. We explore explanations that might drive messenger effects.

5.1 The Teenager's Dilemma: Listen, Ignore, or Rebel

In response to the sex education message, students might listen, not respond, or rebel. If students listen to the sex education message, they will engage in safer sex. If students rebel, they might ignore the message or intentionally engage in riskier sex. We observe effects on pregnancy in line with this prediction, presented earlier in the results section.

The possibility of teenage rebellion has been documented extensively in the psychology and sociology literature.[26] This insight has spawned a technique used by the adept parent: reverse psychology. Another parental favorite is soliciting a friend to give the desired prescription in the hopes their children will listen to someone else. In both cases, the messenger matters.

In the context of economics, we explain this behavior by including a sense of agency in a teenager's utility function, in a variation of the identity economics framework (Akerlof and Kranton, 2000). Agency is defined as the capacity to act independently and make one's own

[26] Milgram, 1963; Brehm, 1966; Pennebaker and Sanders, 1976; Bushman and Stack, 1996.

choices.[27] Agency is threatened when a prescription is given by a person with authority. Adhering to the prescribed action demonstrates lack of agency, since one 'has to,' resulting in disutility. Agency can be demonstrated by ignoring the prescription, demonstrating partial agency, or performing the opposite action, demonstrating full agency, both enhancing utility. Rebellion will occur if more utility is derived from exerting agency by doing the exact opposite of the given prescription, even if the given prescription will make you better off. [28]

When a parent tells their child not to drink or smoke they might now be motivated to try. When a teacher tells you to do your homework 'or else', you might intentionally skip it. In all these instances, there are clear economic payoffs to listening to the given advice, yet teenagers might forgo them and do the exact opposite to exercise maximum agency. This simple conceptual framework illuminates the possibility of rebellion – a phenomenon where people do not listen or do the opposite of a given prescription even when listening is in their best interest. We find suggestive evidence such a response might be triggered by teachers.

5.2 Additional Results

We provide related evidence of a listening versus rebellion response depending on the messenger by examining various dimensions of risky sexual behavior to supplement the pregnancy results.

First, we analyze self-reported sexual behavior. We explore effects on: number of sexual partners, if students have ever had sex in general and with older partners, if students are receiving gifts from a partner, if students feel a sense of self efficacy to say no to older partners, and condom use. Table 12 shows our results. In the near-peer arm girls exhibit greater empowerment and safer sexual behavior. Among girls in junior school this includes greater self-efficacy significant at the 1 percent level, being 18 percent less likely to have one or more sexual partner in junior school and 16 percent less likely to have had sex in the last 12 months. We further observe that rural girls in the near-peer arm are 25 percent less likely to date older partners.[29] In contrast, we see that the relative effect of teachers is always in the opposite direction as near-peers, and is significant and larger in a few cases. Teachers

[27] A possible outcome is that teachers have a null effect on pregnancy rather than increase it. This might be explained by students demonstrating 'partial agency' rather than 'full agency.' The near-peer arm demonstrates that in this particular context and for this particular sex education message belief updating *does* lead to changes in sexual behavior. Thus, it is unlikely environmental factors, such as financial constraints, are preventing information from translating to sexual behavior outcomes. Instead, a null could be driven by students demonstrating partial agency by ignoring teacher advice, rather than demonstrating full agency by rebelling against it.

[28] An interesting question to explore is how much of the rebellion effect might be demonstrating agency for social image versus self-image. While we do not have direct evidence to answer this question, we hypothesize that in our context, social image is likely to be less tractable since rebellious actions taken are not directly and clearly observed. Whereas an action like not getting good grades is visible and often publicly shared in school settings, risky sexual encounters often occur in private, are taboo and rarely discussed publicly.

[29] We further observe no changes in condom usage. This suggests that girls in the near-peer arm are more likely to delay sexual debut and not date at all or delay sexual debut when dating same-age partners, rather than negotiating for condom use.

increase the likelihood of receiving gifts from older partners relative to near-peers by 16.7 percent. While results for near-peers alone are not significant, since results diverge between near-peers and teachers, the relative effect is significant. A similar pattern is observed for the likelihood of ever having sex for rural girls. While we interpret these results with caution given that self-reported behavior is susceptible to social desirability bias, these results support the notion that students listen to near-peers and engage in safer sexual behavior and rebel against teachers, exhibiting slightly riskier sexual behavior.

Second, we explore partner matching. If students listen to the sex education message, they will follow the specific safer sex strategy proposed: dating younger partners. While we do not have reliable data on partner age or observe older partner behavior, we observe behavior among school-going boys. We examine effects for school-going boys to indicate possible age-mate matching. Since boys do not fall pregnant, we use the next best behavioral outcome we have available: attendance at school. We assume that both spending time with and matching with girls is one of young boys' incentives to be present in school. Thus, we expect boys to mimic girls school-going behavior. If girls are in school since they are dating age-mates and haven't fallen pregnant, boys are more likely to be in school. In contrast, if girls shift to date older partners they are more likely to spend time out of school. In turn, their male peers are less likely to come to school since there is no one to spend time or match with. In line with this prediction, we use absenteeism among boys as a plausible corresponding behavioral response to pregnancy.

Table 13 shows our results. Results are not significant in the near-peer or teacher arm on their own. However, since effects go in opposite directions the relative effect of the near-peer versus teacher intervention is significant and shows 2 percentage points less absenteeism for junior school boys which translates into a 25 percent reduction. This behavioral response among boys is consistent with girls shifting to date lower-risk younger male partners whereas girls rebel against teachers and shift towards dating higher-risk older partners.

5.3 Alternative Mechanisms

Teacher risky behavior. — It is possible teachers are stimulated to engage in sexual activity with students by the training, thereby increasing pregnancy. However, this is unlikely for a few reasons. First, as shown in the Table 1, 80 percent of the teachers are female. Thus, most teachers are biologically unable to impregnate school-going girls. Moreover, the gender distribution is nearly equivalent across near-peer and teacher arms. Thus, it is unlikely the gender dimension of the messenger is driving sexual behavior. Most conclusively, in the few schools that had male teachers, there were zero pregnancies.

Teachers distort the message. — Teachers might deliver the message, but distort it. One possibility is that teachers revert to status-quo abstinence messaging. If this were the case, however, this would not drive differences relative to the control group, where the message is also abstinence. Moreover, our results, as highlighted in Figure 11 and 13, provide evidence that even a year later, students in the teacher arm learn that older partners are riskier and update their beliefs along the intensive margin (*relative* rather than *absolute* risks), learning that older partners are riskier than younger partners. To this end, beliefs across the near-peer and teacher arms update in a similar direction and to a similar degree.

A related possibility is that teachers deliver a skewed message, emphasizing the risk of the *oldest* men and de-emphasizing the risk of *slightly older* men, thus inadvertently triggering entry into high-risk relationships with slightly older men. This possibility is captured in the conceptual framework in Dupas (2011) where revision of beliefs downwards might trigger risky behavior in certain instances. For example, teachers might update HIV risk perceptions towards 40-year-olds away from 20-year-olds making 20-year-olds appear safer than students initially thought. In contrast, near-peers might update risk perceptions of all older partners, guarding against all high-risk partners. This pattern of belief updating across arms might explain why near-peers reduce pregnancy whereas teachers increase it. While possible in theory, we find this explanation is unlikely to explain differences between teachers and near-pears in our context. Figure 22 shows that near-peers revise beliefs of 40-year-olds *upwards* relative to teachers and revise risk perceptions of 20-year-olds *downwards* relative to teachers. Thus, near-peers are more likely than teachers to de-emphasize the risk of slightly older men, not teachers, yet still reduce pregnancy.

Proximity. — It is possible students are most likely to act on beliefs when transferred by proximate messengers. Near-peers are more proximate than teachers to students on a bundle of characteristics such as age. Thus, as teachers converge towards near-peers, we might expect effects to converge. We provide descriptive evidence in Table 14 of effects on pregnancy by degree of teacher proximity to students. Since teacher characteristics might be endogenous we interpret these results as descriptive. We observe no relative effect of age and gender on pregnancy. However, we observe that for each additional year of experience a teacher has effects on pregnancy increase. This relationship holds for junior school girls. This indicates that proximate teachers are no more effective, while more entrenched teachers drive pregnancy in the opposite direction. This suggests that since teachers hold power and authority over students they might generate a rebellion response, regardless of proximity.

We further exploit an embedded randomized trial in our endline survey to measure the effects of proximity. At endline, we randomized enumerators to schools to ensure balance across treatment arms. In addition to ensuring balance across arms, this provides an exogenous shock on a bundle of observable enumerator characteristics, such as age and gender.

Thus, we can estimate the causal effect of messenger characteristics and relative proximity in characteristics between recipients and senders of information on survey reporting behavior. While this trial does not strictly translate to messenger effects in the main trial, it is suggestive and can illustrate possible mechanisms. Figure 23 plots the density of enumerator age, showing a range from mid 20s to late 30s. While this age range does not cover the full range of teachers, it overlaps with both near-peers and teachers providing suggestive evidence of how age might translate across messengers in the intervention.

Table 15 reports our results. We compare reporting on non-sensitive questions – such as knowledge of HIV risk – relative to sensitive questions such as reporting having ever had sex with an older partner. We interpret differential reporting on sensitive versus non-sensitive measures as 'freedom of expression' which relates to one's sense of agency. We find no effects of gender on reporting behavior. However, we find that while age does not affect reporting on mechanical HIV knowledge, age does have significant effects on reporting for sensitive questions: the older the enumerator, the lower the likelihood students reveal they have had sex or engaged in sensitive behavior. This indicates that messenger characteristics can affect freedom of expression. However, the more proximate the sender of information and receiver – such as when we restrict effects to students in junior school – effects are nearly identical to the average and the difference between primary and junior school students is not statistically significant. These results suggest that sender characteristics matter, but for reasons other than relative proximity. Moreover, these results reveal the importance of margins beyond beliefs, such as freedom of expression, which could affect whether students listen or rebel to the same set of information depending on the messenger.

Mechanisms across contexts. — It is interesting that we find first-order effects of the messenger whereas Dupas, Huillery, and Seban (2018) find limited effects by messenger for a set of sex education messages in Cameroon. It is challenging to adjudicate exact reasons for this difference. As Table 2 indicates, the contexts in Botswana, Cameroon and Kenya differ substantially, with up to a five times difference across dimensions such as income per capita, HIV rates, and sexual activity, so we are careful not to overstate drivers of variation without more conclusive evidence.

One possibility is that teachers are more trusted messengers in Cameroon. According to Demographic Health Survey (DHS) data condom knowledge is high in Cameroon at 82 percent relative to 65 in Botswana. However, comprehensive HIV knowledge is higher in Botswana at 48 percent relative to 31 percent in Cameroon, suggesting the direction of this effect is ambiguous. We further explore data from the Wellcome Trust Global Monitor Data in 2018 which explicitly measures levels of trust. Among young people aged 15-29 in Botswana 57 percent trust science whereas 69 percent do in Cameroon. In contrast, slightly more youth in Botswana trust external sources such as NGOs than in Cameroon (54 relative

to 47 percent). Since factual information is relatively less trusted in Botswana, but certain social actors are trusted more, this suggests the interaction between the type of information and its source might matter more in Botswana.

It is also possible that specific messenger profiles matter. In Cameroon, for example, many teachers are contract teachers. Data from PASEC in 2011 reveal that 83 percent of primary school teachers in Cameroon are non-permanent (Bashir et al. 2018). In contrast, in Botswana less than 10 percent of teachers are non-permanent (UNICEF 2019). These trends are consistent with significantly higher rates of contract teachers in Francophone Africa than East and Southern Africa (Bashir et al. 2018). Since contract teachers are temporary and often younger and less formally trained, this suggests teachers in Cameroon might in practice converge in characteristics to consultants or near-peers, which might lead to less difference among messengers in Cameroon than we observe in the Botswana trial.

To more fully explore explanations, data could be collected in the future on student perceptions and messenger profiles across contexts, and specific aspects of sender profiles experimentally tested. While such a coordinated multi-study design is rare it is possible as done in the ultra-poor graduation model across six countries (Banerjee et al. 2015). This study highlights that the messenger is first-order in some contexts, which could motivate future large-scale experiments around this topic.

6 Conclusion

Sex education is a critical component of a healthy society. A growing body of evidence exists on the *types* of information that are most effective, such as providing safe sex options rather than promoting abstinence. However, less is known about the *sender* of this information. In this paper, we test the importance of both the type and the sender of information.

We find that when a sex education message is delivered by near-peers, it reduces pregnancy by over 40 percent in Botswana relative to a 28 percent reduction for a similar message in Kenya ten years prior. This suggests that providing safe sex options can reduce risky sex across diverse contexts. Over the last decade, there have been dramatic changes in the cost and of salience HIV with widespread ARV medication. Moreover, there are substantial differences in the opportunity cost of leaving older partners in a Botswana today, a middle-income country, relative to Kenya ten years ago, a low-income country. Both of these parameters mark significant contextual changes across the Kenya and Botswana trials. Despite significant heterogeneity in context, our results suggest that the mechanism that sexual behavior changes on the intensive margin can generalize.

However, we find that within Botswana the messenger can make or break the sex education message. When near-peers deliver the message, pregnancy decreases while when teachers deliver the message there is a null effect and a plausible increase in pregnancy. This

indicates that the sender of information is first-order. We find that neither the compliance or beliefs margin explains these results. Both near-peers and teachers successfully implement and update beliefs as intended. Rather, the sender of information seems to have differential effects along the preferences channel. An analysis of potential mechanisms suggests that the messenger determines whether students listen to the message and respond in-kind or results in the information backfiring, with students rebelling. This is the case regardless of proximity between sender and receiver of information. Younger, female teachers are no more effective. Near-peers seem to be trust-worthy figures who do not threaten students' agency and are successful at persuading students to practice safer sex. Of note, this does not mean teachers do not have a role to play in sex education – they might be best placed to provide essential facts, such as what HIV is and how condoms work.

Our results contribute to a few literatures. First, we contribute to the sex education literature. The importance of the sender of information might help reconcile why despite field experiments in Africa which have found large responses to information, risk responses are much smaller at the population-level. One reason for this discrepancy might be that at the population-level teachers are almost always the default messenger. We also contribute to the literature on when information translates to behavior. The literature has to date focused on responsiveness to information in general. We demonstrate the role of the sender in ensuring the response is intended. We also contribute to the literature on persuasion. We find a persuasion rate of over 98 percent when a sex education message is delivered by near-peers, among the highest in a review by DellaVigna and Gentzkow (2010), emphasizing the first-order importance of the sender of information. We further contribute to the literature on external validity. Given the importance of the sender of information in our context, future research and policy should consider how and whether results translate across messengers, as well as the more often examined dimensions of context-variation such as geography and time. Future research might also expand on theoretical frameworks for how the sender of information can affect behavioral responses, as well as provide causal evidence on the effectiveness of other possible messengers – such as religious leaders and parents.

Our findings have significant policy implications. Teachers are often the default messenger for sex education in schools worldwide. While teachers are effective at transferring information – their core business – they might not reduce risky sex. In contrast, near-peers have the potential to reduce pregnancy by 40 to 65 percent. Many societies have large-scale national service programs which already deploy tens of thousands of near-pears in schools. These schemes present a scalable model for near-peers to deliver sex education, and thus for governments to deliver large welfare gains.

References

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and identity." *The Quarterly Journal of Economics* 115, no. 3: 715-753.

Allcott, Hunt. 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics*. 130.3: 1117-1165.

Alsan, Marcella, Owen Garrick, and Grant Graziani. 2019. "Does diversity matter for health? Experimental evidence from Oakland." *American Economic Review* 109, no. 12: 4071-4111.

Bandura, Albert. 2001. "Social cognitive theory: An agentic perspective." *Annual review of psychology* 52, no. 1: 1-26.

Banerjee, Abhijit V., Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2019. "Using gossips to spread information: theory and evidence from two randomized controlled trials." *Review of Economic Studies.*

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

Bates, Mary Ann, and Rachel Glennerster. 2017. "The generalizability puzzle." *Stanford Social Innovation Review* 15.3.

BenYishay, Ariel, and A. Mushfiq Mobarak. 2018. "Social learning and incentives for experimentation and communication." *Review of Economic Studies*.

Bernard, Tanguy, Stefan Dercon, Kate Orkin, and Alemayehu Taffesse. 2014. "The future in mind: Aspirations and forward-looking behaviour in rural Ethiopia." Centre for Economic Policy Research.

Bettinger, Eric P., and Bridget Terry Long. 2005. "Do faculty serve as role models? The impact of instructor gender on female students." *American Economic Review* 95, no. 2: 152-157.

Bodley-Bond, C., and K. Cronin. 2013. "National youth service, employability, entrepreneurship and sustainable livelihoods: Overview of the national youth service landscape in Sub-Saharan Africa." Washington DC: Innovation in Civic Participation.

Brehm, Jack W. 1996. "A theory of psychological reactance." Oxford: Academic Press.

Bushman, Brad J., and Angela D. Stack. 1996. "Forbidden fruit versus tainted fruit: Effects of warning labels on attraction to television violence." *Journal of Experimental Psychology: Applied* 2.3: 207.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in action: teacher and health worker absence in developing countries." *Journal of Economic Perspectives* 20, no. 1: 91-116.

Chinkhumba, Jobiba, Susan Godlonton, and Rebecca Thornton. 2014. "The demand for medical male circumcision." *American Economic Journal: Applied Economics* 6, no. 2: 152-77.

Chong, Alberto, and Eliana La Ferrara. 2009. "Television and divorce: Evidence from Brazilian novelas." *Journal of the European Economic Association* 7, no. 2-3: 458-468.

Crawford, Vincent, and Joel Sobel. 1982. "Strategic Information Transmission," *Econometrica*, 50, 1431–1451

De Walque, Damien. 2014. *Risking Your Health: Causes, Consequences, and Interventions To Prevent Risky Behaviors.* Washington, DC: World Bank

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210: 2-21.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2019. "From local to global: External validity in a fertility natural experiment." *Journal of Business and Economic Statistics* (forthcoming).

DellaVigna, Stefano, and Matthew Gentzkow. 2010. "Persuasion: empirical evidence." *Annual Review of Economics.* 2, no. 1: 643-669.

Dewatripont, Mathias, and Jean Tirole. 1999. "Advocates." *Journal of Political Economy* 107, no. 1: 1-39.

Downs, Anthony. 1957. "An economic theory of political action in a democracy." *Journal of Political Economy* 65, no. 2: 135-150.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "Education, HIV, and early fertility: Experimental evidence from Kenya." *American Economic Review.*

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using randomization in development economics research: A toolkit." *Handbook of Development Economics*: 3895-3962.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives work: Getting teachers to come to school." *American Economic Review* 102.4: 1241-78.

Dupas, Pascaline, Elise Huillery, and Juliette Seban. 2018. "Risk information, risk salience, and adolescent sexual behavior: Experimental evidence from Cameroon." *Journal of Economic Behavior & Organization* 145: 151-175.

Dupas, Pascaline. 2011. "Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya." *American Economic Journal: Applied Economics* 3.1: 1-34.

Eble, Alex, and Feng Hu. 2019. "Stereotypes, role models, and the formation of beliefs." *CDEP-CGEG Working Paper* 43.

Gallant, Melanie, and Eleanor Maticka-Tyndale. 2004. "School-based HIV prevention programmes for African youth." *Social Science & Medicine* 58.7: 1337-1351.

Gentzkow, Matthew, and Emir Kamenica. 2016. "Competition in persuasion." *The Review of Economic Studies* 84, no. 1: 300-322.

Godlonton, Susan, Alister Munthali, and Rebecca Thornton. 2016. "Responding to risk: circumcision, information, and HIV prevention." *Review of Economics and Statistics* 98, no. 2: 333-349.

Grossman, Sanford J. 1981. "An introduction to the theory of rational expectations under asymmetric information." *The Review of Economic Studies* 48, no. 4: 541-559.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125.1-2: 241-270.

Jalan, Jyotsna, and E. Somanathan. 2008. "The Importance of Being Informed: Experimental Evidence on Demand for Environmental Quality." *Journal of Development Economics*, 87(1): 14–28.

Jensen, Robert. 2010. "The (perceived) returns to education and the demand for schooling." *The Quarterly Journal of Economics* 125.2: 515-548.

Jensen, Robert, and Emily Oster. 2009. "The power of TV: Cable television and women's status in India." *The Quarterly Journal of Economics* 124, no. 3: 1057-1094.

Kahneman, Daniel, Stewart Paul Slovic, Paul Slovic, and Amos Tversky, eds. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian persuasion." *American Economic Review* 101, no. 6: 2590-2615.

Ketting, Everett, and Olena Ivanova. 2018. "Sexuality Education in Europe and Central Asia: State of the Art and Recent Developments." Cologne, Germany.

Kerwin, Jason. 2019. "Scared straight or scared to death? The effect of risk beliefs on risky behaviors." *American Economic Review* (forthcoming).

Kirby, Douglas B. 2008. "The impact of abstinence and comprehensive sex and STD/HIV education programs on adolescent sexual behavior." *Sexuality Research & Social Policy* 5.3: 18.

Kofoed, Michael S. 2019. "The Effect of Same-Gender or Same-Race Role Models on Occupation Choice Evidence from Randomly Assigned Mentors at West Point." *Journal of Human Resources* 54, no. 2: 430-467.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp. 2006. "Toward an understanding of the economics of charity: Evidence from a field experiment." *The Quarterly Journal of Economics* 121, no. 2: 747-782.

Lo, Nathan C., Anita Lowe, and Eran Bendavid. 2016. "Abstinence funding was not associated with reductions in HIV risk behavior in sub-Saharan Africa." *Health Affairs* 35.5: 856-863.

Luke, Nancy, and Kathleen Kurz. 2002. "Cross-generational and transactional sexual relations in sub-Saharan Africa." International Center for Research on Women.

Madajewicz, Malgosia, Alexander Pfaff, Alexander van Geen, Joseph Graziano, Iftikhar Hussein, Hasina Momotaj, Roksana Sylvi, and Habibul Ahsan. 2007. "Can Information Alone Change Behavior? Response to Arsenic Contamination of Groundwater in Bangladesh." *Journal of Development Economics*, 84(2): 731–54.

Meager, Rachael. 2019. "Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments." *American Economic Journal: Applied Economics* 11: 57-91.

Milgram, Stanley. 1963. "Behavioral study of obedience." *The Journal of Abnormal and Social Psychology* 67.4: 371.

Milgrom, Paul R. 1981. "Good news and bad news: Representation theorems and applications." *The Bell Journal of Economics*: 380-391.

Ministry of Basic Education. 2016. "Second Botswana Youth Risk Behavioural and Biological Surveillance Report." Gaborone, Botswana.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer. 2008. "Coarse thinking and persuasion." *The Quarterly Journal of Economics* 123, no. 2: 577-619.

Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at scale." *Journal of Economic Perspectives* 31, no. 4: 103-24.

Munshi, Kaivan. 2004. "Social learning in a heterogeneous population: technology diffusion in the Indian Green Revolution." *Journal of Development Economics* 73, no. 1: 185-213.

Nelson, Philip. 1974. "Advertising as Information." *Journal of Political Economy* 729-754.

Olson, Rick. 2012. "Refocusing the Lens: Recognizing and Enhancing Agency around HIV Risk Avoidance with Adolescent Girls in Eastern and Southern Africa." *Adolescent Girls, Cornerstone of Society*: 165.

Oster, Emily. 2012. "HIV and sexual behavior change: Why not Africa?" *Journal of Health Economics* 31, no. 1: 35-49

Paula, Áureo De, Gil Shapira, and Petra E. Todd. 2014. "How beliefs about HIV status affect risky behaviors: Evidence from Malawi." *Journal of Applied Econometrics* 29, no. 6: 944-964.

Paul-Ebhohimhen VA, Poobalan A, van Teijlingen ER. 2008. "A systematic review of school-based sexual health interventions to prevent STI/HIV in sub-SaharanAfrica." *BMC Public Health;* 8:4

Pennebaker, James W., and Deborah Yates Sanders. 1976. "American graffiti: Effects of authority and reactance arousal." *Personality and Social Psychology Bulletin*: 264-267.

Peters, Jörg, Jörg Langbein, and Gareth Roberts. 2016. "Policy evaluation, randomized controlled trials, and external validity—A systematic review." *Economics Letters* 147: 51-54.

Pritchett, Lant, and Justin Sandefur. 2013. "Context matters for size: why external validity claims and development practice do not mix." *Journal of Globalization and Development*: 161-197.

Rosenzweig, Mark, and Christopher Udry. 2019. "External Validity in a Stochastic World: Evidence from Low-Income Countries." *Review of Economic Studies*.

Ross, David A., Bruce Dick, Jane Ferguson. 2016. "Preventing HIV/AIDS in young people: a systematic review of the evidence from developing countries." World Health Organization.

Rothwell, Peter M. 2005. "External validity of randomised controlled trials: "to whom do the results of this trial apply?" *The Lancet* 365.9453: 82-93.

Statistics Botswana. 2012. "Education Statistics Report 2012." Statistics Botswana.

Statistics Botswana. 2013. "Botswana AIDS Impact Survey IV (BAIS IV): 2013 Summary Results." Gaborone, Botswana.

Stigler, George J. 1961. "The economics of information." *Journal of Political Economy* 69, no. 3: 213-225.

Thornton, Rebecca L. 2008. "The demand for, and impact of, learning HIV status." *American Economic Review* 98.5: 1829-63.

Underhill, Kristen, Don Operario, and Paul Montgomery. 2007. "Abstinence-only programs for HIV infection prevention in high-income countries." *Cochrane Database of Systematic Reviews* (4).

UNAIDS. 2018. "Global HIV & AIDS statistics – 2018 fact sheet." Geneva, Switzerland.

UNESCO. 2015. "Emerging Evidence, Lessons and Practice in Comprehensive Sexuality Education: A Global Review." Paris, France.

UNICEF. 2019. Public Expenditure Review of the Basic Education Sector in Botswana.

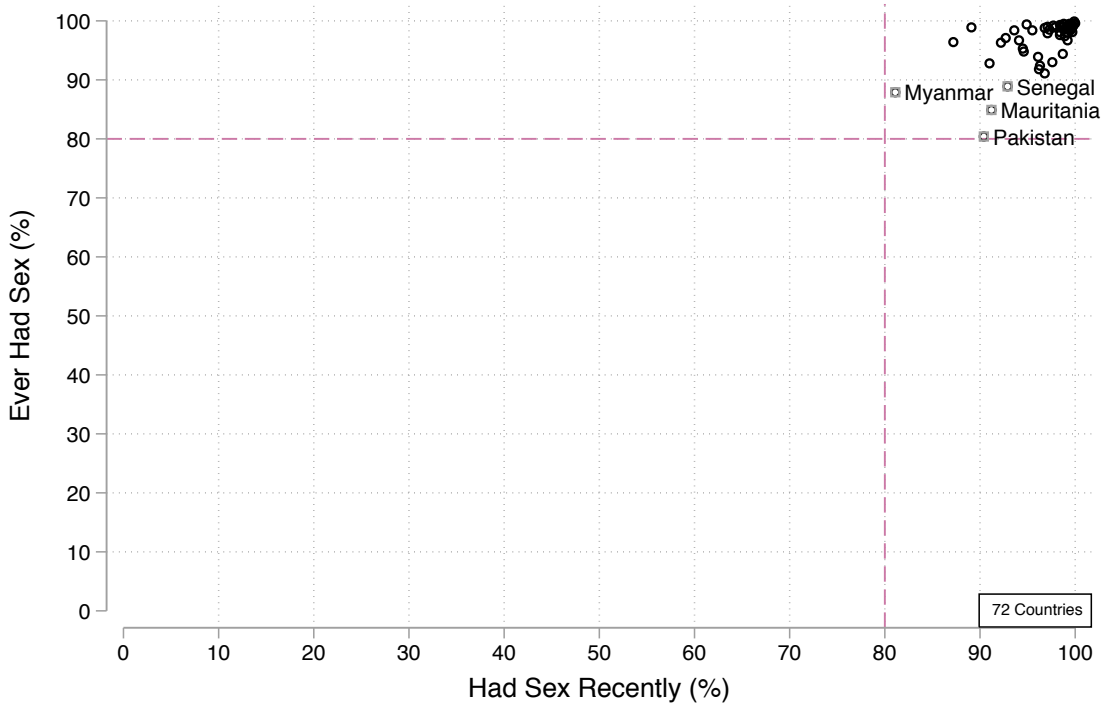
UNFPA. 2013. "Adolescent Pregnancy: A Review of the Evidence." New York.

Vivalt, Eva. 2015. "Heterogeneous treatment effects in impact evaluation." *American Economic Review*, *Papers and Proceedings,* 105.5: 467-70.

Wellcome Global Monitor. 2018. "How does the world feel about science and health."

Figures

Figure 1: Global Sexual Activity



Notes: This figure plots the percent of adults who have ever had sex along two indicators from the Demographic Health Survey (DHS) conducted across 72 countries. The first indicator ('ever had sex') is a question asking the age of first sexual encounter for adults aged 25 to 54. One option is 'never had sex.' We compute the inverse of this as the prevalence of having ever had sex. The second indicator ('had sex recently') is a question on frequency of sex, with an option being 'never had sex.' We again compute the inverse of this as the prevalence of having ever had sex. This indicator is consistently available across all countries for women only from ages 25 to 49. We plot both measures as a robustness test. We compute these measures using the most recent DHS data available for a given country. The median year is 2013. We label a few 'outliers' on the lower end of sexual prevalence on both indicators: Myanmar, Senegal, Mauritania, and Pakistan. We include dotted lines at 80 percent prevalence to illustrate this is the lower bound on sexual prevalence across both indicators.

Figure 2: Study Timeline

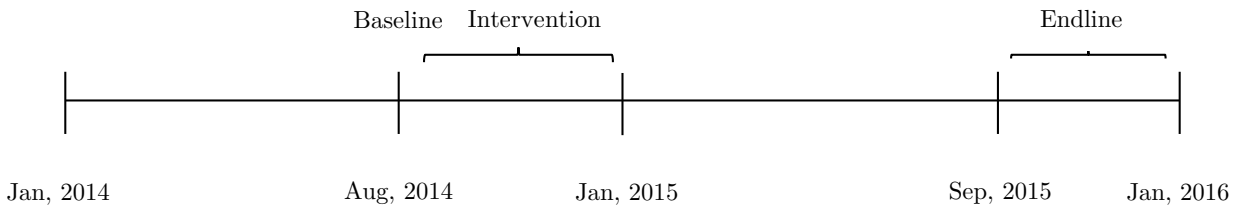
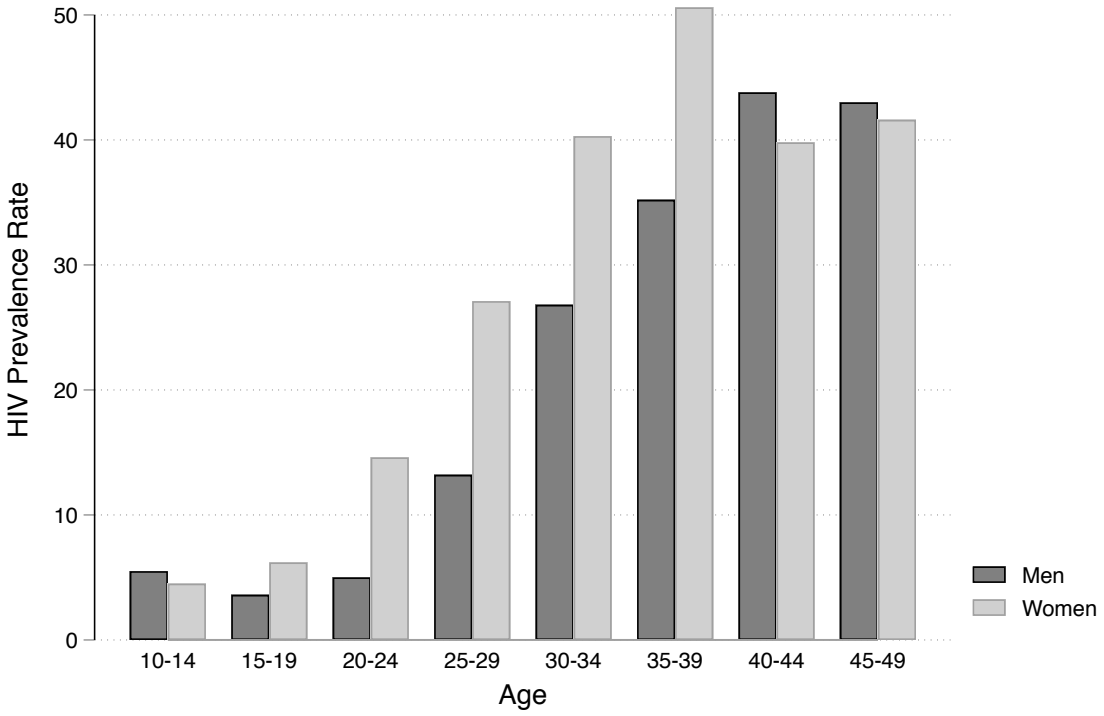
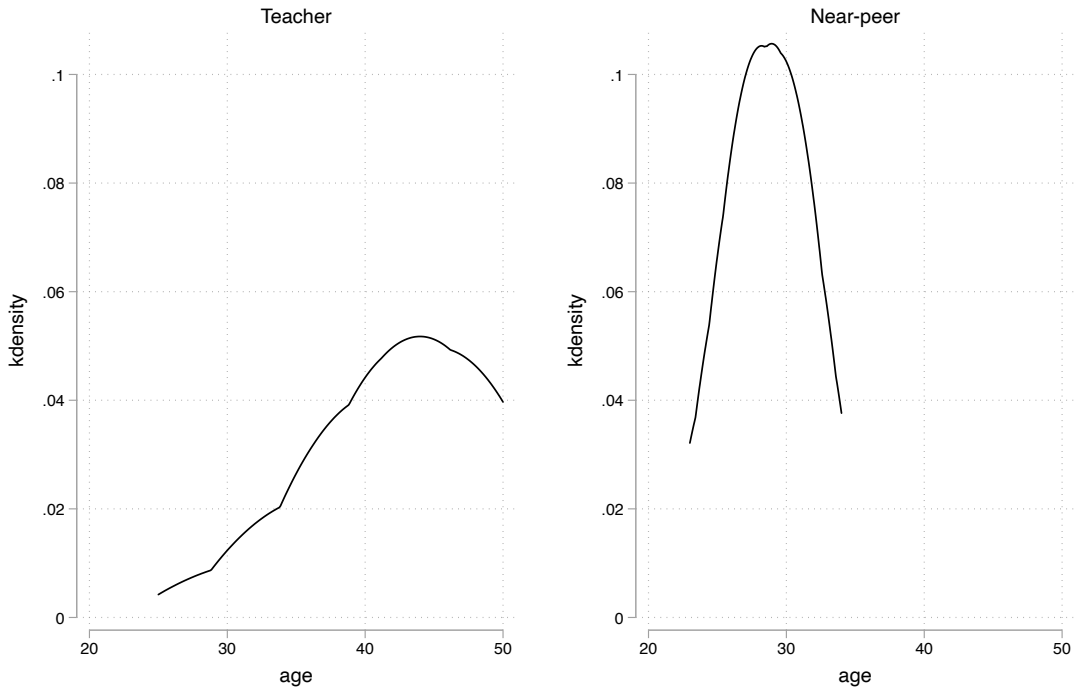


Figure 3: HIV prevalence by gender and age in Botswana



Source: Botswana AIDS Impact Survey (2013)

Figure 4: Messenger Age



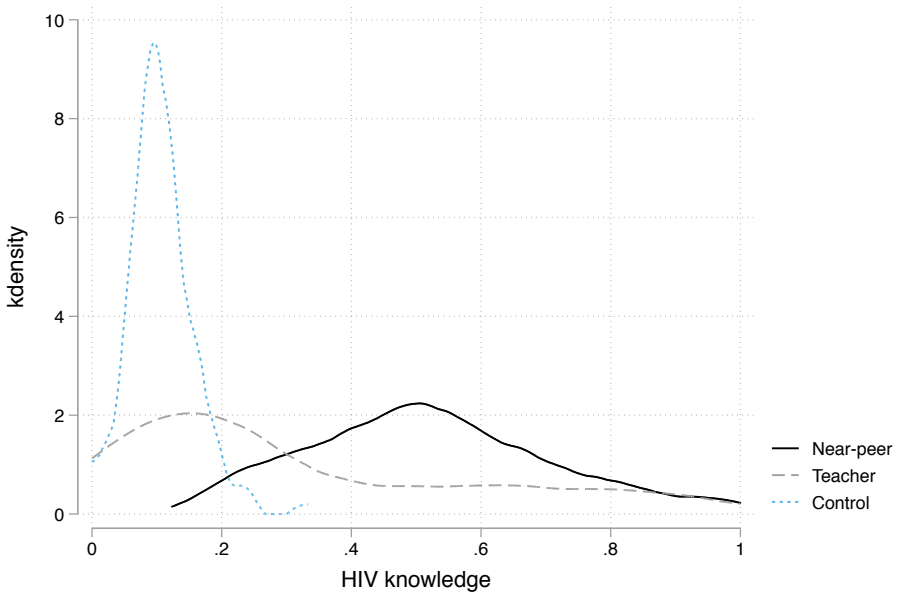
Notes: This figure plots the density of near-peers versus teachers age with bandwidth set equal to each respective standard deviation.

Figure 5: Messengers 'In Action'



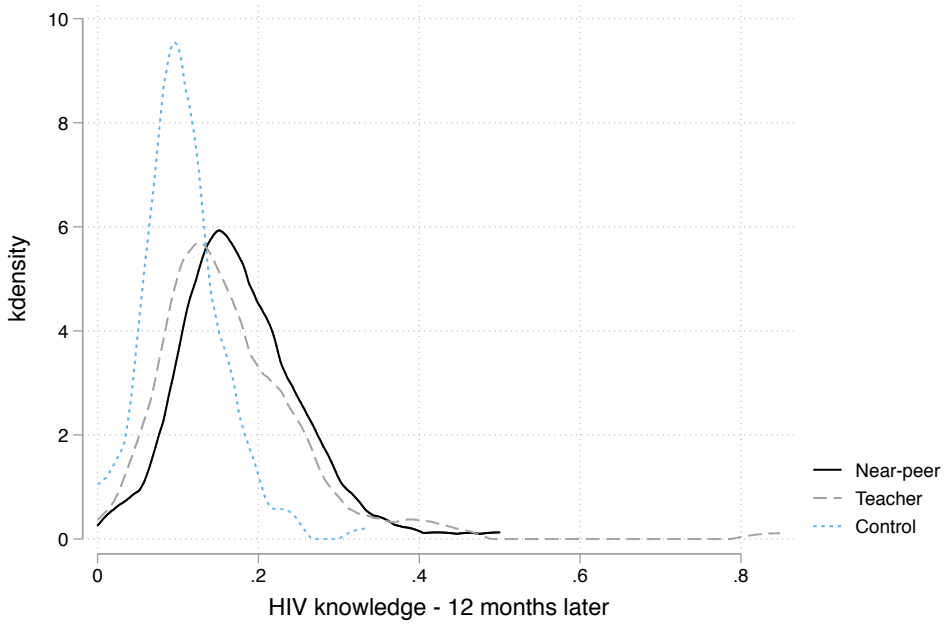
Notes: This figure shows a photo of a teacher and near-peer during training delivering the sex education message.

Figure 6: HIV Beliefs After the Intervention



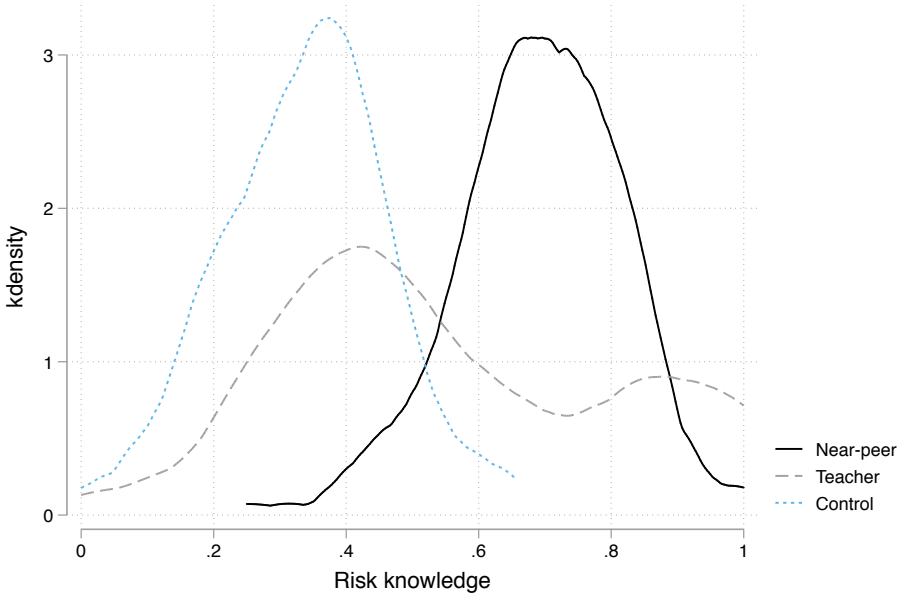
Notes: This figure presents treatment effects on beliefs measured by HIV knowledge immediately after the intervention. HIV knowledge is defined as knowing 40-year-old men are most likely to have HIV. Belief outcomes include all girls.

Figure 7: HIV Beliefs 12 Months Later



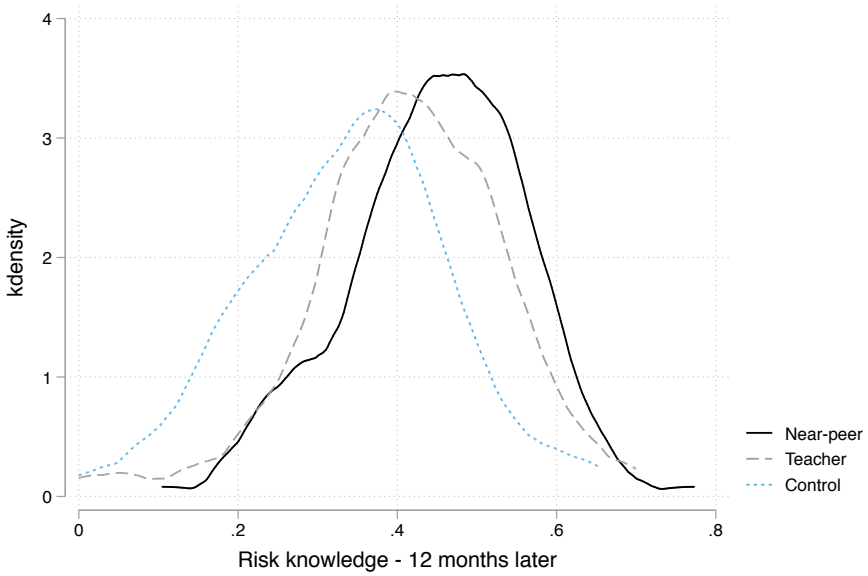
Notes: This figure presents treatment effects on beliefs measured by HIV knowledge 12 months after the intervention. HIV knowledge is defined as knowing 40-year-old men are most likely to have HIV. Belief outcomes include all girls.

Figure 8: Risk Beliefs After the Intervention



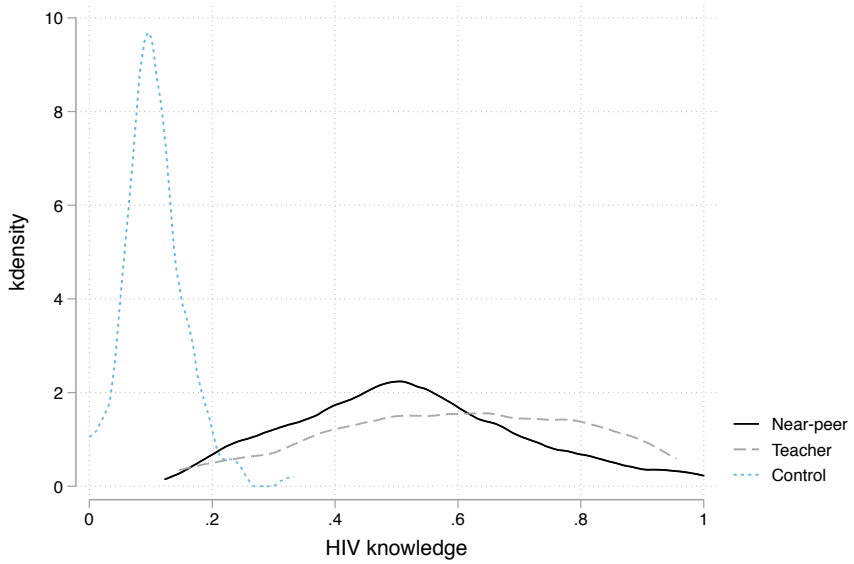
Notes: This figure presents treatment effects on beliefs measured by risk knowledge immediately after the intervention. Risk knowledge is defined as knowing older men are riskiest in general. Belief outcomes include all girls.

Figure 9: Risk Beliefs 12 Months Later



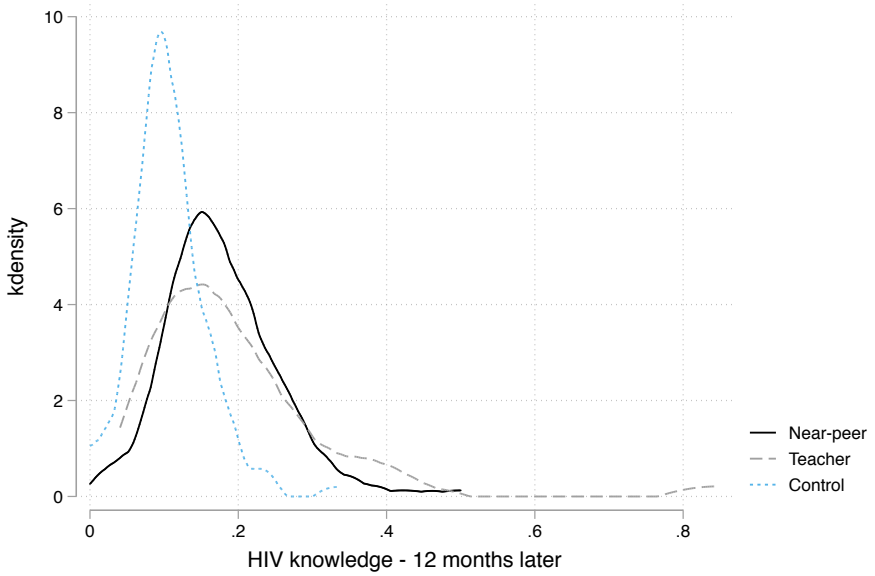
Notes: This figure presents treatment effects on beliefs measured by risk knowledge 12 months after the intervention. Risk knowledge is defined as knowing older men are riskiest in general. Belief outcomes include all girls.

Figure 10: HIV Beliefs After the Intervention, Conditional on Implementation



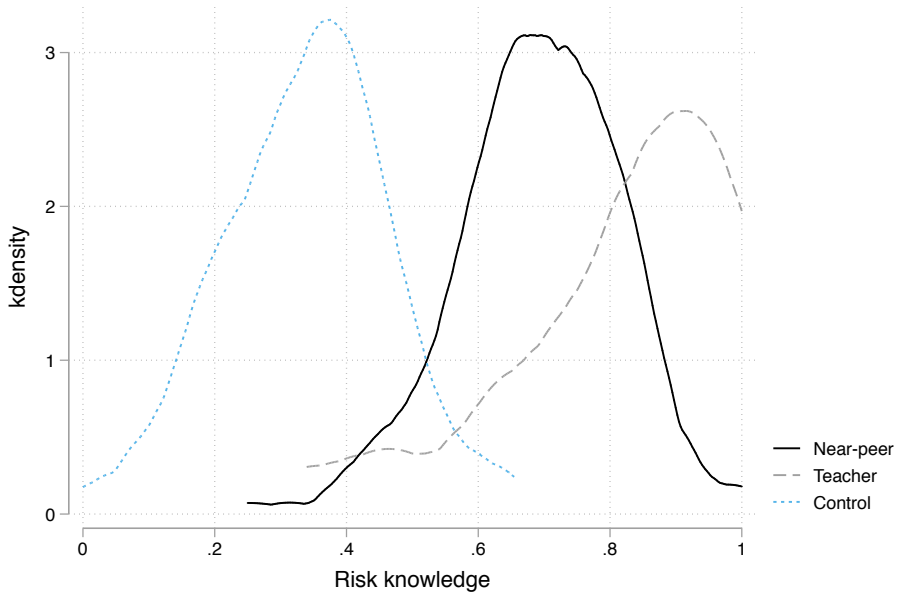
Notes: This figure presents treatment effects on beliefs measured by HIV knowledge immediately after the intervention. HIV knowledge is defined as knowing 40-year-old men are most likely to have HIV. Implementation is coded as submitting evidence of student responses after the intervention in the teacher arm. Belief outcomes include all girls.

Figure 11: HIV Beliefs 12 Months Later, Conditional on Implementation



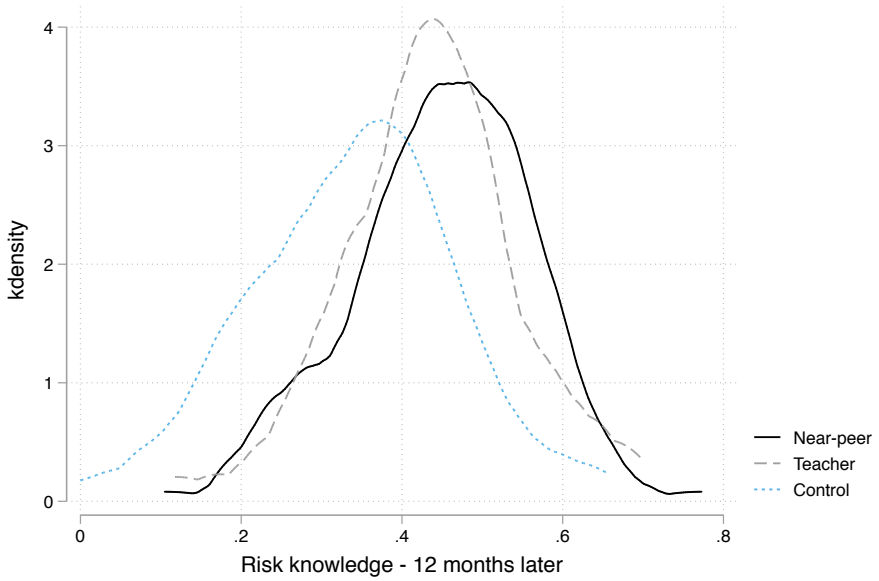
Notes: This figure presents treatment effects on beliefs measured by HIV knowledge 12 months after the intervention. HIV knowledge is defined as knowing 40-year-old men are most likely to have HIV. Implementation is coded as submitting evidence of student responses after the intervention in the teacher arm. Belief outcomes include all girls.

Figure 12: Risk Beliefs After the Intervention, Conditional on Implementation



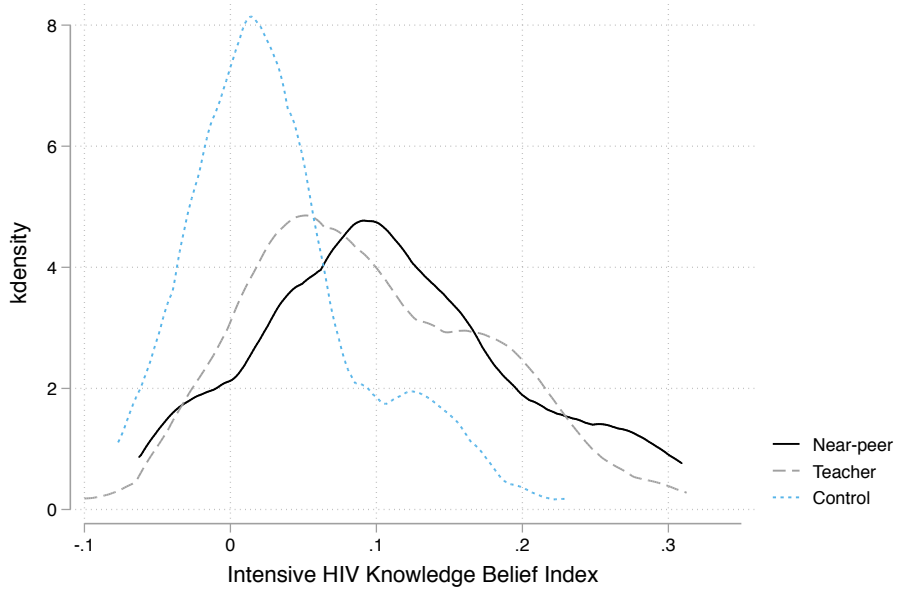
Notes: This figure presents treatment effects on beliefs measured by risk knowledge immediately after the intervention. Risk knowledge is defined as knowing older men are riskiest in general. Implementation is coded as submitting evidence of student responses after the intervention in the teacher arm. Belief outcomes include all girls.

Figure 13: Risk Beliefs 12 Months Later, Conditional on Implementation



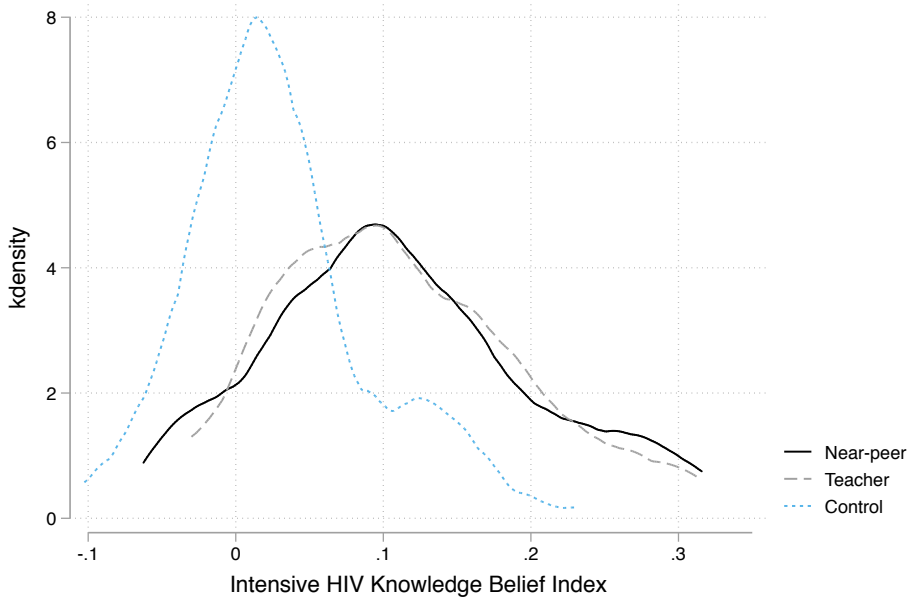
Notes: This figure presents treatment effects on beliefs measured by risk knowledge immediately 12 Months after the intervention. Risk knowledge is defined as knowing older men are riskiest in general. Implementation is coded as submitting evidence of student responses after the intervention in the teacher arm. Belief outcomes include all girls.

Figure 14: Beliefs of Relative Risks



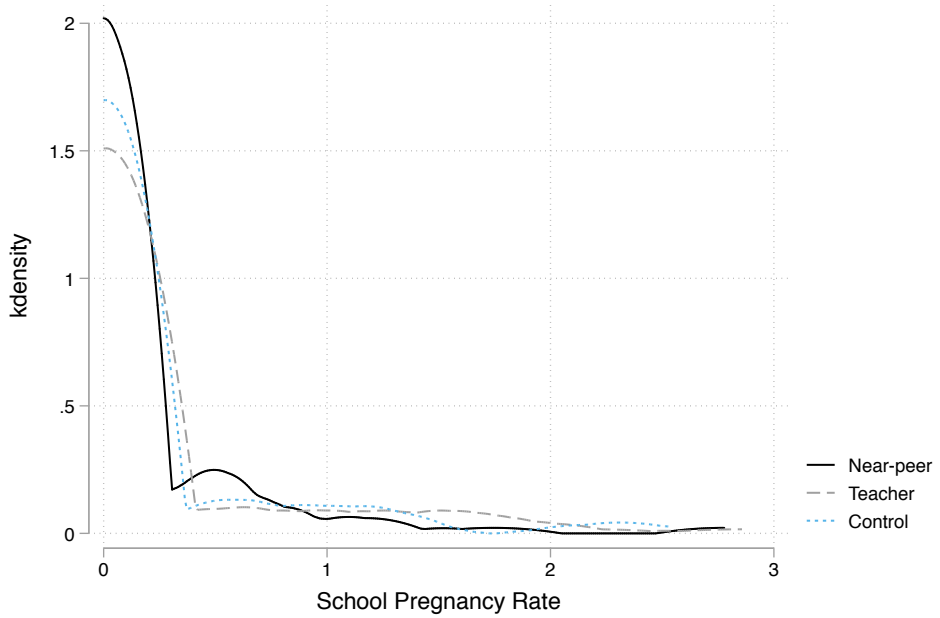
Notes: This figure presents treatment effects on beliefs updating as intended revealing relative risks between younger or older sexual partners. The index of beliefs captures the degree to which students think older partners have higher HIV than younger partners. The index is a linear subtraction of correctly identifying 40-year old men as mostly to have HIV minus incorrectly identifying 10-19-year-old men as most likely to have HIV. Cutoff outliers at 2.5 percentiles for clarity of visual exposition. This figure includes all girls.

Figure 15: Beliefs of Relative Risks, Conditional on Implementation



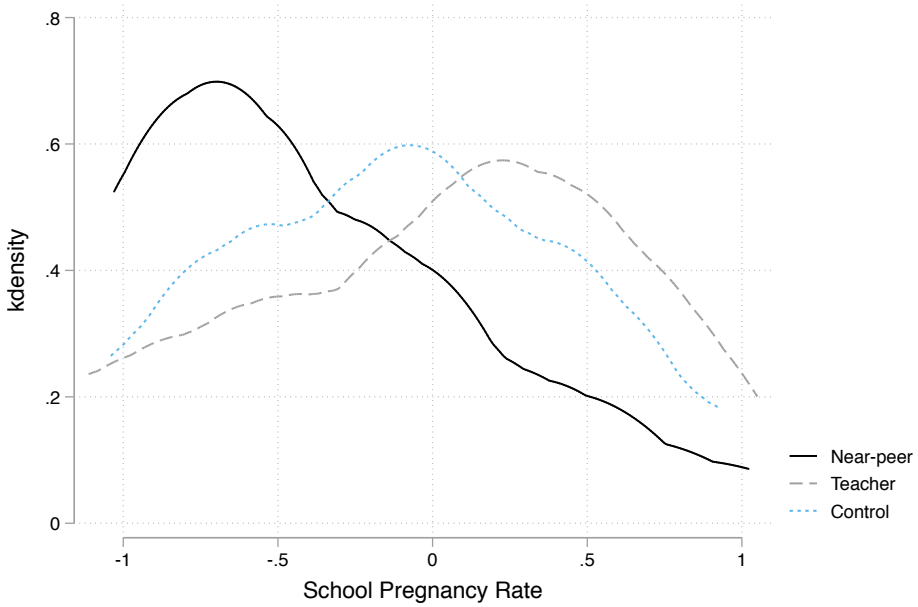
Notes: This figure presents treatment effects on beliefs updating as intended revealing relative risks between younger or older sexual partners. The index of beliefs captures the degree to which students think older partners have higher HIV than younger partners. The index is a linear subtraction of correctly identifying older men as mostly to have HIV minus incorrectly identifying same-age partners as most likely to have HIV. Implementation is coded as submitting evidence of student responses after the intervention in the teacher arm. Cutoff outliers at 2.5 percentiles for clarity of visual exposition. This figure includes all girls.

Figure 16: School Pregnancy Rate for All Girls



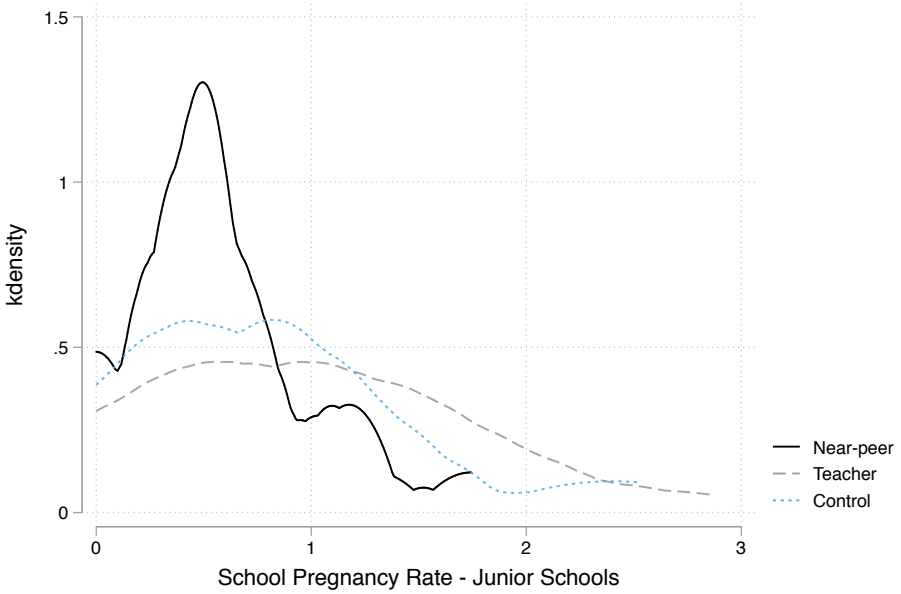
Notes: These figures plot the distribution of treatment effects on school-level pregnancy rates for all girls. School-level pregnancy rates along the x-axis are interpreted as the pregnancy incidence rate.

Figure 17: Log of School Pregnancy Rate for All Girls



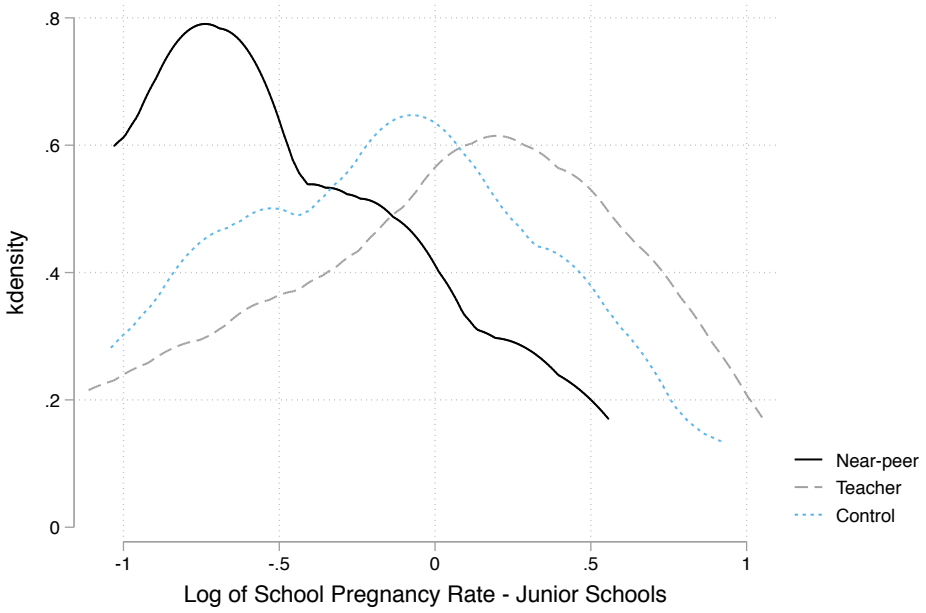
Notes: These figures plot the distribution of treatment effects on school-level pregnancy rates for all girls. School-level pregnancy rates along the x-axis are interpreted as the pregnancy incidence rate. The log of school pregnancy rate on the right panel captures pregnancy conditional on any pregnancy at the school. We remove extreme outliers (above 3 percent incidence rate) for clear visualization. The right panel is also designed to facilitate ease of visualization given the rare nature of the pregnancy outcome.

Figure 18: School Pregnancy Rate for Junior Girls



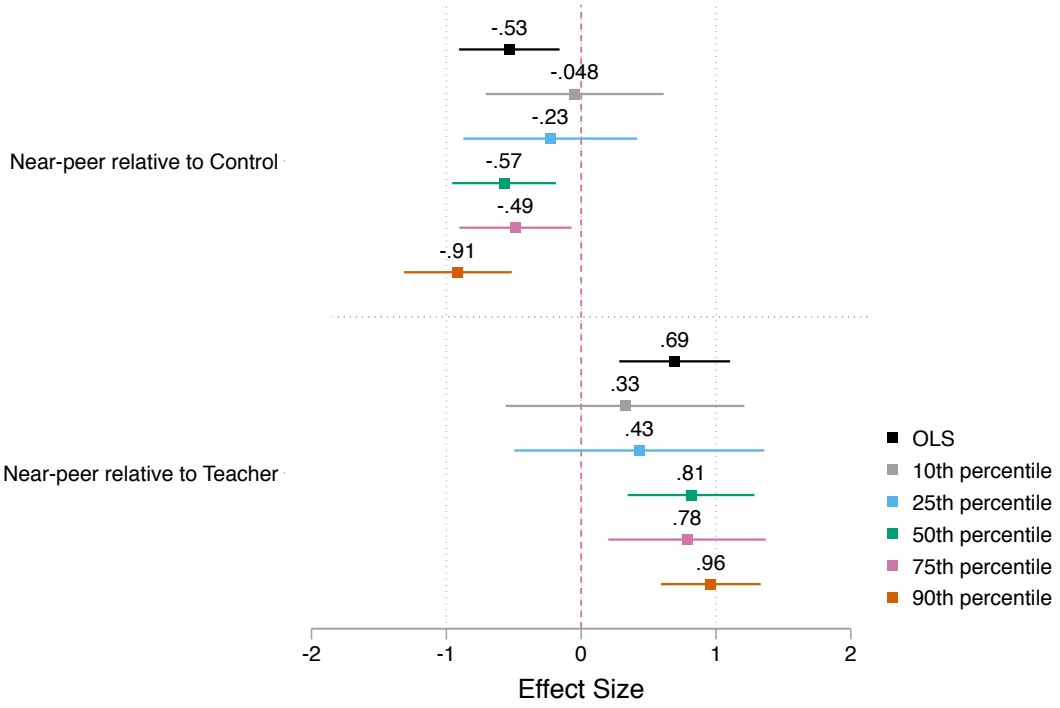
Notes: These figures plot the distribution of treatment effects on school-level pregnancy rates for junior girls. School-level pregnancy rates along the x-axis are interpreted as the pregnancy incidence rate.

Figure 19: Log of School Pregnancy Rate for Junior Girls



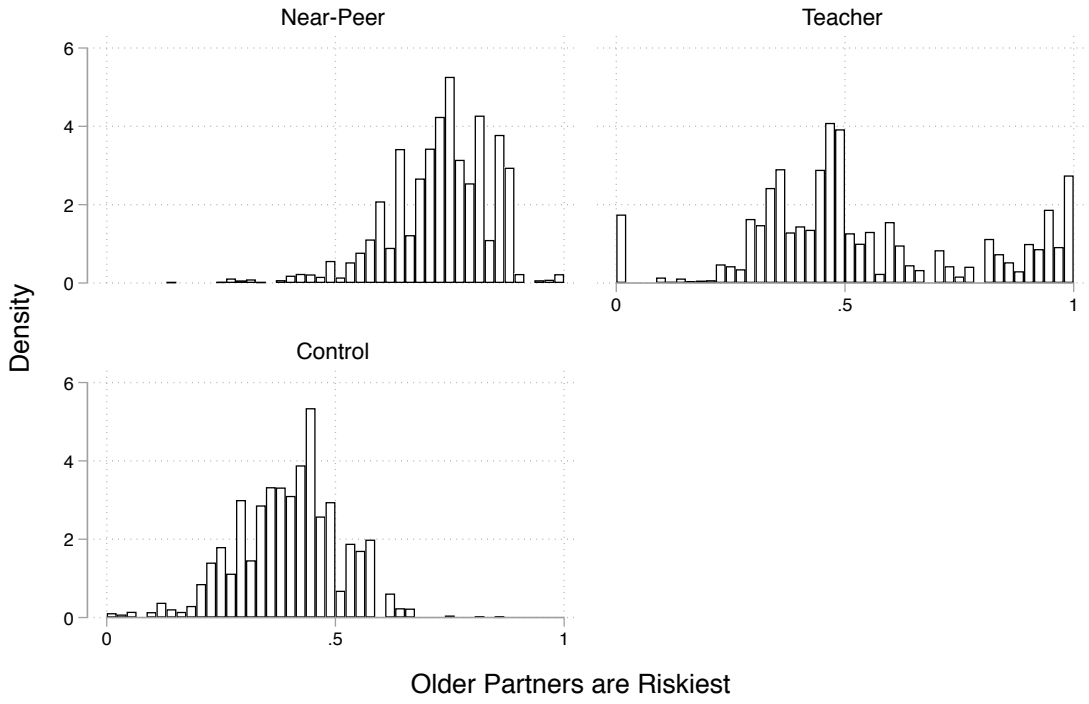
Notes: These figures plot the distribution of treatment effects on school-level pregnancy rates for junior girls. School-level pregnancy rates along the x-axis are interpreted as the pregnancy incidence rate. The log of school pregnancy rate on the right panel captures pregnancy conditional on any pregnancy at the school. The right panel is designed to facilitate ease of visualization given the rare nature of the pregnancy outcome.

Figure 20: Quantile Treatment Effects on Pregnancy



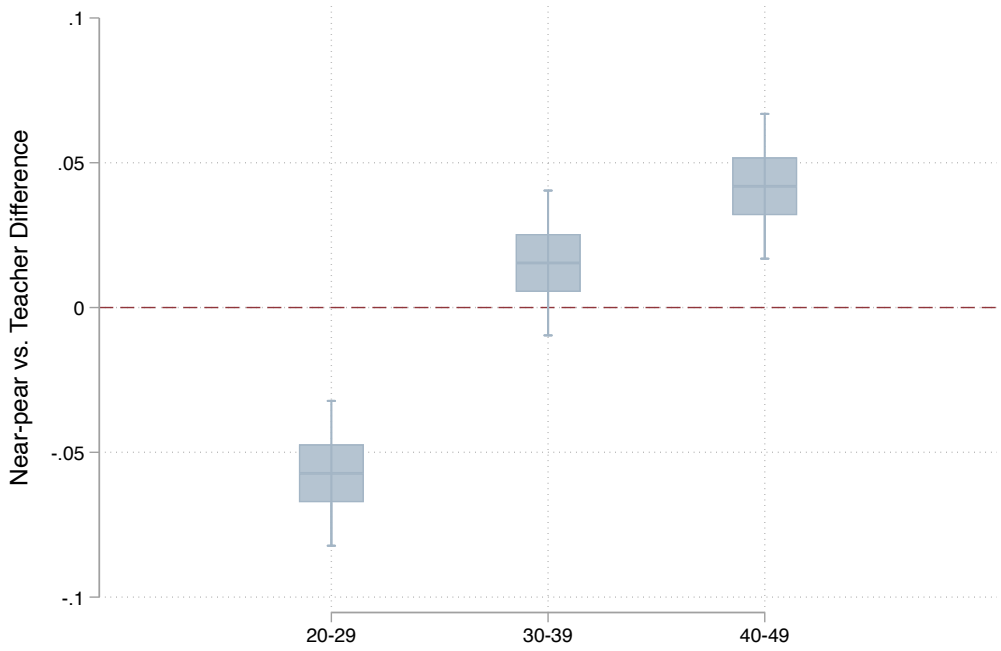
Notes: This figures plots OLS and quantile treatment effects on pregnancy. The dependent variable is a school-level continuous variable for the percentage of girls pregnant. This variable construction differs slightly from individual level pregnancy results with standard errors clustered at the school level. All specifications include strata dummies and a vector of school-level baseline control variables. We scale the dependent variable by a factor of one hundred for ease of interpretation of the coefficient as the direct incidence rate.

Figure 21: Beliefs about Risky Partners



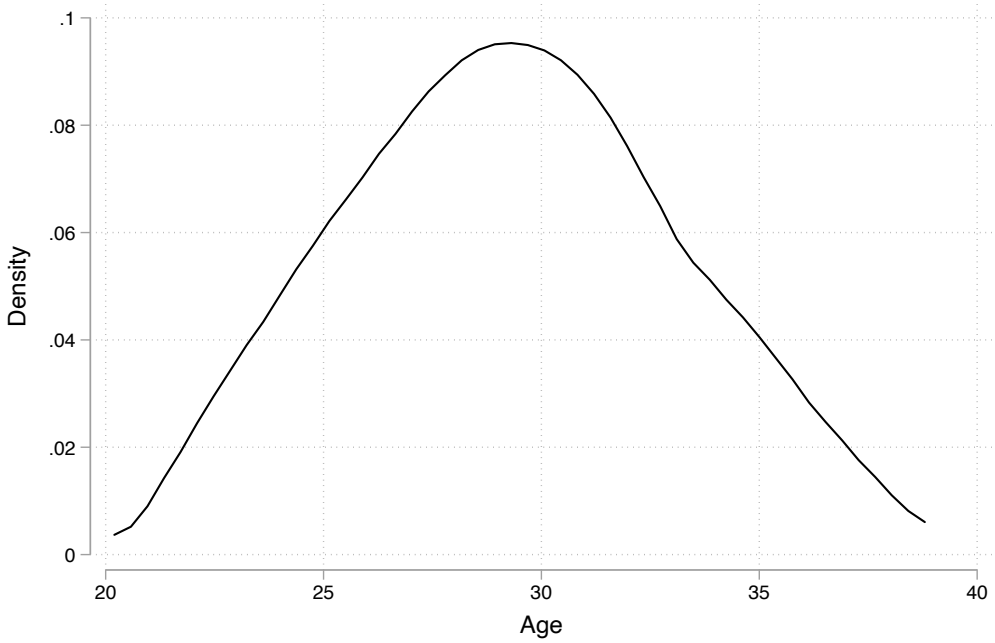
Notes: This figure plots the density of beliefs about risky partners immediately after the intervention. The x-axis represents the school-level rate of students knowing that 'older partners are riskiest.' It is immediately clear that in the near-peer arm over half of students have this belief in nearly all schools, while in the control arm less than half of students have this belief in nearly all schools. For teachers, there is high variance.

Figure 22: HIV Risk Perception of Older Partners



Notes: This figure plots the relative difference in HIV risk belief updating between near-peers and teachers. This is a linear index of risk beliefs in the near-peer arm at endline relative to risk beliefs in the teacher arm for the question "which age group of men is most likely to have HIV." We observe that near-peers update risk perceptions of 30-30-year-olds and 40-49-year-olds upwards relative to teachers and 20-20-year-olds downwards relative to teachers.

Figure 23: Messenger Age for Survey Enumerators



Notes: This figure plots the density of surveyor age across all three treatment arms with a bandwidth set equal to the standard deviation. Surveyors were randomly assigned to schools. This generates exogenous shocks to survey reporting on a bundle of messenger characteristics, such as surveyor age.

Table 1: Messenger Profile

	Teachers				Near-peers			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Female	0.80				0.88			
Age	43.80	5.40	25	50	28.56	2.68	23	34
Years Teaching	20.10	7.20	1.8	33.7	0.00	0.00	0.00	0.00
Highest Degree								
Diploma	0.59				0.25			
Bachelors	0.13				0.31			
Teaching Diploma	0.14				0.00			

Notes: This table reports descriptive statistics by messenger type. We omit redundant standard deviation, min and max descriptive statistics for dummy variables.

Table 2: Botswana Comparison to Kenya and Cameroon Contexts

	Kenya, 2004	Cameroon, 2011	Botswana, 2014
Panel A: National indicators			
GDP per capita (in current USD)	458.88	1403.23	7493.75
Percent of adults aged 15-49 living with HIV	0.07	0.04	0.23
Percent of all people living with HIV accessing ARVs	0.02	0.20	0.69
Total fertility rate (births per woman)	4.90	5.05	2.80
Gross primary enrollment ratio	1.02	1.07	1.05
Net primary enrollment rate	0.70	0.88	0.89
Population density (people per square kilometer)	61.60	44.22	3.80
Share of the population that lives in urban areas	0.21	0.52	0.66
Comprehensive HIV Knowledge	0.42	0.31	0.48
Share Non-permanent or Contract Teachers	0.24	0.83	0.10
Trust in Science	0.71	0.69	0.57
Trust in NGOs	0.66	0.47	0.54
Panel B: Study sample baseline: demographics, knowledge and behavior			
Age (years)	15.10	16.77	13.60
Class size	38.20	23.00	36.80
Share reporting having had sex	0.21	0.22	0.07
Sharing reporting having fallen pregnant	0.14	0.10	0.05
Share that think men over 25 have higher rates of HIV than teenage boys	0.29	0.39	0.50
Share that think condoms can prevent HIV	0.45	0.82	0.65

Source: National-level indicators are compiled by the World Bank drawing on estimates from UNAIDS, UNDP, UIS, FAO, World Bank national accounts data and the Demographic Health Surveys (DHS). Baseline from the study samples are drawn from Dupas (2011) for Kenya, from Dupas, Huillery, and Seban (2018) for Cameroon, and from the baseline in this study for Botswana. For the share of contract or non-permanent teachers this data for Cameroon and Kenya comes from the nearest year available, 2011 and 2007, respectively from SACMEQ and PASEC surveys. For Botswana, this data come from a report by UNICEF from data in 2017. Trust data from the Wellcome Global Monitor is only available in 2018 so we can only compare across countries, but is not specific to years across trials. We sum answers as "yes" or "a lot" and "some" as affirmative responses for the population group aged 15-29.

Notes: We compare national-level indicators and study samples in Kenya, Cameroon and Botswana along the most similar indicators available. For study-specific samples, we draw on numbers reported by Dupas (2011) and Dupas, Huillery, and Seban (2018) directly. Since some indicators are only reported along various dimensions of disaggregation such as gender and by treatment group, we report the most direct analogy from our study. Study indicators are reported for girls only while national-level indicators are reported for both genders. Class size is reported for the control groups in the Kenya and Botswana studies to ensure comparability and on average at baseline for Cameroon. The share of students that think men over 25 have higher rates of HIV than teenage boys is derived from our study by calculating the percentage of students who say 30-year old men and older are most likely to have HIV. While this is not the identical question reported in the other studies it captures a comparable indicator of knowledge of relative HIV risk.

Table 3: Baseline Summary Statistics

	Summary Statistics			Joint F-Test
	Near-peer	Teacher	Control	P-value
Panel A: Survey Data				
Gender	0.51	0.51	0.50	0.292
Age (years)	13.88	13.82	13.77	0.788
Class Size	36.18	36.40	36.76	0.724
School Size	307.88	302.44	297.31	0.727
School Absenteeism	0.05	0.05	0.05	0.518
Grade 6	0.34	0.36	0.37	0.774
Grade 8	0.32	0.31	0.32	0.869
Grade 9	0.31	0.30	0.29	0.674
40 year-old men have the highest HIV rate	0.10	0.10	0.10	0.886
Older partners are riskier	0.44	0.42	0.44	0.288
Self Efficacy	0.74	0.77	0.77	0.806
Ever had sex	0.12	0.12	0.11	0.277
Ever had sex with an older partner	0.06	0.05	0.05	0.749
Ever been pregnant - self reported	0.06	0.05	0.05	0.597
Girls pregnant in class - peer reported	0.10	0.11	0.12	0.044
Girls with older partners - peer reported	1.33	0.58	0.64	0.002
Panel B: Botswana Education Council				
Gender	0.50	0.50	0.51	0.759
Grade	49.05	49.28	48.80	0.868
Percent A	0.04	0.05	0.05	0.776
Percent B	0.13	0.13	0.13	0.793
Percent C	0.31	0.31	0.30	0.788
Panel C: Ministry of Education				
Female Enrollment	301.96	299.07	294.78	0.947
Male Enrollment	298.28	296.75	304.75	0.880
Total Enrollment	600.24	595.82	599.53	0.995
Pregnancy Dropout	0.32	0.31	0.34	0.932

Notes: This table reports summary statistics and p-values on a joint F-test across baseline characteristics from survey and administrative data for all students, including boys and girls. All standard errors are robust and clustered at the school level. For the joint F-test we control for classes that were joined together during surveying. Pregnancy dropout is interpreted as the direct incidence rate (e.g. .32 percent). All other dummy variables are interpreted as a percent (e.g. Percent A indicates 4 percent of students in the near-peer arm received an A on their last exam).

Table 4: Beliefs

	Beliefs Post Intervention		Beliefs 12 Months Later	
	(1)	(2)	(3)	(4)
	HIV Older	Risk Older	HIV Older	Risk Older
Panel A: Randomized				
Near-peer	0.442***	0.330***	0.088***	0.086***
	(0.017)	(0.014)	(0.007)	(0.013)
Teacher	0.174***	0.120***	0.042***	0.031**
	(0.038)	(0.037)	(0.009)	(0.016)
Observations	22,602	22,602	22,773	22,773
R-squared	0.532	0.452	0.282	0.280
Control Mean	0.100	0.399	0.100	0.399
P-value: near-peer vs. teacher	0.000	0.000	0.000	0.000
Panel B: Implemented				
Near-peer	0.439***	0.327***	0.087***	0.084***
	(0.015)	(0.012)	(0.007)	(0.013)
Teacher	0.528***	0.435***	0.077***	0.062*
	(0.034)	(0.024)	(0.019)	(0.033)
Observations	17,184	17,184	17,355	17,355
R-squared	0.813	0.806	0.339	0.340
Control Mean	0.100	0.399	0.100	0.399
P-value: near-peer vs. teacher	0.010	0.000	0.636	0.374

Notes: This table reports girls' beliefs immediately after the intervention and 12 months later by random assignment and for those who implement the intervention. Column (1) and (3) report beliefs that older man are more likely to have HIV and Columns (2) and (4) report beliefs that older partners are riskier in general. Implementation is coded if teachers hand back evidence of student responses post implementation. We further report selection on baseline school-level characteristics to explore selection effects for those who implement. All standard errors are robust and clustered at the school-level. All specifications are ordinary least squares, include strata fixed effects and a vector of school-level baseline control variables. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1

Table 5: Quantifying Selection for Teachers Who Implement

	Baseline Average School Characteristics				Teacher Characteristics				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Female Students	Test Scores	Girls Pregnant	Girls Dropped Out	Age	Female	Years of Experience	Bachelors	Diploma
Teacher Arm	-0.001	0.501	-0.108	-0.158	0.819	0.012	0.195	-0.001	-0.023
	(0.011)	(0.748)	(0.122)	(0.383)	(0.659)	(0.047)	(0.808)	(0.038)	(0.052)
By Implementation Degree									
Attended Training	0.003	0.839	-0.117	-0.008	0.870	-0.021	0.468	0.013	-0.016
	(0.011)	(0.775)	(0.126)	(0.397)	(0.681)	(0.048)	(0.837)	(0.039)	(0.054)
Commitment Card	-0.006	0.506	-0.062	0.148	0.328	0.010	0.034	-0.005	-0.027
	(0.012)	(0.838)	(0.137)	(0.428)	(0.733)	(0.052)	(0.901)	(0.042)	(0.058)
Student Responses	0.019	0.033	-0.083	0.385	0.558	0.061	-0.361	-0.043	-0.104
	(0.016)	(1.110)	(0.180)	(0.563)	(0.965)	(0.068)	(1.201)	(0.056)	(0.077)
Observations	339	339	343	343	340	337	336	339	339
Control Mean	0.500	52.38	0.403	1.965	43.03	0.789	20.80	0.132	0.610

Notes: This table reports effects of teacher selection by baseline average school characteristics as well as teacher characteristics. Baseline data on dropouts and pregnancy are from the Ministry of Education and baseline test scores are from the Botswana Education Council. Teacher characteristics data comes from endline surveys of guidance and counselling school teachers across all arms. We code non-implementation by each measure of implementation as a zero. We see no evidence of selection across all indicators, with no differential characteristics by any degree of implementation. All specifications include strata dummies. Each row is a separate regression, comparing teachers in general and teachers who implement to the near-peer and control groups to identify potential selection effects. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Beliefs of Relative Risks

	All Girls		Junior Girls		Rural Girls	
	(1)	(2)	(3)	(4)	(5)	(6)
	HIV Knowledge	Risk Knowledge	HIV Knowledge	Risk Knowledge	HIV Knowledge	Risk Knowledge
Panel A: Randomized						
Near-peer	0.110***	0.102***	0.103***	0.084***	0.103***	0.117***
	(0.009)	(0.016)	(0.011)	(0.024)	(0.011)	(0.018)
Teacher	0.054***	0.034**	0.028**	0.003	0.052***	0.035*
	(0.011)	(0.017)	(0.013)	(0.021)	(0.012)	(0.019)
Observations	21,041	21,175	13,502	13,587	16,086	16,196
R-squared	0.011	0.008	0.011	0.007	0.010	0.011
Control Mean	0.019	0.279	0.019	0.279	0.019	0.279
P-value: near-peer vs. teacher	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: Implemented						
Near-peer	0.106***	0.098***	0.101***	0.081***	0.101***	0.113***
	(0.009)	(0.017)	(0.012)	(0.027)	(0.011)	(0.018)
Teacher	0.089***	0.078**	0.080***	0.063	0.093***	0.092***
	(0.021)	(0.032)	(0.025)	(0.040)	(0.022)	(0.030)
Observations	15,923	16,008	9,993	10,044	12,316	12,384
R-squared	0.013	0.011	0.013	0.008	0.012	0.015
Control Mean	0.019	0.279	0.019	0.279	0.019	0.279
P-value: near-peer vs. teacher	0.340	0.180	0.271	0.175	0.644	0.201

Notes: This table reports treatment effects on belief updating as intended: revealing relative risks between younger or older sexual partners. The relative risk belief is calculated as a linear index of beliefs subtracting identification of older partners as risky from younger partners as risky in general. A similar index is calculated for knowledge of HIV risk. Results presented are for beliefs 12 months after the intervention as a conservative lower bound for beliefs immediately after the intervention. We report results by random assignment and for those who implement the intervention. Implementation is coded by whether teachers handed in student responses after the intervention. All standard errors are robust and clustered at the school-level. All specifications are ordinary least squares, include strata fixed effects and a vector of school-level baseline control variables. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Treatment Effects on Pregnancy

	OLS		
	(1)	(2)	(3)
	Girls	Junior Girls	Rural Girls
Sex Education	-0.214**	-0.444***	-0.365***
	(0.095)	(0.146)	(0.115)
Sex Education x Teacher	0.280**	0.592***	0.438***
	(0.114)	(0.163)	(0.125)
Observations	22,903	14,751	17,458
R-squared	0.003	0.002	0.003
Control Mean	0.493	0.737	0.566

Notes: This table reports average treatment effects on pregnancy incidence. All standard errors are robust and clustered at the school-level. All specifications include strata dummies and a vector of school-level baseline control variables. In OLS regressions we scale the dependent variable by a factor of one hundred for ease of intuitive interpretation of the coefficient as the direct incidence rate (e.g. .214 percent reduction in pregnancy). Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 8: Quantile Treatment Effects on Pregnancy

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	.1	.25	.5	.75	.9
Sex Education	-0.533***	-0.048	-0.230	-0.572***	-0.487**	-0.914***
	(0.187)	(0.337)	(0.329)	(0.196)	(0.212)	(0.204)
Sex Education x Teacher	0.694***	0.325	0.431	0.815***	0.785***	0.961***
	(0.206)	(0.452)	(0.473)	(0.240)	(0.298)	(0.189)
R-squared	0.321	0.061	0.161	0.271	0.247	0.225

Notes: This table reports quantile treatment effects on pregnancy. The dependent variable is a school-level continuous variable for the percentage of girls pregnant. This variable construction differs slightly from individual level pregnancy results with standard errors clustered at the school level. We present results for junior school girls using OLS and at various quintiles. Results from the OLS specification are similar to those presented in Table 6 for junior girls indicating that variable construction has minor effects as expected. All specifications include strata dummies and a vector of school-level baseline control variables. We scale the dependent variable by a factor of one hundred for ease of intuitive interpretation of the coefficient as the direct incidence rate. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 9: Effects on Pregnancy by Degree of Belief Updating

	(1) Girls	(2) Junior Girls	(3) Rural Girls	(4) Girls	(5) Junior Girls	(6) Rural Girls
Sex Education	0.208	0.508	0.113			
	(0.160)	(0.371)	(0.193)			
Sex Education x Teacher	-0.857**	-1.840*	-0.652*	-0.690**	-1.462*	-0.705**
	(0.357)	(0.939)	(0.380)	(0.305)	(0.751)	(0.330)
Risk Belief	-0.964**	-2.043**	-0.940**	-1.080**	-2.366*	-1.336**
	(0.373)	(1.016)	(0.415)	(0.527)	(1.349)	(0.640)
Sex Education x Teacher x Risk Belief	1.515***	2.895**	1.340**	1.611**	3.159**	1.726**
	(0.543)	(1.185)	(0.571)	(0.648)	(1.465)	(0.737)
Observations	22,602	14,377	17,282	15,357	9,654	11,807
R-squared	0.003	0.001	0.003	0.003	0.001	0.003

Notes: This table reports treatment effects on pregnancy incidence. All standard errors are robust and clustered at the school-level. All specifications include strata dummies. In OLS regressions we scale the dependent variable by a factor of one hundred for ease of intuitive interpretation of the coefficient as the direct incidence rate. We code 'Risk Belief' as knowing that older partners are riskiest right after the intervention. For the control group we code this as beliefs at endline since no beliefs are transferred and we observe near-identical beliefs in the control group at endline as baseline. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 10: Effects on Pregnancy by Degree of Implementation

	Attended			Comittment Card			Student Responses		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Girls	Junior Girls	Rural Girls	Girls	Junior Girls	Rural Girls	Girls	Junior Girls	Rural Girls
Sex Education	-0.244***	-0.492***	-0.383***	-0.230**	-0.473***	-0.354***	-0.221**	-0.478***	-0.379***
	(0.090)	(0.132)	(0.105)	(0.095)	(0.131)	(0.112)	(0.093)	(0.129)	(0.107)
Sex Education x Teacher Implementation	0.253*	0.574***	0.437***	0.295**	0.792**	0.538***	0.426**	1.049***	0.524***
	(0.130)	(0.215)	(0.147)	(0.138)	(0.308)	(0.140)	(0.191)	(0.384)	(0.196)
Observations	22,903	14,751	17,458	22,903	14,751	17,458	22,903	14,751	17,458
R-squared	0.003	0.002	0.003	0.003	0.002	0.003	0.003	0.002	0.003
Control Mean	0.493	0.737	0.566	0.493	0.737	0.566	0.493	0.737	0.566

Notes: This table reports reduced form IV estimates of treatment effects on pregnancy by the degree of implementation. Implementation is coded as "Attended" if teachers attended training, "Commitment Card" if teachers signed a card scheduling implementation and "Student Responses" if teachers handed back written students' responses to the regional ministry authority post implementation. The bounds on these measures of implementation range from 90 percent to 30 percent. *Teacher Implementation* corresponds to each of the degrees of implementation: *Attended*, *Commitment Card* and *Student Responses* and is instrumented for by random assignment to the teacher arm. All standard errors are robust and clustered at the school-level. All specifications are 2SLS and include strata dummies and a vector of school-level baseline control variables. We scale the dependent variable by a factor of one hundred for ease of intuitive interpretation of the coefficient as the direct incidence rate. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 11: HIV Beliefs by Degree of Implementation

	All		Attended		Comittment Card		Student Responses	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Post Intervention	12 Months Later	Post Intervention	12 Months Later	Post Intervention	12 Months Later	Post Intervention	12 Months Later
Sex Education	0.442***	0.088***	0.442***	0.088***	0.443***	0.088***	0.439***	0.087***
	(0.017)	(0.007)	(0.017)	(0.007)	(0.018)	(0.007)	(0.014)	(0.007)
Sex Education x Teacher	-0.268***	-0.046***	-0.225***	-0.039***	-0.178***	-0.029**	0.126***	0.006
	(0.040)	(0.010)	(0.045)	(0.011)	(0.058)	(0.014)	(0.039)	(0.025)
Observations	22,602	22,773	21,265	21,436	19,271	19,442	16,688	16,859
R-squared	0.532	0.282	0.555	0.285	0.601	0.306	0.822	0.359
Control Mean	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100

Notes: This table reports reduced treatment effects on beliefs on identifying older partners are most likely to have HIV by the degree of implementation. Implementation is coded as "Attended" if teachers attended training, "Commitment Card" if teachers signed a card scheduling implementation and "Student Responses" if teachers handed back written students' responses to the regional ministry authority post implementation. The bounds on these measures of implementation range from 90 percent to 30 percent. All standard errors are robust and clustered at the school-level. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 12: Self-Reported Sexual Behavior

	(1) More than 1 Sexual Partner	(2) Ever Had Sex	(3) Had Sex with Older Partner	(4) Receiving Gifts	(5) Self Efficacy	(6) Condom Use
Panel A: All Girls						
Sex Education	-1.423	-0.918*	-0.510	-0.531	1.824	-0.965
	(0.943)	(0.551)	(0.321)	(0.532)	(1.169)	(0.617)
Sex Education x Teacher	1.493	1.457**	0.560	1.042	-0.804	0.733
	(1.071)	(0.698)	(0.403)	(0.676)	(1.207)	(0.752)
Observations	21,148	21,114	20,991	20,944	21,030	20,981
R-squared	0.018	0.013	0.003	0.012	0.039	0.017
Control Mean	10.99	7.362	2.976	6.703	80.48	6.491
Panel B: Junior Girls						
Sex Education	-2.215**	-1.517*	-0.275	-0.590	3.277***	-1.253
	(0.988)	(0.800)	(0.471)	(0.771)	(0.948)	(0.802)
Sex Education x Teacher	2.304**	1.636*	0.310	1.369*	-2.657**	0.929
	(1.078)	(0.841)	(0.527)	(0.754)	(1.074)	(0.888)
Observations	13,575	13,545	13,438	13,399	13,456	13,424
R-squared	0.023	0.008	0.003	0.008	0.021	0.012
Control Mean	11.93	9.552	3.250	8.184	83.61	8.178
Panel C: Rural Girls						
Sex Education	-1.841	-1.149	-0.823**	-0.543	1.788	-1.144
	(1.219)	(0.702)	(0.348)	(0.716)	(1.471)	(0.804)
Sex Education x Teacher	1.662	1.633*	1.008**	1.199	-0.746	0.967
	(1.380)	(0.871)	(0.458)	(0.900)	(1.592)	(0.950)
Observations	16,164	16,142	16,040	16,012	16,083	16,040
R-squared	0.015	0.014	0.003	0.011	0.032	0.015
Control Mean	12.46	7.902	3.264	7.269	78.31	7.293

Notes: This table reports treatment effects on self-reported sexual behavior and attitudes for girls after 12 months. All standard errors are robust and clustered at the school-level. All specifications include strata dummies and a vector of school-level baseline control variables. We scale the dependent variable by a factor of one hundred for ease of intuitive interpretation of the coefficient as the direct incidence rate. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 13: Treatment Effects on Boys' Absenteeism

	OLS			Probit		
	(1)	(2)	(3)	(4)	(5)	(6)
	Boys	Junior Boys	Rural Boys	Boys	Junior Boys	Rural Boys
Sex Education	-0.416	-0.759	-0.800	-0.027	-0.049	-0.052
	(0.766)	(0.969)	(0.899)	(0.051)	(0.065)	(0.060)
Sex Education x Teacher	0.880	2.038*	0.939	0.059	0.130*	0.064
	(0.732)	(1.085)	(0.803)	(0.048)	(0.070)	(0.054)
Observations	22,288	13,993	17,159	22,288	13,993	17,159
R-squared	0.004	0.007	0.003	0.011	0.003	0.006
Control Mean	8.077	8.410	8.284	8.077	8.410	8.284

Notes: This table reports treatment effects on boys absenteeism from school. Boys absenteeism is calculated using the 'roll-call' method where enumerators read off a list of all students names from baseline to see who was missing from school at endline. All standard errors are robust and clustered at the school-level. All specifications include strata dummies and a vector of school-level baseline control variables. We scale the dependent variable by a factor of one hundred for ease of intuitive interpretation of the coefficient as the direct incidence rate. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 14: Treatment Effects on Pregnancy by Teacher Proximity

	All Girls					Junior Girls				
	(1)	(3)	(2)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Teacher	-0.502	-0.332	0.159	0.143	0.051	-1.804***	-1.001*	-1.682	-1.652	0.099
	(0.392)	(0.452)	(0.551)	(0.554)	(0.117)	(0.429)	(0.525)	(2.089)	(2.143)	(0.201)
Experience	-0.032**	-0.014		-0.001		-0.100***	-0.046*		-0.004	
	(0.016)	(0.023)		(0.011)		(0.015)	(0.024)		(0.011)	
Teacher x Experience	0.027	0.019				0.099***	0.053*			
	(0.017)	(0.020)				(0.024)	(0.029)			
Age		-0.149*	-0.163**	-0.163**			-0.432***	-0.658***	-0.643***	
		(0.083)	(0.064)	(0.076)			(0.116)	(0.199)	(0.218)	
Teacher x Age			-0.021	-0.018				0.294	0.288	
			(0.096)	(0.096)				(0.377)	(0.389)	
Gender					0.094					0.100
					(0.138)					(0.209)
Observations	14,850	14,850	14,913	14,850	14,870	9,304	9,304	9,304	9,304	9,304
R-squared	0.004	0.004	0.004	0.004	0.004	0.003	0.004	0.004	0.004	0.002

Notes: This table reports treatment effects on pregnancy by teacher proximity to students on a bundle of characteristics including age (in years), gender and teaching experience (in years). Since teacher characteristics might be endogenous we interpret these results as descriptive and suggestive. We report effects at junior school. Standard errors are robust and clustered at the school-level. All specifications include strata dummies and a vector of school-level baseline control variables. We scale the dependent variable by a factor of one hundred for ease of intuitive interpretation of the coefficient as the direct percentage. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 15: Survey Reporting Effects by Messenger

	Knowledge	Sensitive Behaviors		
	(1)	(2)	(3)	(4)
	HIV risk: age 40-49	More than 1 sex partner	Ever sex with older partner	Ever pregnant
Panel A: All				
Age	-0.111	-0.208	-0.119**	-0.189**
	(0.087)	(0.116)	(0.046)	(0.058)
Female	-0.005	-1.322	0.023	-0.326
	(0.480)	(0.885)	(0.279)	(0.403)
Observations	39,249	39,494	39,213	39,136
Control Mean	13.74	17.32	4.909	5.261
Panel B: Junior				
Age	-0.169	-0.206	-0.128*	-0.207**
	(0.108)	(0.133)	(0.062)	(0.072)
Female	-0.569	-0.649	0.146	-0.213
	(0.569)	(1.146)	(0.351)	(0.549)
Observations	24,490	24,621	24,364	24,335
Control Mean	13.06	18.87	5.664	5.424
P-Value: Primary vs. Junior	0.179	0.576	0.788	0.830

Notes: This table reports effects of messenger characteristics on non-sensitive versus sensitive survey reporting behavior. Surveyors were randomly assigned to schools providing an exogenous shock of messenger characteristics, such as age and gender. Standard errors are robust and clustered at the school-level. All specifications include strata dummies and a vector of school-level baseline control variables. We scale the dependent variable by a factor of one hundred for ease of intuitive interpretation of the coefficient as the direct percentage. Significance levels are denoted as follows: *** p<0.01, ** p<0.05, * p<0.1.

Appendix A: Additional Figures

Figure 24: Map of Schools by Treatment Group

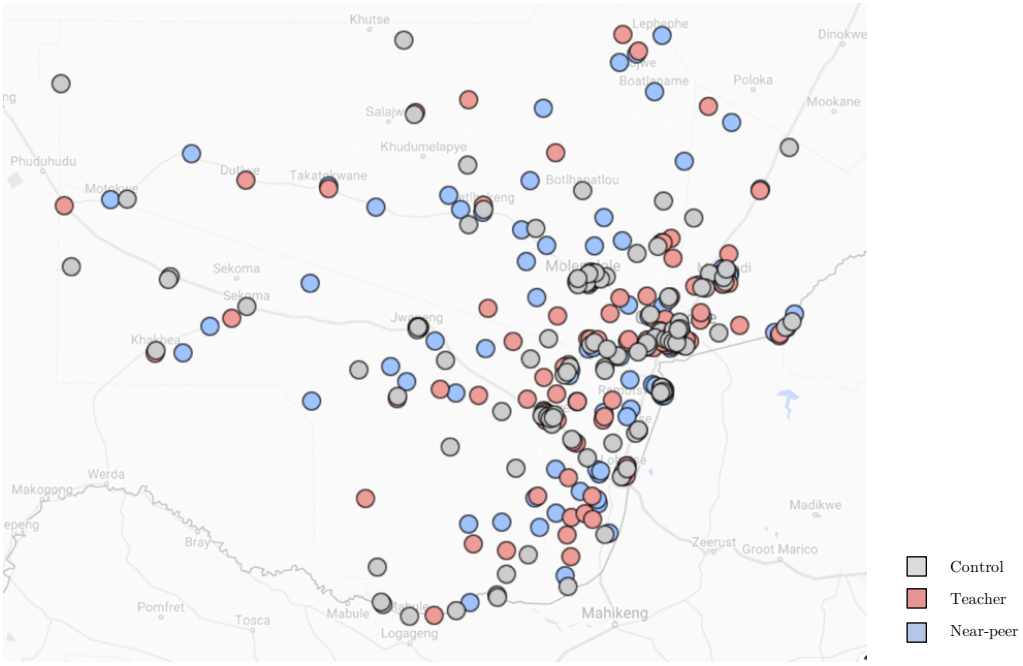
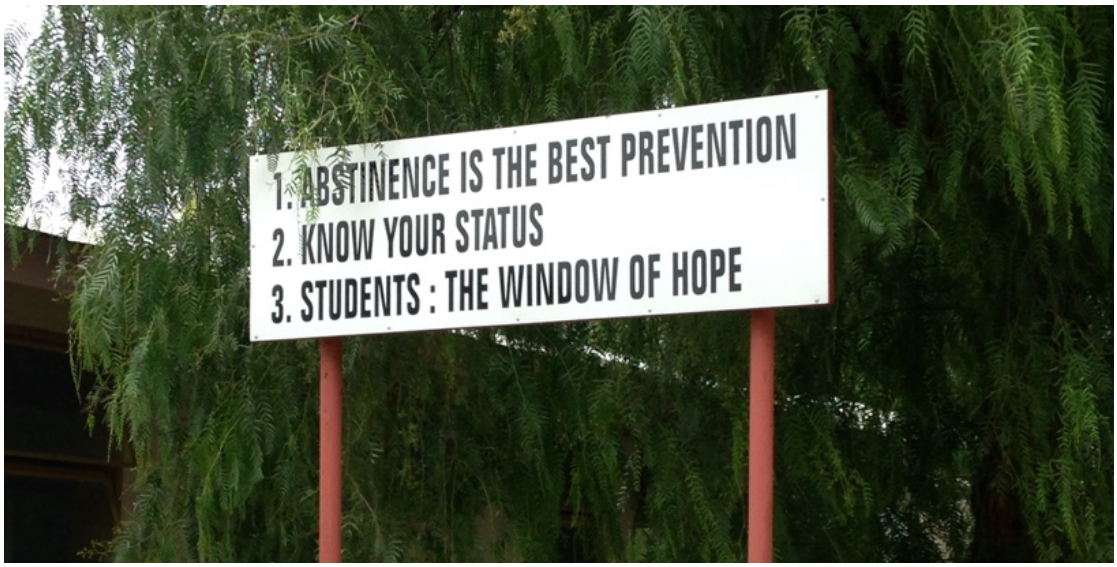


Figure 25: Typical Sign at a School



Appendix B: At-Home Verification of Pregnancy

We conducted household visits to verify pregnancy. Once at the household, enumerators observed whether the girl had a visibly large stomach (a 'tummy') or a baby. This was recorded. The enumerator would then give the girl a survey to fill out in written form followed by an oral interview. Finally, the enumerator would interview the girl's caregiver. In total this produced five measures of verification for pregnancy that had been identified at the school by classmates and teachers. Of all pregnancies identified at the school roughly 90 percent had at least one measure of the five indicate the girl was pregnant. This gives us confidence that our pregnancy measure is accurate and robust.

We compute the Cronbach's alpha to assess reliability of our various measures of pregnancy. We code tummy and baby as a joint observational measure. Across all measures, including written, oral and observational by the caregiver, girl and enumerator, we find a Cronbach's alpha of .65, which falls in the .60 to .70 range considered reliable in the measurement literature.

Table 16: Reliability of At-Home Measures

Item	Sign	Item-test Correlation	Alpha
Observational	+	0.619	0.638
Caregiver	+	0.657	0.598
Written - Girl	+	0.73	0.598
Oral - Girl	+	0.79	0.491
Chronbach's alpha			0.654

Notes: We combine tummy and baby observational measures. While the 'tummy test' has not been validated in the literature, it provides a potentially objective measure of pregnancy to complement observations of babies. However, since there are multiple interpretations of how tummy and baby measures move together, we combine them into an observational measure. For example, tummies might indicate early pregnancy whereas babies indicate post-pregnancy. Given the ambiguities of how to interpret each measure independently we combine them into an observational measure which we corroborate against our three additional measures: interviews with girls and caregivers and survey responses from girls.

MEASURING HUMAN CAPITAL

NOAM ANGRIST, SIMEON DJANKOV, PINELOPI K. GOLDBERG,
AND HARRY A. PATRINOS*

Students around the world are going to school but learning is limited – an emerging gap in human capital formation. To understand this gap, this paper introduces a new data set measuring learning in 164 countries. The data cover 98 percent of the world's population from 2000 to 2017. The data set will be publicly available and updated at regular intervals by the World Bank. The paper presents several stylized facts in a first application of the data: (a) although enrollment has increased worldwide, learning has stagnated; (b) girls outperform boys on learning – a positive gender gap – in contrast to a negative gender gap observed for schooling; (c) learning is associated with growth on a global scale; (d) associations with growth are heterogenous; and (e) human capital accounts for between a fifth to half of cross-country income differences – a middle ground in the recent development accounting literature. These stylized facts demonstrate the potential of the data to reveal new insights into the relationship between human capital and economic development.

(JEL I20, O40, O15, H40, H52, J24, P50)

*Angrist: University of Oxford, and World Bank, 1818 H Street NW, Washington, DC 20433 (e-mail: noam.angrist@bsg.ox.ac.uk); Djankov: London School of Economics, and World Bank, 1818 H Street NW, Washington, DC 20433 (e-mail: sdjankov@worldbank.org); Goldberg: Yale University, and World Bank, 1818 H Street NW, Washington, DC 20433 (e-mail: pgoldberg@worldbank.org); Patrinos: World Bank 1818 H Street NW, Washington, DC 20433 (e-mail: hpatrinos@worldbank.org).
We are grateful to Syedah Aroob Iqbal and Husein Abdul-Hamid for research support. This work builds on co-authored work with Nadir Altinok. Particular thanks to Aart Kraay for detailed comments and contributions to the methodology. Valuable comments were provided by Eva L. Baker, Felipe Barrera-Osorio, Eduardo Cascallar, Paul Collier, Stefan Dercon, Deon Filmer, Roberta Gatti, Rachel Glennerster, Daniel Koretz, Silvia Montoya, George Psacharopoulos, Simon Quinn, Heiner Rindermann, Halsey Rogers, Jaime Saavedra, Shwetlena Sabarwal, Justin Sandefur, and Eduardo Velez. This paper benefited from seminars held at the World Bank, World Congress of Cliometrics, American Economic Association, IDB, Oxford, USAID and FHI360. A great number of individuals and organizations supplied us with data. A special thanks to Rebecca Rhodes for access to the microdata for many EGRA learning assessments. The views expressed here are those of the authors and should not be attributed to the World Bank.

I. Introduction

The notion of human capital – resources imbedded in people – was alluded to as early as 1776 by Adam Smith and formalized two centuries later by Becker (1962). Ever since, the literature has explored the role of human capital in economic development. For decades, this literature proxied human capital with measures of schooling. Examples include Nelson and Phelps (1966), Mincer (1984), Lucas (1988), Barro (1991), Mankiw, Romer, and Weil (1992) and Gennaioli, La Porta, Lopez-de-Silanes, and Shleifer (2013).

However, proxying human capital with schooling assumes that being in school translates into learning. Evidence suggests that this is often not the case (Pritchett 2013). A recent analysis reveals that six out of ten adolescents worldwide cannot meet basic proficiency levels in math and reading (UNESCO 2017). The gap between schooling and learning is particularly acute in developing countries. In Kenya, Tanzania, and Uganda three-quarters of grade 3 students cannot read a basic sentence such as "the name of the dog is Puppy." In rural India, half of students in grade 3 cannot solve a two-digit subtraction problem such as 46 minus 17 (World Bank, World Development Report 2018).

These stylized facts demonstrate a gap in human capital formation: students are in school but learning is limited. Figure I demonstrates the contrast between rising enrollment and stagnating learning.[2] Closing this gap presents a significant margin to drive economic development. Several papers have argued that when human capital is measured by schooling it fails to deliver the returns predicted by growth models (Krueger and Lindahl 2001; Pritchett 2006; Hanushek and Woessmann 2008). However, when measured by learning, human capital is more strongly

[2] The average patterns observed for learning in Figure I might be driven by a few factors, including limited learning for the same cohort of students or selection as marginal students enroll in school. Later in the paper, we explore more detailed trends and drivers of learning.

associated with growth (Hanushek and Woessmann 2012a). This result is robust to improved measures of schooling such as those produced by Cohen and Soto (2007). Thus, when human capital is measured by a central output of schooling – learning – predictions from economic growth models prove more robust.

However, to date much of the analysis of human capital when measured by learning has focused on advanced economies. This is due to the absence of comparable measures of learning in developing countries. This excludes a significant portion of the global distribution, in particular countries with the most potential to gain from human capital accumulation.

In this paper, we bridge this gap. We introduce a database of globally comparable learning outcomes for 164 countries covering 98 percent of the global population from 2000 to 2017.[3] The database will be updated at regular intervals and made available for public use. We hope this database will enable tracking of human capital formation, a critical ingredient for development. A large-scale effort to do this is the World Bank's new Human Capital Index. This database is a key ingredient in the index.[4] This database will enable deeper understanding of factors correlated with human capital formation and with potential causal links to economic development.

We present stylized facts in a first application of the database to demonstrate its utility. We show that learning has stagnated while schooling has increased. We further find that gender gaps in learning are large and positive with girls outperforming boys. This points in the opposite direction of the gender gap for schooling which is negative. In an illustrative growth exercise, we find that human capital measured by learning is associated with growth. This finding reinforces recent literature on a global scale for the first time. The expanded coverage of our

[3] There are two administrative regions: Hong Kong and Macao, which we refer to as countries for simplicity.
[4] The World Bank Human Capital Index includes additional measures of human capital, such as combining measures of school enrollment with learning, as well as including measures of survival and health (Kraay 2019).

database makes possible heterogeneity analysis of previously underexplored dimensions, enabling new insight. We find significant heterogeneity in growth patterns by initial and current income status.

We also contribute to a development accounting literature which accounts for the role of human capital in cross-country income differences. To date the literature has relied on years of schooling data, and in the absence of global learning data, explored an array of techniques to infer school quality. This has resulted in a wide range of accounting estimates. Rather than infer, we provide a direct measure of school quality and human capital. We find that human capital accounts for between a fifth to half of cross-country income differences – a middle ground in a literature that provides estimates ranging from nearly all (Jones 2014) to potentially zero (Caselli and Ciccone 2018). These results are consistent with more comprehensive formulations of human capital, such as models capturing the role of educated entrepreneurs (Gennaioli, La Porta, Lopez-de-Silanes, and Shleifer 2013).

We compare our measure of human capital to alternatives. We find that our measure is highly correlated with other learning measures while tripling coverage. Moreover, our measure of human capital has a strong association with growth while schooling and the measure of human capital in the Penn World Tables (Feenstra, Inklaar, and Timmer 2015) are more limited. This indicates that years of schooling and the Penn World Tables measures might understate the role of human capital. However, their use remains standard practice in part since these data have the broadest coverage. By constructing learning data across 164 countries we fill a key gap: broad coverage over nearly two decades and a measure of human capital with strong links to economic development.

The rest of this paper is organized as follows. Section II outlines the methodology. Section III describes the database. Section IV showcases descriptive statistics. Section V compares our data to various alternatives. Section VI presents several stylized facts and Section VII concludes.

II. METHODOLOGY

We leverage the growth of international assessments to construct globally comparable learning outcomes. These tests are derived from assessments conducted in the United States since the 1960s such as the SAT and the National Assessment of Educational Progress (NAEP). The tests are psychometrically designed, standardized assessments of cognitive skills. Since the 1990s, international assessments have been conducted by organizations such as the OECD. Two high profile examples are PISA and TIMSS which covered 71 and 65 countries in 2015. These assessments enable credible global comparison of learning across countries and over time.

To date, a series of international assessments have been included in analysis of human capital and development (Hanushek and Kimko 2000; Barro and Lee 2001; Hanushek and Woessmann 2012a). These analyses focus on OECD countries and cover few developing countries. For example, Hanushek and Woessmann (2012a) include 50 countries, around half of which are from the OECD. This limits the distribution of countries represented and has implications for our understanding of the link between human capital and economic development.

We include developing countries by constructing a learning 'exchange rate' between international assessments and their regional counterparts. Regional assessments cover much of Sub-Saharan Africa and Latin America. Thus, through construction of a cross-test exchange rate between international and regional assessments we quantify the difference between them, adjust for this difference, and then place learning outcomes from regional assessments on a global scale.

Our approach builds on seminal work by Hanushek and Kimko (2000) and Barro and Lee (2001). Barro and Lee (2001) use regression to obtain different constants between various tests, allowing for test differences. Hanushek and Kimko (2000) create comparable scores over time using the United States as an anchor, since the

United States participates in consistent national assessments as well as various international assessments. Hanushek and Woessmann (2012b), Barro and Lee (2015), and Altinok, Angrist and Patrinos (2018) use a related approach across countries to including a subset of developing countries.

II.A. Construction of a Learning Exchange Rate

The central intuition behind the construction of globally comparable learning outcomes is the production of a learning 'exchange rate' between international and regional assessments. This exchange rate can be produced for countries that participate in a given pair of assessments and captures the difference in difficulty between the two assessments. This exchange rate can then be used to place scores for countries that only participate in regional assessments on the international scale. This enables construction of globally comparable learning outcomes.

We present an illustrative example. Lesotho has data from a regional assessment in 2007. We want to place Lesotho's regional score on an international scale. To do this, we can produce an exchange rate between the regional assessment and an international assessment. We produce this exchange rate first by comparing scores for countries that participate in both assessments in the same subject and adjacent years. In this case, Botswana participated in the same regional assessment as Lesotho as well as an international assessment in 2007. Since in Botswana the same population of students took two different tests at the same time, we can derive an exchange rate. We can then apply this exchange rate to Lesotho's score on the regional assessment to place it on the international scale.

In this example, Botswana has a mean score of 364 in math in 2007 on an international assessment, TIMSS. Botswana also participated in a regional assessment, SACMEQ, in the same year and subject and scored 520. The exchange rate between SACMEQ and TIMSS in math in 2007 is therefore 0.7. Lesotho's

regional score was 477 in math in 2007. Applying the exchange rate of 0.7 to Lesotho's regional score of 477 would produce a score of 334 on the international scale.

More formally, suppose that a population of students takes two different assessments *X* and *Y*. We assume that any differences in the score distributions on *X* and *Y* can be attributed to the assessments with constant group ability. In its most straightforward form, we convert a score from regional test *X* to an international test *Y* as follows:

$$y = x * e \tag{1}$$

where e is an exchange rate; y is a score on international test Y and x is a score on regional test X.

The success of this approach hinges on three key assumptions. First, linked tests must capture the same underlying population. This assumption is satisfied by using sample-based assessments representative at the national level where a country participated in both a regional and international assessment. This ensures that the underlying population tested is the same on average and we capture differences between tests.

Second, tests should measure similar proficiencies. To this end, we link within subjects (math, reading and science) and schooling levels (primary and secondary) to ensure overlap.

Third, the exchange rate should capture differences between tests rather than country-specific effects. This assumption is most likely to hold the larger the number of countries which participate in a given pair of tests being linked. To ensure this last assumption holds, we fix the exchange rate over the entire interval. This increases the sample size used to produce the exchange rate, increasing the likelihood that we capture test-specific rather than country-specific differences. In

fixing the exchange rate, we assume that the relationship between tests stays constant across rounds. This assumption is reasonable and is true by design since the mid-1990s when assessments started to use a standardized approach and to link testing rounds with overlapping test items. A related advantage of fixing the exchange rate is that it guarantees that any changes in test scores over this interval are due to realized progress in learning rather than changing exchange rates.

Of note, every update of the database increases the number of countries participating in a given pair of assessments. Thus, each update both expands coverage as well as enhances the reliability of all estimates by enabling construction of a more robust exchange rate.

Below we capture a level of precision needed to satisfy the above assumptions. We produce a fixed exchange e_{sl} within subjects and schooling levels (primary and secondary) from test X to test Y:

(2)
$$e_{sl} = \frac{1}{r}\sum_{r=1}^{r} \frac{\mu(y_{rsl})}{\mu(x_{rsl})} \quad \text{where}$$

$$\mu(x_{rsl}) = \frac{1}{n_{rsl}} \sum_{i \in Y_{rsl} \cap X_{rsl}} x_{irsl}$$

$$\mu(y_{rsl}) = \frac{1}{n_{rsl}} \sum_{i \in Y_{rsl} \cap X_{rsl}} y_{irsl}$$

where i is a country in the set of n_{rsl} countries that participate in both tests X and Y in a given testing round r, subject s, and schooling level l such that $Y_{rsl} \cap X_{rsl}$. The upper value of r is the number of testing rounds over the period of the fixed exchange rate.

Mean scores, $\mu(x)$ and $\mu(y)$, are calculated across countries for a given round r, subject s and schooling level l. We consider tests to be in the same round if they are five years apart and optimize to have the rounds as tight as possible. Most often this

translates to within one to two years. In some cases, this extends to three to five years apart. In a few exceptions, we average adjacent years across one another. This minimizes the likelihood that test differences are a function of time, proficiency, schooling level, or data availability and are an accurate reflection of test difficulty.

We apply this exchange rate to a country *j* that participates in test X but not test Y to produce a comparable score (referred to as a Harmonized Learning Outcome (HLO) in the database):

$$y_{jtsl} = x_{jtsl} * e_{sl} \tag{3}$$

where *t* is the official year of test X. Means can be calculated for various disaggregated groups, for example, by gender. We can also aggregate scores across levels and subjects by averaging across subjects and then across levels, weighting each equally.

In line with this level of precision, we explore another demonstrative example. We want to know Brazil's primary math score in 2013. While Brazil does not participate in an international assessment at the primary level, it does participate in a regional assessment, LLECE. To place this score on the international scale, we produce an exchange rate for countries that participate in this regional test and an international counterpart at the primary level in math, TIMSS.

In the latest testing round, these countries were Chile and Honduras. Chile's regional math score was 581 and Honduras's score was 480. On the international assessment, they score 462 and 397. Thus, the average regional score is 530 and the average international score is 429, and the resulting exchange rate in the latest testing round is 0.81. We next examine the prior testing round to see if the exchange rate is stable across rounds and to produce a fixed exchange rate. In this round, the countries participating in the regional and international assessment were Colombia and El Salvador. Colombia scored a 493 and El Salvador scored a 472 on the

regional assessment. On the international assessment they scored 356 and 330, respectively. Thus, the exchange rate was 0.71. The two exchange rates of 0.71and 0.81 are relatively stable across a 15-year period and variation across participating countries. Averaging across both rounds, we produce a fixed exchange rate of 0.76. Applying this exchange rate to Brazil's regional primary math score of 520 in 2013, an internationally comparable score is 395 for math in 2013.

A potential challenge in applying this methodology is that it is in principle sensitive to the scale of the test. For example, given score scales have no absolute zero, in theory we can add 300 points to the mean of each test and preserve the properties of the scale, but will alter the exchange rates. We address this potential issue by having strict inclusion criteria for the underlying tests: they must have a similar scale with a mean of 500 and standard deviation of 100. Exchange rates are derived using the same scale and applied on the same scale. Thus, while in theory changing score scales might bias results, *by design* this is never the case. This ensures we capture differences in test difficulty rather than in scaling. We run a series of robustness tests in the appendix to ensure scaling does not bias scores.[5]

Another important feature of the methodology is accounting for the changing distribution of countries in each testing round. Given that the underlying tests used are scaled to a mean of 500 and standard deviation of 100, this might produce variation in scores as a result of changes in the underlying distribution of countries rather than true learning progress. However, in practice this is not an issue, since we use tests which already account for this variation. For example, each new PISA round is made comparable to prior rounds to account for the changing distribution of countries by the testing agency prior to publishing results. A detailed review of the procedure used to make data comparable over time can be found in the PISA

[5]Additional methods for linking test scores in the psychometric literature include linear and equipercentile linking as well as Item-Response Theory (IRT) (Kolen and Brennan 2014). We demonstrate that our approach results in scores with over a .98 correlation with all of these approaches. Figures A1 and A2 in the Appendix shows these robustness tests.

technical report (OECD 2015). This approach is analogous across underlying tests used for this database. We use learning data *ex-post* this transformation has already been conducted by the underlying international and regional assessments. Hanushek and Woessmann (2012a) introduce an additional layer of assurance to ensure changing distributions of countries participating in assessments do not drive results. The authors express performance in terms of standard deviations. They then project the standard deviations of a relatively homogenous and stable group of OECD countries and transform these standard deviations into scores using a standardized scale for an international assessment. However, as the authors acknowledge, this approach does not apply for countries far from the OECD distribution of scores, limiting its relevance when including developing countries on a global scale.

Our methodology enables us to compare learning outcomes on a global scale. On this scale, 625 represents advanced attainment and 300 represents minimum attainment. This interpretation is derived by taking establish benchmarks already used on international and regional assessments. For the high-performance benchmark on the upper end of the distribution, we use the TIMSS benchmark of 625. For the low-performance benchmark on the lower end of the distribution, we use 300, which is the equivalent on the HLO scale of the minimum benchmarks on regional assessments such as LLECE and PASEC. This approach enables us to capture performance across the distribution and accounts for floor and ceiling effects that would be introduced by taking either international or regional benchmarks on both ends of the distribution. Additional details of the methodology are included in the Appendix, including information on countries that participate in multiple assessments to enable derivation of exchange rates in Table A4.

II.B. Measures of Uncertainty

We include measures of uncertainty to quantify the degree of confidence around our estimates. We capture two sources of uncertainty: scores on the original test and uncertainty in the calculation of the exchange rate across tests. We calculate the variance by bootstrapping. We consider each average country score on a given subject, test, and schooling level as a random variable with a mean – the score itself – and a standard deviation which captures the sampling variation across students. This distribution of scores is asymptotically normal by virtue of the central limit theorem. We take 1,000 draws from the distribution of subject-level average test scores for each testing regime. We do this as a computational shortcut, rather than bootstrapping subsamples of students from each test. We create exchange rates and scores from each bootstrapped sample. We take the 2.5^{th} and 97.5^{th} percentiles of the distribution and use this to construct lower and upper bounds of uncertainty.

We find small uncertainty intervals overall, with an average of 3.5 points and ranging from 1 to 11 points. This is consistent with original standard errors from each respective testing regime. We find larger uncertainty for our estimates relative to original scores when testing regimes have fewer countries participating in a given pair of tests. Figure A3 decomposes standard errors due to within-test sampling variation as well as variance in the exchange rates. This figure shows that for tests where there is no need to produce an exchange rate, or many pair-wise countries which we can use to produce this exchange rate, the final standard errors remain similar to standard errors from the original test. For tests with fewer pair-wise countries, the exchange rate has more uncertainty, such as PASEC, where the average standard error increases from 5.3 on the original test to 7.5 for the HLO. By quantifying the degree of uncertainty, we can more reliably bound our estimates.

II.C. Limitations

A potential limitation is the representativeness of the data of the total stock of cognitive skills in a given country. While the tests used are nationally representative, they are conducted at the school. To this end, learning data might be affected by enrollment patterns, and we advise users of the data to analyze learning outcomes alongside enrollment trends. For example, as marginal students enter the schooling system, average test scores might be driven by selection rather than true learning progress. While this is a potential concern, it is mitigated for a few reasons. First, primary enrollment rates are relatively high, reaching 90 percent on average, and above 75 percent even in the furthest behind regions, such as Sub-Saharan Africa. Second, the direction of the bias is likely to yield a conservative upper bound of learning in a given country. If all students enrolled, the average test score would be even lower, since the marginal students would pull the average down. Since most countries at the bottom of the distribution of learning are also those with relatively lower enrollments, it is unlikely this will alter substantive conclusions – the lowest preforming countries will be revealed to be even lower performing. In addition, data at the primary level should be largely unaffected, since at this level students are being taught basic skills, such as reading "the name of the dog is Puppy." Thus, even if marginal students enter the system, these students should still be expected to attain basic skills by the time they are tested in later primary school grades. Of note, in future work, we aim to include household-based learning data to sign and quantify the degree of selection present in school-based testing. However, current household-based data is limited and not yet comparable across a significant number of countries.

A second limitation regards data availability. While this is the largest learning outcomes database to date, data are still sparse for some countries. This introduces bias if data availability is correlated with education quality or progress. For example, if countries that perform worse have data only in later years (because they were later to introduce assessments), their average score will be likely biased

upwards, as the test scores will reflect more recent testing, not stronger performance. Since we provide year-by-year scores this can be accounted for.

Relatedly, when averaging data across subjects, levels and over time, there is a possibility that averages reflect the availability of data rather than learning gains. For example, let's examine a case where a country has a score of 500 in 2000 in math and jumps to 550 in 2005. If this country added reading in 2005 and scored 450, the average score across subjects in 2005 would be 500, reflecting no learning progress since average scores would be 500 in both years. However, an apples-to-apples comparison in math shows learning grains from 500 to 550. To address this issue, we construct disaggregated measures by subject and schooling levels as well as aggregated ones. This enables analyses at each level considering the trade-offs.

A point of emphasis is that while learning measures human capital better than prior proxies, such as enrollment, learning does not capture the concept of human capital in its totality. Moreover, assessments do not capture only cognitive skills. For example, recent evidence suggests test scores pick up as differential effort as well as cognitive ability (Gneezy, List, Livingston, Sadoff, Qin, and Xu 2017). We refer to learning outcomes as human capital in this paper with these caveats in mind.

III. DATA

This section describes the achievement tests we use to construct our database. The first set of assessments consists of international standardized achievement tests; the second are regional standardized achievement tests; and the third is the Early Grade Reading Assessment (EGRA). Each test covers between 10 and 72 countries. By combining these assessments and making them comparable we include 164 countries covering 98 percent of the global population. Appendix A includes a detailed description of all data included in the database.

The database was produced through a combination of methodological innovation as well as a large-scale effort by the World Bank to identify, collect and collate student assessment data worldwide across eight testing regimes. We include data from the Early Grade Reading Assessment for the first time. This adds 48 countries to the database with at least one data point in the past 10 years, including large developing economies such as Bangladesh, Nigeria and Pakistan.

The final database includes mean scores for all 164 countries from 2000 to 2017. We do not extend the time series prior to 2000 since data quality is low and since this does not significantly affect country coverage, with an addition of just two territories: Zanzibar and Puerto Rico. The codebook is summarized in Table I.

Scores are disaggregated by schooling level (primary and secondary), subject (reading, math and science) and gender (male and female). We include standard errors for each measure to quantify uncertainty. We include year-by-year data, an indicator for whether the data are nationally representative, as well as the source test from which the scores were derived. Finally, we include the exchange rates used to convert an original score to a globally comparable score. The database will be made publicly available by the World Bank and will be updated regularly.

Of note: while this is the largest learning database to date, data are still sparse. This introduces a few potential biases when aggregating data to maximize country coverage as described in the limitations section of the methodology. The inclusion of year-by-year and disaggregated data provides transparency and granularity to enable informed analysis.

IV. Descriptive Statistics

Several statistics demonstrate the coverage and detail of the database. Table II presents coverage for country-year observations by region. The database includes 2083 observations across all countries from 2000-2017. Disaggregation by gender

is available for nearly all the data, with 2,020 country-year observations. Most data come from math scores, with 744 country-year observations, followed by reading scores, with 692 and lastly by science scores, with 647. Almost two-thirds of observations are secondary school scores. Latin America and the Caribbean and Sub-Saharan Africa make up nearly a quarter of all available data. This provides the largest representation of developing countries to date in a learning database – the countries that have the most potential to gain from human capital accumulation.

Table III presents country-subject-level observations by year. The data are spread over time, slightly weighted towards recent years since countries are increasingly participating in assessments. A related feature of the data is a large influx of data in particular testing years. This is more prevalent for developing regions which participate in sporadic assessment.

Figure II presents mean human capital scores for 164 countries from 2000-2017. Table A3 in the appendix provides a corresponding table for all countries and includes gender disaggregation. Figure II makes the global coverage of the database immediately apparent with typically excluded regions from international tests such as PISA and TIMSS included in our database. This includes the vast majority of Sub-Saharan Africa, Latin America and the Caribbean, and South Asia – economies with significant potential to harness human capital for economic development. A few trends emerge when comparing our measure of human capital across regions. Advanced economies far outpace developing economies in terms of human capital; Sub-Saharan African lags behind all regions; within Sub-Saharan Africa, a few countries such as Kenya and Tanzania lead, on par with many countries in Latin America; within Latin America, a few countries such as Chile are on par with European counterparts; the Middle East performs similarly or worse than Latin America; many Asian countries outperform North American and European counterparts, while a few South Asian countries such as India and Bangladesh perform on par with many Sub-Saharan African countries.

Figure III summarizes human capital for selected countries for the last decade. We observe a few case studies consistent with the trends described above. The Russian Federation outperforms the United States – a surprising finding. Chile outperforms Eastern European countries such as Georgia. Saudi Arabia places near the bottom outperforming only African countries. The gap between Morocco and Singapore is nearly double. Singapore and Finland have low variation due to a potential plateau on the upper end of performance. Rwanda has low variation due to limited data. Russia has high variation due to improving human capital, whereas South Africa has high variation due to declining human capital.

Figure IV plots human capital for each country by the log of their GDP per capita. This graph illuminates cases where countries have managed to improve human capital despite a lack of resources, as well as cases where countries have resources to invest in to date unrealized human capital accumulation. Former or current centrally planned economies display better human capital than their income would suggest, such as Singapore, Poland, Bulgaria, Cuba and Vietnam. Countries in the Middle East and Africa reach lower human capital levels than predicted by income, such as Qatar, Kuwait, United Arab Emirates, South Africa, Nigeria and Ghana. We also highlight large developing countries: India, China, Mexico, and Brazil. China outperforms its counterparts, Mexico and Brazil perform slightly below where their income would predict, and India and South Africa trail far behind.

Figure V contrast years of schooling with learning for the most recent year in which data is available. A few trends emerge from this graph. First, the exponential shape of the line of best fit indicates that there is a tipping point where countries realize more learning after about ten years of schooling. Second, there is high variance in learning conditional on years of schooling, with some countries reaching 8 years of expected schooling yet only just over 300 on learning while others with the same years of schooling reach over 450 on learning. Moreover, this graph reveals that many developing countries have successfully achieved high

schooling, but have not yet realized high rates of learning. A few notable examples with high schooling but low learning include India, South Africa and Ghana. South Africa, for example, has 9.3 expected years of schooling, yet an HLO score of just 343. India has 10.2 expected years of schooling, yet an HLO score of just 355.

V. Comparison to Alternative Measures

We compare our measure of human capital to a few alternatives. First, we compare our measure to related learning measures of human capital. We find high levels of consistency with other learning measures. Next, we compare our measure to schooling measures of human capital. While coverage is similar, learning is consistently associated with growth, whereas the association with schooling is more limited. Finally, we compare our measure of human capital to a measure included in the recent version of the Penn World Tables. We find that our measure is consistently associated with growth, whereas the Penn World Tables measure, like schooling, is limited. In summary, the learning measure of human capital in this database is highly correlated with other learning measures, while tripling coverage to include 164 countries on a global scale for the first time. Moreover, our measure of human capital has a strong association with growth in contrast with schooling and Penn World Tables measures.

V.A. Comparison to Learning Measures of Human Capital

Our database complements alternative learning measures of human capital. We correlate our measures with a few notable alternatives. Figure VI below shows direct comparisons to learning data used in growth regressions by Hanushek and Kimko (2000) and Hanushek and Woessmann (2012a). We find correlations of 0.794 and 0.925, indicating high consistency.

In Table IV, we also compare our data to learning outcomes produced using Item Response Theory (IRT) – the technique used by psychometricians to generate scores for each respective international and regional assessment. IRT models the probability a given pupil answers a given test item correctly as a function of pupil and item-specific characteristics (Mislevy, Beaton, Kaplan, and Sheehan 1992; Holland and Dorans 2006). While this methodology is used to construct the underlying tests we use, to use it to compare learning across assessments would require enough overlap in the test items. This is not true for a large enough set of tests and time periods to create a globally comparable panel data set.

When this overlap is small, standard maximum likelihood estimates will reflect both true variance and measurement error, overstating the variance in the test score distribution. Das and Zajonc (2010) elaborate on the various challenges of estimating IRT parameters with limited item-specific overlap. Sandefur (2018) shows that IRT produces unreliable results in a context with limited item overlap between an international and regional assessment in East and Southern Africa.

While IRT might not be a reliable approach when there is limited item-by-item overlap, we conduct a few comparisons where overlap is larger. We compare our results to the *Linking International Comparative Student Assessment* (LINCS) project which uses IRT methods and has significant overlap in items for a subset of international studies focused on reading at primary school (Steinmann, Strietholt**,** and Bos 2014). We find that our database can produce similar results to IRT methods for average scores where there is overlap with a correlation coefficient above 0.92. Figure A2 compares scores for the same subject (reading), schooling level (primary) and time period (2000-2010), with a correlation of over .98.

Table IV summarizes the results. In all cases the correlation is roughly .8 or higher. These comparisons indicate that as we expand coverage from 50 to 164 countries, we maintain high consistency with alternative measures where there is overlap.

V.B. Comparison to Schooling and Penn World Tables Human Capital Measures

We compare our measure of human capital with a schooling measure of human capital from Barro-Lee (2013) as well as a new measure of human capital in the Penn World Tables. The measure of human capital in the Penn World Tables is a combination of years of schooling from Cohen-Leker (2014) and Barro-Lee (2013) as well as returns to schooling estimates from Psacharopoulos (1994). We include these measures of human capital in growth regressions in Table V.

Each measure is significant on its own in columns (1)-(3). However, columns (4)-(7) show that when we include the measures together, the only measure that is consistently significant and associated with growth across multiple specifications is learning while schooling has smaller, less significant and sporadic association with growth. We also observe a larger R-squared of .26 when including learning.

This illustrative analysis indicates that the measure of human capital introduced in this database is associated with growth and development while schooling measures such as Barro-Lee (2013) are more limited. Thus, while schooling databases have global coverage, the association with growth is most robust when human capital is measured by learning. This finding is consistent with the recent literature across a smaller sample of 50 countries (Hanushek and Woessmann 2012a). We demonstrate that this stylized fact holds on a global scale across over 100 countries. This does not mean that schooling is not useful, but that it might lead to growth largely through the channel of learning.

This analysis further reveals that the human capital variable in the Penn World Tables might underestimate the role of human capital in economic development by relying mostly on a schooling dimension rather than learning.

VI. Stylized Facts

We present a series of stylized facts relevant to human capital and economic development. First, we observe that as schooling has increased over the last decade, learning has stagnated. Second, we observe a positive gender gap for learning in contrast to the negative gender gap observed for schooling. Third, we find human capital measured by learning is consistently associated with growth on a global scale, whereas schooling is not. Fourth, we find significant heterogeneity in growth patterns by income status. The wide coverage of our database enables meaningful heterogeneity analysis for previously underexplored levels of granularity. Fifth, we conduct an illustrative development accounting exercise and find that human capital accounts for between a fifth to half of cross-country income differences, roughly in the middle of the recent literature.

VI.A. Differences in Schooling versus Learning

Figure VII explores the contrast between changes in schooling and learning over time. We measure schooling using primary school enrollment rates. We compare this to the learning measure of human capital in our database for the years 2000-2010. We use data from 2000-2010 since this is the period with the most overlap of schooling and learning measures.

We see a clear trend towards increased schooling, while learning has stagnated and, in some instances, declined. Part of this decline, for example, in East Asia and Pacific is reflective of low-income countries being included in our database for the first time. In other cases, stagnation reflects a stall in learning. In both scenarios the data highlight an emerging human capital gap: students are increasingly in school, but learning is limited. These stylized facts reinforce recent emphasis by the World Bank and UNESCO of a 'learning crisis' (UNESCO 2017; World Bank 2018).

VI.B. The Gender Gap Is Flipped

Figure VIII disaggregates learning and enrollment data by gender. We see gender gaps for schooling and learning across most regions. Gender gaps are both larger and have more variance across years and regions for learning than for schooling. We further observe that the gender gaps move in opposite directions. In terms of enrollment, girls lag while on learning girls outperform boys.

This finding is consistent with the literature (Muralidharan and Prakash 2017). This flip in the gender gap might be due to selection. In regions where enrollment is low, only high-achievers might be taking assessments. However, in Sub-Saharan Africa, where enrollment is second lowest, the gender gap is null to negative. Moreover, across Europe and Central Asia there is little gender gap in enrollment yet meaningful gender gaps in learning. These stylized facts indicate selection is unlikely the only driver.

VI.C. Human Capital When Measured by Learning Is Associated with Growth

We examine the relationship between human capital and economic growth. We include a global distribution of countries in an analysis where human capital is measured by learning for the first time. A prior analysis by Hanushek and Woessmann (2012a) included 50 countries, about half of which were from the OECD. We analyze more than double this number of countries and include the largest number of developing countries to date – the countries with the most potential to gain from human capital accumulation.

We focus on the 2000-2010 interval which has the largest overlap in data on human capital as measured by both schooling and learning. This ensures that when making comparisons between these two types of measures of human capital we capture differences in the explanatory power of the human capital variable rather than time effects.

The full sample of countries for which we have growth, years of schooling and learning data is 132. We exclude two countries, the Syrian Arab Republic and the Republic of Yemen, due to recent civil wars as well as Zimbabwe which had a historic inflation crisis during this time period. We further exclude 12 countries for which more than 25 percent of their GDP is comprised of rents from natural resources. This cutoff is arbitrary, and the findings are robust to various cutoffs, including 15 and 20 percent. These economies have highly volatile growth rates and are driven by factors orthogonal to human capital, for example, shocks to oil and mineral prices. This is consistent with the approach taken in standard cross-country growth regressions (Mankiw, Romer, and Weil 1992; Hanushek and Woessmann 2012a) and is applied systematically using data from the World Bank on rents from natural resources as a share of GDP (Lange, Wodon, and Carey 2018). It is also typical to exclude former communist countries. We do not show results with this exclusion. When this exclusion is made the associations are slightly stronger. The final sample includes 117 countries. Table VI presents the results and Figure IX presents an associated added variable plot.

We find that learning is highly associated with growth, even when we control for initial GDP and years of schooling. The association is large and significant varying between the 1 percent and 10 percent level across nearly all specifications. This result is consistent with the recent literature (Hanushek and Woessmann 2012a). This similarity in findings is noteworthy on two fronts. First, with regards to the time period. We find broadly consistent results with prior analyses from 1960 to 2000. Second, with regards to the distribution of countries. We include 117 countries, including a significant number of developing countries. This indicates that prior associations between learning and growth are unlikely idiosyncratic - they hold even when a global distribution of countries is considered.

These findings are robust to multiple specifications. We include the average of years of schooling across the time period. We find that results persist when

controlling for the log of initial GDP at the 10 percent level. This allows for conditional convergence as well as variation in competing growth theories with advocates of the neoclassical model preferring this specification (Lucas 1988; Romer 1990; Mankiw, Romer, and Weil 1992). We find robust findings regardless of the model specified.

VI.D. Heterogeneity in Growth

A significant contribution of our database is the ability to conduct heterogeneity analysis since we include learning data for 164 countries. Whereas Hanushek and Woessmann (2012a) have one country in the low-income subgroup, we are now able to include 20. Our database enables growth analysis for Sub-Saharan Africa, Europe and Central Asia, and South Asia, as well as substantially improves coverage in Latin America and the Caribbean and the Middle East and North Africa. To this end, our database enables us to uncover precise trends within subgroups.

Table VII presents results by income status as classified by the World Bank as well as by initial quartile of income at the start of the growth period in 2000. These results are robust to multiple specifications similar to Table VI. We present only the main specification in the table.

We find a positive correlation between learning and growth in all cases. However, results are only significant for current upper middle-income countries. For upper middle-income countries, the association is over triple the average and significant at the 1 percent level. The R-squared is also triple the average at 0.635. When exploring results by initial income levels at the start of the period, we find the largest and most significant associations are for the second quartile with more than triple the average association between human capital and growth.

These results together indicate a potential human capital 'sweet spot' where investment in human capital is most associated with growth as countries move from the bottom to upper end of middle-income status, which has important implications for development research and policy. One potential interpretation of this stylized fact is that human capital might need to be accompanied by labor markets that value skills to harness human capital into productive assets, which is less often the case in the low-income countries but might be the case in middle-income countries. There are many alternative possible explanations for this stylized fact. We do not aim to adjudicate explanations, and simply present this fact to motivate the importance of heterogeneity and future research into these issues.

By expanding coverage to 164 countries our database enables heterogeneity analysis at a previously underexplored level of disaggregation. Such analyses have the potential to shed new insights into the relationship between human capital and economic development.

VI.E. Development Accounting

The development accounting literature studies the relative contribution of human capital in cross-country income differences. However, this question remains unsettled, in part due to difficulties in measuring human capital. While direct measures of years of schooling exist, the quality of schooling has often been inferred or measurement of quality has covered only a limited sample of countries.

Several approaches have emerged to estimate the quality of schooling, including cross-country differences in Mincerian returns (Hall and Jones 1999; Caselli 2005), immigrant returns (Schoellman 2011), and cross-country skill premia (Caselli and Coleman 2006). However, these approaches have encountered challenges such as the substitutability between skilled and unskilled workers (Jones 2014; Caselli and Ciccone 2018; Jones 2018). The challenges in measuring quality have contributed

to substantial variation in estimates of the role of human capital in accounting for cross-country income differences, from just one-fifth all the way to four-fifths (Hendricks and Schoellman 2017).

In this paper, we provide a direct and reliable measure of the quality of schooling on a global scale. We construct human capital stocks using our learning outcome data and produce development accounting results in a motivating application. We follow the literature (see Caselli (2005) for a review) and begin with a standard aggregate production function in its per-capita form following Klenow and Rodriguez-Clare (1997):

$$y = Ah\left(\frac{k}{y}\right)^{\frac{\alpha}{1-\alpha}} \tag{4}$$

where y represents output per worker, k denotes the capital-labor ratio, h denotes the level of human capital per capita, and A captures the residual, usually attributed to Total Factor Productivity (TFP). Taking the log on both sides decomposes cross-country income differences into three proximate sources: capital-output ratios, total factor productivity, and average human capital. Since we are only interested in the share of income differences that can be explained by variation in human capital, we provide decompositions of our baseline accounting results for the human capital share. In Table VIII, we provide decompositions as direct analogies to Schoellman (2011), comparing his estimates to measures relying on quantity (Halls and Jones 1999; Hendricks 2002). In Table IX we include an additional decomposition: $\frac{\ln(h_{90})-\ln(h_{10})}{\ln(y_{90})-\ln(y_{10})}$ which provides direct comparisons with the more recent literature using various quality measures (Hanushek and Woessmann 2012b; Jones 2014; Hendricks and Schoellman 2017; Caselli and Ciconne 2018).

To measure human capital, we extend the standard Mincer specification that weights education by its micro labor market returns to consider learning as well as schooling in line with Hanushek and Woessmann (2012b):

$$h = e^{rS+wL} \tag{5}$$

where S is the quantity of schooling and L is a measure of learning, and r and w are their respective returns. For years of schooling, we use Barro-Lee (2010) data. For learning measures, we use the data presented in this paper. We assume rates of return based on the microeconomic literature following Hanushek and Woessmann (2012b). We take the value r = .10 for the rate of return per school year based on Psacharopoulos and Patrinos (2004) and w = .20 per standard deviation increase in learning based on analysis conducted by Hanushek and Zhang (2009). The .20 value is based on U.S. data. However, we might expect that returns to skills will be higher in developing countries, where the supply of skills is lower, as is the case in the returns to schooling literature. Significant work has been done to identify this parameter value. For the purpose of this paper, our intention is not to provide a final result, but rather to motivate the use of the data for future use in the development accounting literature. To this end, we take parameter values as given and conduct sensitivity analyses with values w = .15 and w = .25. We include 131 countries in this development accounting exercise.[6]

Table VIII shows our results in comparison to Schoellman (2011), Hall and C. Jones (1999) and Hendricks (2002).We find that when our human capital measure only captures quantity (w = 0), we produce results similar to Hall and C. Jones (1999) and Hendricks (2002) where the share of human capital account for roughly 10 percent of output per worker differences. However, when we include quality,

[6] This includes all countries which have both Barro-Lee data as well as learning data. We do not exclude countries such as Syria, Yemen and Zimbabwe or resource-rich countries, which we excluded in earlier growth regressions since while shocks to growth might bias growth rates, they are less likely to bias income levels.

we find this share goes up to 15–22 percent. This estimate is in line with Schoellman (2011) estimates of 21-26 percent, which also account for both quantity and quality. These results suggest that measuring human capital with quality substantially increases the role of human capital in explaining cross-country output per worker differences.

In Table IX, we show results using the following decomposition: $\frac{\ln(h_{90})-\ln(h_{10})}{\ln(y_{90})-\ln(y_{10})}$ We compare results to the recent literature, which varies from nearly all (Jones 2014) to potentially none (Caselli and Ciccone 2018). We find that when including our measure of quality, the share of human capital varies between 39 to 49 percent. These estimates are in the vicinity of similar exercises measuring quality in Latin America (Hanushek and Woessmann 2012b) and using immigrant returns (Hendricks and Schoellman 2017).

This development accounting exercise is designed to be suggestive. Our central contribution is in providing a direct measure of schooling quality, and thus a better measure of human capital. We find that human capital accounts for between a fifth to half of cross-country income differences – a middle ground in the literature.

VII. Conclusion

To understand and track human capital formation, a critical ingredient for development, there is need for globally comparable measurement of learning. The growth of international standardized achievement tests is a significant step in this direction. However, many of the countries that participate in these tests are often already rich. This limits the ability to track, compare or understand education patterns in developing countries – the countries that often have the most potential to gain from human capital formation.

We build on the literature on comparable learning outcomes to construct a globally comparable database of 164 countries from 2000-2017, representing more than 98 percent of the global population. Approximately two-thirds of these countries are developing economies.

We contribute to the literature in several ways. This is the most current and expansive cross-country data set on learning, including the most developing countries. We hope this data set can be used to reveal important descriptive trends in human capital formation across developed and developing countries. Descriptive trends suggest learning is stagnating and that there is a female learning premium.

We also hope this database enables analysis of factors correlated with, and that have plausible causal links to, the formation of human capital and economic growth. In a first application on a global scale, we find that human capital is consistently associated with growth, reinforcing the recent literature. We also find significant heterogeneity, indicating scope for generation of new insights about the relationship between human capital and development. We further find that this database provides a measure of human capital that is more explanatory than schooling measures and the current measure included in the Penn World Tables 9.0.

We also contribute to a development accounting literature. To date the literature has explored an array of techniques to infer school quality in the absence of globally comparable data. This has contributed to a wide range of estimates of the role human capital plays in cross-country income differences, from nearly all to potentially none. We provide a direct measure of school quality to build human capital stocks. We find that human capital accounts for between a fifth to half of cross-country income differences – a middle ground in the literature. We hope our measure of human capital can further inform this debate.

This database comes at a moment when a series of global efforts have been launched to measure and track learning on a global scale. A notable effort to do this is the World Bank's Human Capital Index which compares countries' levels of

human capital around the world (Kim 2018a; Kraay 2019; World Bank 2019). The Human Capital Index includes learning outcomes from this database as one of its core ingredients in addition to health (Kim 2018b). The database introduced in this paper will be updated regularly, will be made publicly available, and feature in future large-scale efforts to understand and track human capital and economic development.

Future iterations of this database will expand coverage across countries and time as countries join existing assessments and by including additional assessments such as early grade reading and mathematics assessments. Moreover, we aim to build a database that enables over-time isolation of value-added learning by including variables which can account for various selection effects. This includes adding measures of enrollment and retention across schooling levels. We also aim to enable further identification of the link between human capital and economic growth by including variables such as comparable estimates on the returns to education. This database and future iterations will enable a deeper understanding of mechanisms driving human capital formation and the link to economic development.

UNIVERSITY OF OXFORD AND WORLD BANK
LONDON SCHOOL OF ECONOMICS AND WORLD BANK
YALE AND WORLD BANK
WORLD BANK

APPENDIX

A. International Standardized Achievement Tests (ISATs)

In the mid-1990s, there was an emergence of standardized, psychometrically-robust and relatively consistent ISATs. Below we describe the major ISATs we use in this database.

*TIMSS***.** The Trends in International Mathematics and Science Study (TIMSS) is conducted by the IEA. Five TIMSS rounds have been held to date in Math and Science subjects covering grades 4 and 8. The first, conducted in 1995, covered 45 national educational systems and three groups of students.[7] The second round covered 38 educational systems in 1999, examining pupils from secondary education (grade 8). The third round covered 50 educational systems in 2003, focusing on both primary and secondary education (grades 4 and 8). In 2007, the fourth survey covered grades 4 and 8 and more than 66 educational systems. In 2011, the survey covered 77 educational systems across grades 4 and 8. The last round was performed in 2015 and covered 63 countries/areas. The precise content of the questionnaires varies but remains systematic across countries.

[7] IEA assessments define populations relative to specific grades, while PISA assessments focus on the age of pupils. In IEA studies, three different groups of pupils were generally assessed: pupils from grade 4, grade 8 and from the last grade of secondary education. In 1995, two adjacent grades were tested in both primary (3-4) and secondary schools (7-8). To obtain comparable trends, we restricted the sample to grades 4 and 8. Some Canadian provinces and states in the United States of America have occasionally taken part in the IEA surveys.

*PIRLS***.** The Progress in International Reading Literacy Study (PIRLS) survey is also conducted by the IEA. The PIRLS tests pupils in primary schools in grade 4 in reading proficiency. Four rounds of PIRLS have been held to date in 2001, 2006, 2011 and 2016.

In 2006, PIRLS included 41 countries/areas, two of which were African countries (Morocco and South Africa), 4 lower-middle-income countries (Georgia, Indonesia, Moldova, Morocco) and 8 upper-middle-income countries (Bulgaria, Islamic Republic of Iran, Lithuania, Macedonia, Federal Yugoslavian Republic, Romania, Russian Federation, South Africa). The 2011 round of PIRLS was carried out alongside TIMSS and included 60 countries/areas. The newest round of PIRLS in 2016 includes 50 countries.

*PISA***.** The Organization for Economic Co-operation and Development (OECD) launched the Programme for International Student Assessment (PISA) in 1997 to provide comparable data on student performance. Since 2000, PISA has assessed the skills of 15-year-old pupils every three years. PISA concentrates on three subjects: mathematics, science and literacy. The framework for evaluation remains the same across time to ensure comparability. In 2009, 75 countries/areas participated; in 2012, 65 countries/areas participated and in 2015, 72 countries/areas participated. An important distinction between PISA and IEA surveys is that PISA assesses 15-year-old pupils, regardless of grade level, while IEA assessments assess grade 4 and 8.

B. Regional Standardized Achievement Tests (RSATs)

In addition to the above international assessments, a series of regional assessments have been conducted in Africa and Latin America and the Caribbean.

*SACMEQ***.** The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ). SACMEQ is a psychometrically designed, standardized test which generally assesses math, reading and English in grade 6 pupils. The first SACMEQ round took place between 1995 and 1999. SACMEQ I covered seven different countries and assessed performance only in reading. The participating countries were Kenya, Malawi, Mauritius, Namibia, United Republic of Tanzania (Zanzibar), Zambia and Zimbabwe. The studies shared common features (instruments, target populations, sampling and analytical procedures). SACMEQ II surveyed pupils from 2000-2004 in 14 countries: Botswana, Kenya, Lesotho, Mauritius, Malawi, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania (Mainland), Tanzania (Zanzibar), Uganda, and Zambia. Notably, SACMEQ II also collected information on pupils' socioeconomic status as well as educational inputs, the educational environment and issues relating to equitable allocation of human and material resources. SACMEQ II also included overlapping items with a series of other surveys for international comparison, namely the *Indicators of the Quality of Education* (Zimbabwe) study, TIMSS and the 1985-94 IEA *Reading Literacy Study*. The third SACMEQ round (SACMEQ III) spans 2006-2011 and covers the same countries as SACMEQ II plus Zimbabwe. SACMEQ collected its latest round of data in 14 countries in East and Southern Africa from 2012-2014. These include Botswana, Kenya, Lesotho, Mauritius, Malawi, Mozambique, Namibia, Seychelles, South Africa, Tanzania, Uganda, Zambia, Zanzibar and Zimbabwe. SACMEQ was designed and scaled to be comparable to past rounds. We include microdata from prior rounds, and estimates from reports for the latest round of SACMEQ since the microdata are pending.

*PASEC***.** The "Programme d'Analyse des Systèmes Éducatifs" (PASEC, or "Programme of Analysis of Education Systems") was launched by the Conference of Ministers of Education of French-Speaking Countries (CONFEMEN). These surveys are conducted in French-speaking countries in Sub-Saharan Africa in primary school (grades 2 and 5) in math and French. Each round includes 10 countries. PASEC I occurred from 1996 to 2003; PASEC II from 2004 to 2010 and PASEC III was conducted in 2014. Of note, PASEC has not always been conducted simultaneously across countries and participation has varied considerably since 1994.[8] The most recent PASEC in 2014 uses Item Response Theory (IRT). Ten countries participated, including Benin, Burkina Faso, Burundi, Cameroon, Chad, Republic of Congo, Côte d'Ivoire, Niger, Senegal and Togo. We include these countries using available microdata. Madagascar also participated in 2015 and was scaled to the PASEC 2014 round. We include Madagascar in our database using estimates from reports. To this end, inclusion of the recent PASEC data adds countries as well as enhances the quality of data from Sub-Saharan Africa. This marks significant improvement over past data sets. To provide a link to past PASEC rounds, which used classical test theory, we create an inter-temporal comparison using an exchange rate derived based on Togo, which participated in all rounds of PASEC. However, given that PASEC did not conduct intertemporal scaling calibration directly, intertemporal comparisons for PASEC should be analyzed with this caveat in mind.

*LLECE***.** The Latin American Laboratory for Assessment of the Quality of Education (LLECE) was formed in 1994 and is coordinated by the UNESCO

[8] The following is a list of participating countries in chronological order: Djibouti (1994), Congo (1994), Mali (1995), Central African Republic (1995), Senegal (1996), Burkina Faso (1996), Cameroon (1996), Côte d'Ivoire (1996), Madagascar (1997), Guinea (2000), Togo (2001), Mali (2001), Niger (2001), Chad (2004), Mauritania (2004), Guinea (2004), Benin (2005), Cameroon (2005), Madagascar (2006), Mauritius (2006), Republic of Congo (2007), Senegal (2007), Burkina Faso (2007), Burundi (2009), Côte d'Ivoire (2009), Comoros (2009), Lebanon (2009), Togo (2010), Democratic Republic of Congo (2010), and Chad (2010). Additional countries took a slightly different test between 2010 and 2011 (Lao PDR, Mali, Cambodia and Vietnam).

Regional Bureau for Education in Latin America and the Caribbean. Assessments conducted by the LLECE focus on achievement in reading and mathematics in primary school. The first round was conducted in 1998 across grades 3 and 4 in 13 countries. These countries include: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Honduras, Mexico, Paraguay, Peru and Venezuela. The second round of the LLECE survey was initiated in 2006 in the same countries as LLECE I. In round two, called the Second Regional Comparative and Explanatory Study (SERCE), pupils were tested in grade 3 and grade 6. The Third Regional Comparative and Explanatory Study (TERCE), was done in 2013 across grades 3 and 6 and included 15 Latin American and Caribbean countries. We only include SERCE and TERCE data in this database, since these assessments are most similar and cover comparable grades.

*MLA***.** A joint UNESCO and UNICEF project, *Monitoring Learning Achievement* (MLA), covers more than 72 countries and ranges from early childhood, basic and secondary education to non-formal adult literacy. In our database, we use data across two subjects, mathematics and reading/literacy. We include data from reports for MLA I across 11 African countries (Botswana, Madagascar, Malawi, Mali, Morocco, Mauritius, Niger, Senegal, Tunisia, Uganda and Zambia).

C. The Early Grade Reading Assessment (EGRA)

The Early Grade Reading Assessment (EGRA) is a basic literacy assessment conducted in early grades. The assessment is conducted most often in grades 2-4. Since 2006, EGRA has been conducted in over 65 countries. EGRA was developed by RTI and is typically implemented by USAID, RTI and local partners (Gove 2009).

The assessment is a short oral assessment conducted with a child one-on-one. EGRA is designed to be flexible and adapted across countries and contexts, while

maintaining core modules and similarities. EGRA is a timed test, enabling uniformity in how it is conducted. The tests often represent the most common features of the local language and align with the expectations of the grade level. EGRA includes up to thirteen subtasks, such as 'oral reading fluency', 'vocabulary', 'diction', and 'reading comprehension.' Multiple questions are included in each subtask to test proficiency. Of the thirteen subtasks, there are a few subtasks encouraged to be delivered across all countries and contexts (Dubeck and Gove 2015).

We compile and include data from the 'reading comprehension' indicator in EGRA from 48 countries. This indicator is available in nearly all EGRA data sets and is less sensitive to differences in context, implementation and language. It also and has a strong conceptual link to RSATs and ISATs (Abadzi 2008; Dubeck and Gove 2015) which also measure reading comprehension. To ensure robustness to language effects, we only include data when students took the test in their language of instruction. We use data for grades 2-4, which EGRA is designed for, although certain countries will participate out of this range. We restrict data used for our database to grades 2-4 to be consistent with the design of EGRA. We scale the EGRA microdata to a mean of 500 and standard deviation of 100. This scale corresponds to the scale used by RSATs and ISATs. We include all EGRA data from 2007-2017 as one round. This ensures our scaling is not biased by changing distributions of countries. In the future, we will consider new EGRA data as part of a future round and will conduct intertemporal comparisons using a similar approach to PISA (OECD 2015). Patrinos and Angrist (2018) provide additional detailed analysis and robustness checks on the inclusion of EGRA data.

The inclusion of EGRA adds 48 countries to the database with at least one data point in the past 10 years, nearly all of which are developing economies. Of the 48 countries, nearly two-thirds (31 countries) have data that is nationally representative. Exchange rates for EGRA are derived using countries with

nationally representative data only, to ensure the assumptions underlying the construction of the exchange rate hold. We include countries with non-representative data only when the alternative is no data. We include a dummy variable indicating when the data is not nationally representative to enable users of the database to analyze the data accordingly.

D. Summary of Assessments Included in the Database

We include eight learning assessments in our database. Table A1 in the Appendix summarizes the assessments included. Table A2 further describes the distribution of source assessments included in our database by country-level-year observations. Most regional assessments are done at the primary level. Moreover, regional assessments comprise nearly 25 percent of country-level-year observations, marking substantial representation of developing countries.

E. Additional Methodological Parameters

Over-time Comparability.— ISATs and RSATs have been designed to be comparable since the late 1990s and early 2000s. Thus, the use of these modern assessments enables comparability over time from this time period onwards.

Time Intervals.— While this is one of the largest and most comprehensive comparable learning outcomes database produced to date, it is still sparse given limited test frequency. In other databases, the data is often disaggregated over 5-year periods. This produces continuously spaced intervals, is designed to reduce noise by averaging results within these intervals, and is comparable to the Barro-Lee approach for years of schooling. In this database, we have moved away from this approach. We now provide the year of test as documented in official reports.

This enables greater granularity and precision of the data and enables the users of the database to make trade-offs at their discretion.

Schooling Levels.— We construct a score for each grade and pool across grades within each schooling level to produce primary and secondary school scores. We distinguish primary from secondary schooling since enrollment rates drop off between levels in many developing countries. This introduces a potential selection term in secondary school scores, with the highest performing students progressing in the system, biasing scores up due to selection rather than actual learning.

Conceptually, the broader categories of 'primary' and 'secondary' scores enable us to categorize learning at schooling levels across assessments which span multiple grades and age groups. If the test is designed for an age group (for example, PISA) we code it at the relevant schooling level (for example, secondary for PISA). We specify an approach to including specific grade levels to ensure we have a tight grade interval within one to two years to minimize scope for grade-fixed effects. While the interval is relatively small, it still leaves room for grade-fixed effects rather than test-fixed effects when linking tests. For example, linking PIRLS 2001 grade 4 with SACMEQ 2000 grade 6 might capture a grade difference in PIRLS in addition to difficulty. However, to enable greater country coverage, we put up with the need to expand beyond single grade level intervals. Moreover, these differences are often small and since exchange rates are applied to all tests being linked, original ranks will be preserved. An analysis of EGRA in Figure A4 demonstrates sensitivity to grade. We run a regression with and without grade-fixed effects comparing mean scores relative to a country which participates across all three grade levels 2-4. We find small differences, with near complete overlap in the confidence intervals on the grade and non-grade-fixed estimates. This sensitivity analysis increases our confidence that the EGRA data, and other regional assessment data, is robust to data availability by grade. We also include a variable with grade information in the database to make this transparent.

Subjects.— We construct exchange rates specific to reading math, and science. While the proficiency is not granular at the test item level, this ensures that there is significant proficiency overlap when tests are being put on a global scale.

Subsamples.—When calculating the HLO by gender we apply the average exchange rate to each subsample, rather than constructing subsample specific exchange rates. While performance is likely to vary across subsamples in a given test, the relationship between pair-wise tests being linked is unlikely to vary across subsamples nor relative to the full sample.

Metadata versus Aggregate Data.— Our database is disaggregated by subject, schooling level, grade, year and source test. We call this version the 'metadata.' The final data series used in the Human Capital Index (HCI) aggregates the metadata presented in this paper. The aggregation used in the HCI is described in depth in Kraay (2019). There are multiple ways to aggregate the data. For example, the HCI averages data across schooling levels and subjects and uses the most recent year available. The HCI further combines data differently depending on the testing source, for example, including EGRA data in the final time series only when no other data is available. This implicitly weights the importance of testing source over schooling level or subject. Alternative aggregations of the metadata are possible. We present the metadata in this database to enable users to make judgements based on the purpose of their analysis and for maximum transparency. In the analyses presented in this paper we use the time series most consistent with the HCI procedure.

Exceptions.— In unusual cases, the procedures practiced for a given international or regional test are adapted for the country context. For example, sixth grade students in Botswana took TIMSS instead of fourth grade in 2011. Another example

includes India and China, where only certain states and cities participated in PISA. These variations are acknowledged by the underlying tests and the data is caveated with an asterisk in published reports. We preserve this information in our data, and include notes in the metadata for each case. We describe each case in detail in Patrinos and Angrist (2018). In the case of India, we verify that the state data is likely to be nationally representative using national assessments.

We make an adjustment beyond the underlying tests in the case of China given the likelihood that China's current PISA data is biased. The China HLO based on 2015 PISA data is from four cities (Beijing, Shanghai, Jiangsu, and Guangdong) and is 532. However, this data is likely biased upwards since the cities participating are urban and rich relative to the national average. We adjust the score based on socioeconomic information by city and across the nation and produce an average HLO of 462 at the secondary level, which is plausibly representative at the national level. The detailed procedure is described in Patrinos and Angrist (2018).

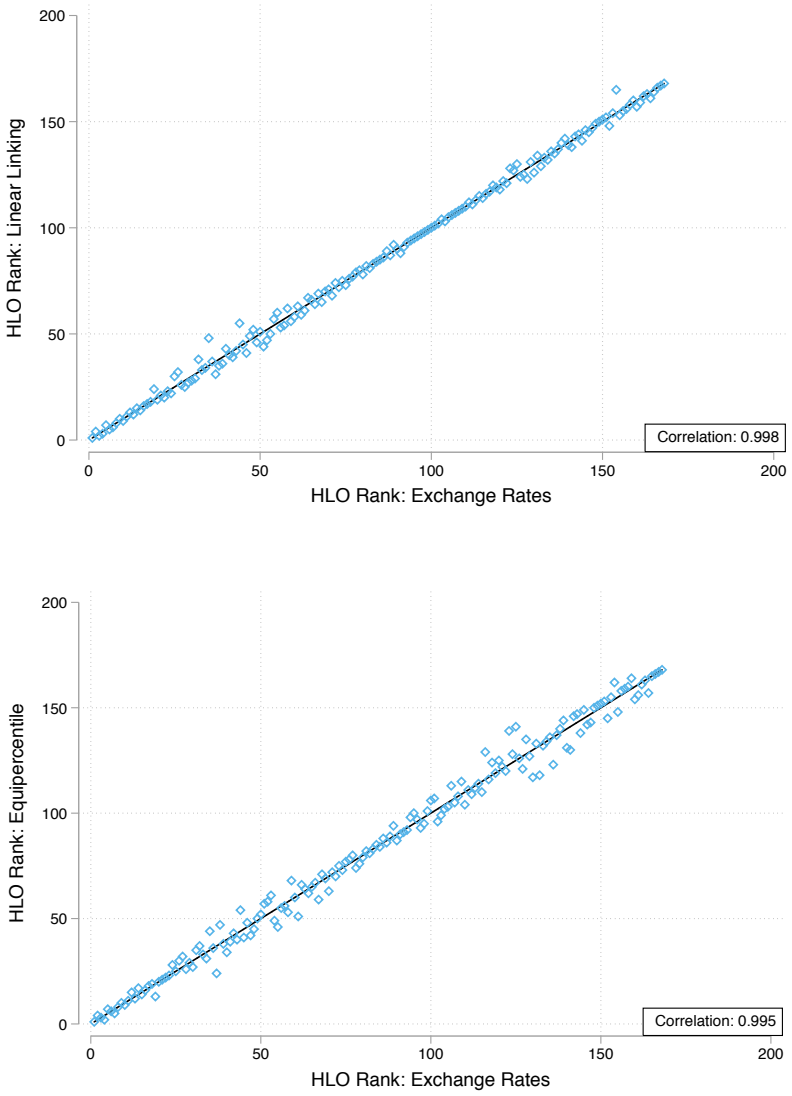


FIGURE A1

Comparison of Ranks to Classical Linking Methods (2000-2015)

Notes: In an older dataset (Altinok, Angrist, and Patrinos 2018) we produced scores using the exchange rate definition used in this paper as well as two classical methods including linear linking as well as equipercentile linking. This figure shows the spearman rank correlation of scores for both of these methods. We use average scores over the period 2000-2015 and produce country ranks based on these scores. We see correlations above .995. This indicates the methodology proposed is robust to score scales.

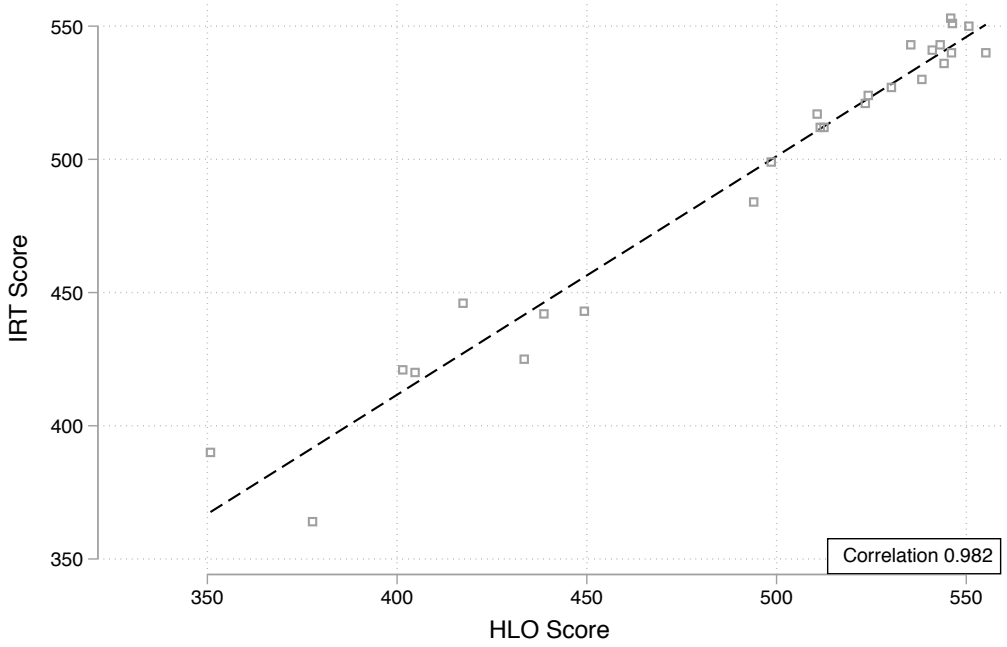


FIGURE A2

Comparison to Item Response Theory (IRT) Linking (2000-2010)

Notes: We compare our data to the *Linking International Comparative Student Assessment* (LINCS) project which uses Item-Response Theory (IRT) methods to link test score data. IRT methods are used to create scores for the underlying international and regional assessments used. However, to compare across assessments, IRT would require enough overlap in the test items. This is not true for a significant enough set of tests and time intervals to create a globally comparable panel data set. However, for a subset of tests, this is the case, such as a series of international studies focused on reading at primary school. The LINCS project produces scores on a globally comparable scale using this subset of data (Steinmann, Strietholt**,** and Bos 2014). We compare results on the HLO as well as the LINCS data for primary school reading on average between 2000-2010.

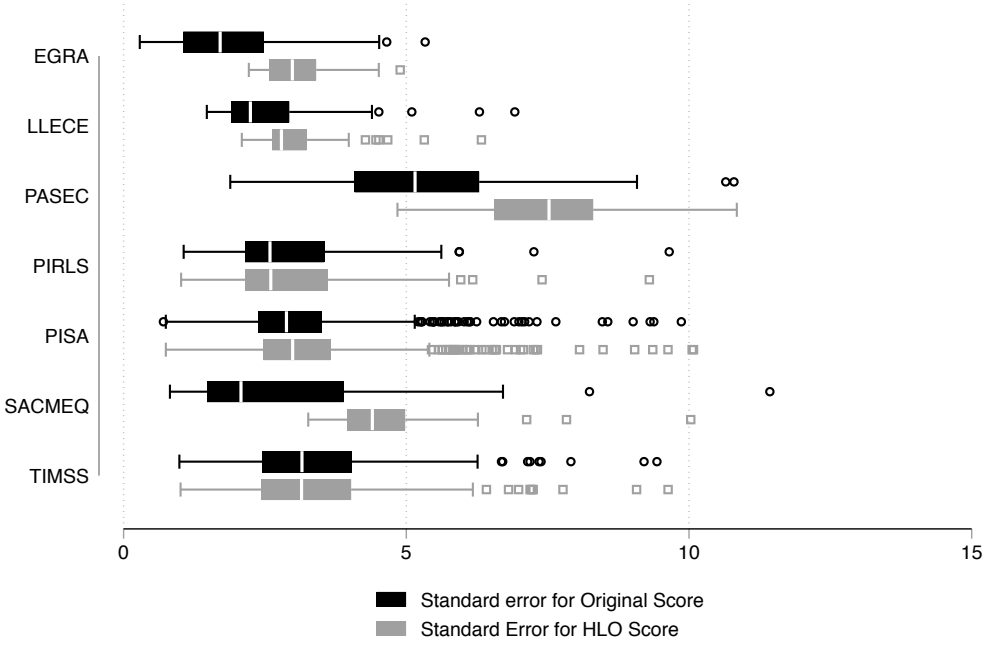


FIGURE A3

Standard Errors by Test

Notes: We decompose standard errors on the overall HLO score versus the original test. This reveals the source of uncertainty derived from sampling variation from the original test relative to variation introduced from the production of exchange rates.

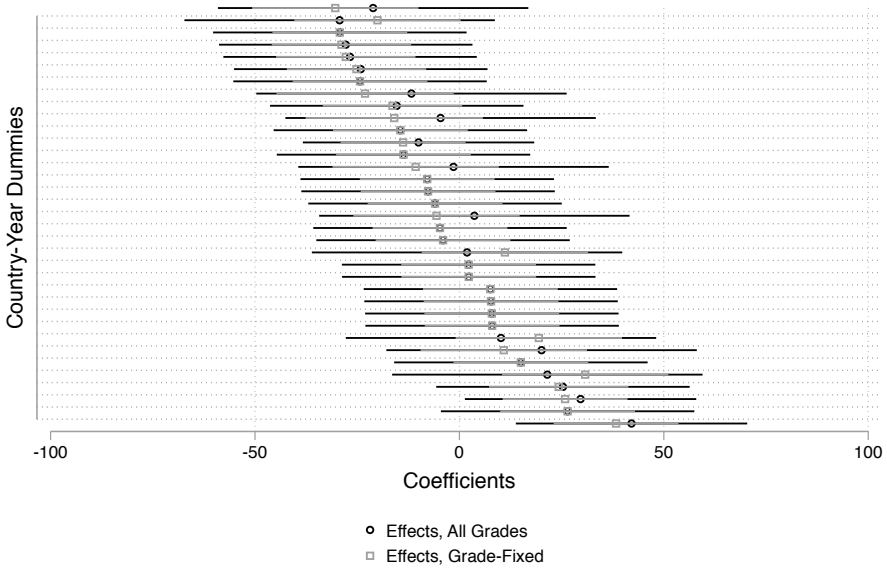


FIGURE A4

EGRA, Grade-Fixed Effects

Notes: We run a regression with and without grade-fixed effects comparing mean scores relative to a given country which participates across all three grade levels 2-4 using EGRA data. We only include nationally representative data. The darker "All Grades" bar represents the confidence interval without grade-fixed effects. The lighter "grade-fixed effects" represents the confidence interval with grade-fixed effects.

TABLE A1

REVIEW OF STUDENT ACHIEVEMENT TESTS

Organization	Abbr.	Year	Subject	Countries/ Areas	Grade/Age
IEA	TIMSS	Every four years since 2003 (latest round is 2015)	M,S	38, 26, 48, 66, 65	4,8
UNESCO	MLA	2000	M,S,R	72	6,8
UNESCO	LLECE	2006, 2013	M,S,R	13, 16 (only 6 for science)	3,6
UNESCO	SACMEQ	2000, 2003, 2007, 2013	M,R	7, 15, 16	6 6
CONFEMEN	PASEC	2006, 2014	M,R	22 (before 2014), 10	Until 2014: 2,5 After 2014: 3, 6
IEA	PIRLS	Every five years since 2001 (latest round is 2016)	R	35, 41, 55	4
OECD	PISA	Every three years since 2000 (latest round is 2015)	M,S,R	43, 41, 57, 74, 65, 71	Age 15
RTI/USAID	EGRA	2007-2017	R	65	2,3,4

Notes: When denoting subjects, M=math; S=science; and R=reading.

TABLE A2

DISTRIBUTION OF SOURCE TEST FOR HLO

	All		Primary		Secondary	
Test	Frequency	Percent	Frequency	Percent	Frequency	Percent
EGRA	72	6.75	72	14.20	0	0.00
LLECE	62	5.81	31	6.11	0	0.00
MLA	25	2.34	25	4.93	0	0.00
PASEC	40	3.75	30	5.92	0	0.00
PIRLS	161	15.09	160	31.56	0	0.00
PISA	345	32.33	0	0.00	345	66.60
SACMEQ	39	3.66	39	7.69	0	0.00
TIMSS	323	30.27	150	29.59	173	33.40

Notes: We include country-year-level observations by source test based on the metadata. We do not include country-year-level-subject observations in this table. Features of the data by subject are captured in Tables 2 and 3.

TABLE A3

AVERAGE SCORES, BY GENDER (2000-2017)

Country	HLO	HLO - Female	HLO - Male	Country	HLO	HLO - Female	HLO - Male	Country	HLO	HLO - Female	HLO - Male
	Tercile 1				Tercile 2				Tercile 3		
Singapore	571.47	576.94	566.28	Gabon	456.13	454.44	457.92	Jamaica	387.13	403.77	371.97
Korea, Rep.	566.78	566.42	566.96	China	456.00	456.00	456.00	Tuvalu	386.90	392.32	380.96
Hong Kong SAR, China	562.56	564.68	560.62	Mauritius	450.03	444.20	444.53	Paraguay	386.34	386.54	386.95
Finland	558.13	567.36	549.04	United Arab Emirates	448.55	462.19	434.76	Peru	386.34	385.12	387.89
Japan	555.64	556.40	554.89	Trinidad and Tobago	448.03	462.65	433.46	Philippines	385.07	392.38	377.31
Taiwan, China	554.15	556.73	551.66	Tajikistan	444.44	448.00	440.94	Mozambique	384.90	387.90	380.82
Canada	544.38	548.38	541.03	Chile	444.06	440.32	447.65	Congo, Rep.	384.40	383.67	385.11
Netherlands	540.41	541.46	539.37	Seychelles	443.73	415.31	452.84	Kiribati	383.09	394.42	371.60
Estonia	538.35	543.48	533.37	Thailand	443.56	452.60	432.37	Qatar	382.65	399.40	366.49
Macao SAR, China	534.33	536.98	531.81	Costa Rica	442.72	440.49	445.23	Uganda	381.67	375.23	372.72
Liechtenstein	533.96	533.69	534.19	Uruguay	439.07	441.60	436.31	Benin	380.55	377.18	382.81
Switzerland	529.86	531.59	528.38	Azerbaijan	437.20	442.56	432.34	El Salvador	380.03	374.92	385.54
Australia	529.65	531.78	527.65	Georgia	435.63	445.37	426.47	Sudan	379.63	389.37	371.29
New Zealand	529.16	534.44	524.16	Bahrain	435.24	453.83	416.95	Algeria	377.83	383.09	373.05
Ireland	529.00	531.86	526.33	Mongolia	434.62	434.46	434.91	Papua New Guinea	377.83	380.84	374.84
Cuba	526.59	530.71	522.84	Kenya	434.26	428.27	419.52	Kuwait	376.55	398.68	353.60
Germany	526.53	528.80	524.58	Iran, Islamic Rep.	430.21	438.91	422.09	Côte d'Ivoire	374.91	373.97	375.70
Vietnam	526.28	530.30	521.91	Mexico	427.58	428.02	427.04	Kosovo	374.80	380.88	368.91
Belgium	526.19	529.03	523.95	Venezuela, RB	427.19	425.70	428.98	Tonga	373.39	385.85	360.90
United Kingdom	524.39	525.48	523.62	Jordan	426.24	443.23	409.57	Namibia	371.76	353.06	355.47
Denmark	522.61	523.40	521.97	Cambodia	426.10			Timor-Leste	368.86	373.71	364.32
Sweden	521.43	527.54	515.67	Burundi	423.75	425.81	422.33	Nepal	368.56		
Czech Republic	520.01	522.97	517.39	Montenegro	423.07	431.78	414.63	Chad	368.21	363.80	370.50
United States	519.96	521.17	518.72	Myanmar	421.51	428.98	415.09	Lao PDR	368.14	373.62	362.57
Austria	519.94	520.94	519.20	Swaziland	419.93	409.42	410.03	Lesotho	366.67	351.62	355.12
Slovenia	518.40	525.48	511.63	Lebanon	419.29	416.56	422.68	Iraq	363.43	376.71	353.08
Poland	517.92	523.08	512.69	Honduras	416.21	416.76	415.69	Solomon Islands	362.19	373.08	351.37
France	516.98	520.80	513.19	Macedonia, FYR	416.14	426.24	406.94	Gambia, The	360.90	358.83	363.42
Hungary	516.91	519.82	514.13	Colombia	415.53	412.80	418.47	Ethiopia	359.37	355.77	362.93
Lithuania	512.31	520.60	504.34	Cameroon	415.05	418.19	412.48	Kyrgyz Republic	357.50	366.89	347.55
Latvia	511.44	520.57	502.44	Senegal	413.60	408.49	418.64	Vanuatu	355.62	360.09	351.17
Russian Federation	511.43	516.77	506.07	Argentina	411.91	417.41	405.46	India	354.99	361.76	347.08
Iceland	510.59	519.90	501.73	Tanzania	410.60	419.10	402.41	Congo, Dem. Rep.	354.82	351.06	358.09
Slovak Republic	507.04	509.91	504.24	Guinea	408.25	397.35	416.87	Afghanistan	354.76	350.15	357.79
Italy	506.77	508.65	505.08	Brazil	405.24	405.11	405.53	Dominican Republic	353.59	356.00	352.49
Norway	505.74	512.09	500.01	Albania	404.16	417.59	390.85	Rwanda	350.75	356.33	345.31
Spain	504.05	505.41	502.95	West Bank and Gaza	403.63	415.26	391.26	Bangladesh	348.24	349.89	346.49
Croatia	503.29	507.68	499.13	Syrian Arab Republic	403.59	397.86	407.20	Guyana	346.37	356.22	337.77
Luxembourg	503.19	504.19	502.52	Oman	403.29	426.76	379.39	Morocco	345.93	352.51	340.28
Portugal	501.94	504.47	499.49	Burkina Faso	402.70	398.67	407.87	South Africa	345.47	337.51	324.15
Israel	487.38	490.77	484.28	Indonesia	402.15	405.79	398.52	Malawi	342.65	337.76	330.50
Greece	483.64	491.53	475.96	Sri Lanka	400.00			Zambia	342.50	337.18	332.09
Kazakhstan	482.88	487.15	478.64	Togo	396.93	396.18	397.52	Mauritania	342.09	343.34	340.48
Cyprus	481.36	485.94	476.89	Ecuador	396.82	398.39	395.44	Haiti	337.80	343.49	332.96
Ukraine	481.30	480.73	481.88	Zimbabwe	395.17	392.21	395.73	Niger	337.63	335.16	339.14
Bulgaria	479.25	489.64	469.42	Madagascar	392.37	392.84	391.84	Liberia	337.51	333.33	341.02
Serbia	475.73	479.20	472.27	Botswana	392.31	388.54	379.15	Pakistan	335.71	340.73	330.74
Malta	472.93	480.69	465.24	Comoros	392.23	386.86	399.83	Mali	335.38	332.38	337.66
Armenia	464.52	469.55	459.63	Tunisia	390.63	392.48	388.90	South Sudan	334.55		
Bosnia and Herzegovina	460.80	460.30	461.28	Saudi Arabia	390.51	407.31	374.23	Belize	326.83	341.73	314.56
Romania	459.43	465.20	453.57	Egypt, Arab Rep.	389.68	396.40	382.97	Angola	325.97	324.92	326.81
Moldova	458.99	467.91	450.26	Panama	387.53	389.05	386.44	Nigeria	325.25	320.94	329.34
Turkey	458.63	464.44	453.28	Guatemala	387.37	388.47	386.99	Sierra Leone	315.88	313.94	317.84
Malaysia	456.19	461.98	450.10	Nicaragua	387.30	386.63	388.54	Ghana	299.38	288.58	308.60
								Yemen, Rep.	292.51	294.81	292.40

Notes: We average scores by country over all subjects and levels from the period 2000-2017, split by terciles.

TABLE A4

TEST LINKING ARCHITECTURE

Test X	Test Y	Subject	Level	Overlapping Countries
PISA	TIMSS	Math, Science	Secondary	Australia, Bulgaria, Canada, Chile, Chinese Taipei, Colombia, Czech Republic, Finland, Georgia, Hong Kong – China, Hungary, Indonesia, Israel, Italy, Japan, Jordan, Kazakhstan, Korea, Republic of, Latvia, Lebanon, Lithuania, Macedonia F.Y.R., Malaysia, Malta, Netherlands, New Zealand, Norway, Qatar, Romania, Russian Federation, Serbia, Singapore, Slovakia, Slovenia, Sweden, Thailand, Tunisia, Turkey, USA, UAE.
SACMEQ	PIRLS	Reading	Primary	Botswana
SACMEQ	TIMSS	Math	Primary	Botswana
LLECE	PIRLS	Reading	Primary	Colombia, Chile, Honduras
LLECE	TIMSS	Math, Science	Primary	Colombia, Chile, Honduras, El Salvador
PASEC Round 1	SACMEQ	Reading, Math	Primary	Mauritius
PASEC Round 2	PASEC Round 1	Reading, Math	Primary	Togo
EGRA	PIRLS	Reading	Primary	Egypt, Honduras, Indonesia
MLA	SACMEQ	Math, Reading	Primary	Botswana, Malawi, Mauritius, South Africa, Uganda, Zambia

Notes: For ease of representation, we include countries used at any point in time for each test linking procedure. In some rounds, some countries are not included, since we specify that for a given round to be linked, tests should be administered in adjacent years. A more detailed architecture by year is available on request.

REFERENCES

Abadzi, Helen. "Efficient learning for the poor: New insights into literacy acquisition for children." *International Review of Education* 54, no. 5-6 (2008): 581-604.

Altinok, Nadir, Noam Angrist, and Harry A. Patrinos. *Global Dataset on Education Quality 1965-2015*. World Bank Policy Research Working Paper No. 8314, 2018.

Barro, Robert J. "Economic growth in a cross section of countries." *Quarterly Journal of Economics* 106, no. 2 (1991): 407-443.

Barro, Robert J., and Jong-Wha Lee. "International data on educational attainment: updates and implications." *Oxford Economic Papers* 53, no. 3 (2001): 541-563.

Barro, Robert J., and Jong Wha Lee. "A new data set of educational attainment in the world, 1950–2010." *Journal of Development Economics* 104 (2013): 184-198.

Barro, Robert J., and Jong-Wha Lee. *Education matters. Global Schooling Gains From the 19th to the 21st Century*. Oxford University Press, 2015.

Becker, Gary S. "Investment in human capital: A theoretical analysis." *Journal of Political Economy* 70, no. 5, Part 2 (1962): 9-49.

Caselli, Francesco. "Accounting for Cross-country Income Differences." *Handbook of Economic Growth*, 1 (2005), pp.679-741.

Caselli, Francesco and Antonio Ciccone. "The Human Capital Stock: A Generalized Approach. Comment." *American Economic Review*, 109, no. 3 (2019): 1155-74

Caselli, Francesco and John Coleman. "The World Technology Frontier." *American Economic Review* 96.3 (2006): 499-522.

Cohen, Daniel and Laura Leker. *Health and Education: Another Look with the Proper Data*. Manuscript, 2014.

Cohen, Daniel and Marcelo Soto. "Growth and Human Capital: Good Data, Good Results." *Journal of Economic Growth* 12, no.1 (2007): 51-76.

Das, Jishnu and Tristan Zajonc. "India Shining and Bharat Drowning: Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement." *Journal of Development Economics* 92, no.2 (2010): 175–187

Dubeck, Margaret M. and Amber Gove. "The Early Grade Reading Assessment (EGRA): Its Theoretical Foundation, Purpose, and Limitations." *International Journal of Educational Development* 40 (2015): 315-322.

Feenstra, Robert. C., Robert Inklaar and Marcel P. Timmer. "The next generation of the Penn World Table." *American Economic Review,* 105 no. 10 (2015): 3150-82.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. "Human capital and regional development." *The Quarterly Journal of Economics* 128, no.1 (2013): 105-164.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Sally Sadoff, Xiangdong Qin, and Yang Xu. *Measuring success in education: the role of effort on the test itself*. No. w24004. National Bureau of Economic Research, 2017.

Gove, Amber. *Early Grade Reading Assessment Toolkit.* RTI International, USAID and the World Bank, 2009.

Hall, Robert E. and Charles I. Jones. "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" *Quarterly Journal of Economics*, *114*, no.1 (1999): 83-116.

Hanushek, Eric A. and Dennis D. Kimko. "Schooling, Labor-force Quality, and the Growth of Nations." *American Economic Review* 90, no.5 (2000): 1184-1208.

Hanushek, Eric A. and Ludger Woessmann. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46.3 (2008): 607-68.

Hanushek, Eric A., and Ludger Woessmann. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17, no.4 (2012a): 267-321.

Hanushek, Eric A., and Lei Zhang. "Quality-consistent estimates of international schooling and skill gradients." *Journal of Human Capital* 3, no. 2 (2009): 107-143.

Hanushek, Eric A., and Ludger Woessmann. "Schooling, educational achievement, and the Latin American growth puzzle." *Journal of Development Economics* 99, no. 2 (2012b): 497-512.

Hendricks, Lutz. "How Important is Human Capital for Development? Evidence from Immigrant Earnings." *American Economic Review 92*, no.1 (2002), 198-219.

Hendricks, Lutz. and Todd Schoellman. "Human Capital and Development Accounting: New Evidence from Wage Gains at Migration." *Quarterly Journal of Economics*, *133* no. 2 (2017): 665-700.

Holland, Paul W. and Neil J. Dorans. "Linking and Equating." *Educational Measurement* 4 (2006): 187-220.

Kolen, Michael J., and Robert L. Brennan. *Nonequivalent groups: Linear methods. Test equating, scaling, and linking*. (2014): 103-142.

Patrinos, Harry Anthony, and Noam Angrist. *Global Dataset on Education Quality: A Review and Update (2000–2017)*. The World Bank, 2018.

Psacharopoulos, George, and Harry Anthony Patrinos. "Returns to investment in education: a further update." *Education Economics* 12, no. 2 (2004): 111-134.

Jones, Benjamin F. "The Human Capital Stock: A Generalized Approach." *American Economic Review* 104, no.11 (2014): 3752-77.

Jones, Benjamin F. "The Human Capital Stock: A Generalized Approach: Reply." *American Economic Review* 109, no. 3 (2019): 1175-95.

Klenow, Peter J., and Andres Rodriquez-Clare. *The neoclassical revival in growth economics: has it gone too far?*" NBER Macroeconomics Annual, MIT Press, Cambridge, MA (1997): 83–103.

Kim, Jim Yong. "The Human Capital Gap: Getting Governments to Invest in People." *Foreign Affairs* 97 (2018): 92

Kim, Jim Yong. "Eliminating Poverty in the 21st Century: The Role of Health and Human Capital." *JAMA* 320, no.14 (2018): 1427-1428.

Kraay, Aart. "The World Bank Human Capital Index: A Guide." *The World Bank Research Observer* 34, no. 1 (2019): 1-33

Krueger, Alan B. and Mikael Lindahl. "Education for Growth: Why and For Whom?" *Journal of Economic Literature* 39, no.4 (2001): 1101-1136.

Lange, Glenn-Marie, Quentin Wodon and Kevin Carey. *The Changing Wealth of Nations 2018: Building a Sustainable Future.* The World Bank, 2018.

Lee, Jong-Wha, and Hanol Lee. "Human Capital in the Long Run." *Journal of Development Economics* 122 (2016): 147-169.

Lucas, Robert E. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22, no.1 (1988): 3-42

Mankiw, N. Gregory, David Romer and David N. Weil. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics* 107, no. 2 (1992): 407-437.

Mincer, Jacob. "Human Capital and Economic Growth." *Economics of Education Review* 3, no.3 (1984): 195-205.

Mislevy, Robert J., Albert E. Beaton, Bruce Kaplan and Kathleen M. Sheehan. "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses." *Journal of Educational Measurement* 29, no.2 (1992): 133-161.

Muralidharan, Karthik, and Nishith Prakash. “Cycling to School: Increasing Secondary School Enrollment for Girls in India.” *American Economic Journal: Applied Economics*, 9, no. 3 (2017): 321-50.

Nelson, Richard R. and Edmund S. Phelps. “Investment in Humans, Technological Diffusion, and Economic Growth.” *American Economic Review* 56, no.1/2 (1966): 69-75.

OECD. *PISA 2015 Technical Report.* OECD Publishing, 2015.

OECD. *PISA 2009 Results: Overcoming Social Background: Equity in Learning Opportunities and Outcomes (Volume II).* OECD Publishing. 2010.

Psacharopoulos, George. “Returns to Investment in Education: A Global Update.” *World Development*, *22*, no.9 (1994): 1325-1343.

Pritchett, Lant. “Does Learning to Add Up Add Up? The Returns to Schooling in Aggregate Data.” *Handbook of the Economics of Education* 1 (2006): 635-695.

Pritchett, Lant. *The rebirth of education: Schooling ain't learning*. CGD Books, 2013.

Romer, Paul M. “Endogenous Technological Change.” *Journal of Political Economy* 98, no.5, Part 2 (1990): S71-S102.

Sandefur, Justin. “Internationally Comparable Mathematics Scores for Fourteen African Countries.” *Economics of Education Review* 62 (2018): 267-286.

Schoellman, Todd. “Education Quality and Development Accounting.” *The Review of Economic Studies* 79, no. 1 (2011): 388-417.

Smith, Adam. *An Inquiry into the Nature and Causes of the Wealth of Nations (Volume One)*. London: printed for W. Strahan; and T. Cadell, 1776.

Steinmann, Isa, Rolf Strietholt and Wilfried Bos. *Linking International Comparative Student Assessment.* LINCS Technical Report, 2014.

UNESCO. *More Than One-half of Children and Adolescents are not Learning Worldwide*. UIS Fact Sheet No. 46, 2017.

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

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

TABLE I

DATABASE CODEBOOK

Variable	Description
country	Country
cntabb	World Bank country code
test	Original assessment
year	Year in which the original assessment was taken
n_res	Dummy variable indicating if EGRA data is nationally representative
subject	Subject (math, reading or science)
level	Schooling level (primary, secondary)
score	Score from source assessment
se	Standard error from source assessment
score_m	Score from source assessment for males
se_m	Standard error from source assessment for males
score_f	Score from source assessment for females
se_f	Standard error from source assessment for females
HLO	Harmonized Learning Outcome (HLO)
HLO_se	Harmonized Learning Outcome (HLO) standard error
HLO_m	Harmonized Learning Outcome (HLO) for males
HLO_m_se	Harmonized Learning Outcome (HLO) standard error for males
HLO_f	Harmonized Learning Outcome (HLO) for females
HLO_f_se	Harmonized Learning Outcome (HLO) standard error for females
e_index	Exchange rate used to convert original scores to HLO scores

Notes: This table describes the key variables that will be publicly available in the online database.

TABLE II

COUNTRY-YEAR OBSERVATIONS BY DISAGGREGATION AND REGION

Region	Total	Male	Female	Reading	Math	Science	Primary	Secondary
East Asia & Pacific	336	332	332	102	118	116	103	233
Europe & Central Asia	954	948	948	301	328	325	261	693
Latin America & Caribbean	247	247	247	88	83	76	111	136
Middle East & North Africa	262	256	256	62	102	98	95	167
North America	56	56	56	18	19	19	17	39
South Asia	16	9	9	11	4	1	13	3
Sub-Saharan Africa	212	172	172	110	90	12	189	23
Total	2083	2020	2020	692	744	647	789	1294

Notes: This table presents coverage for country-year observations by region. The database includes 2083 observations across all countries from 2000-2017.

TABLE III

COUNTRY-SUBJECT-LEVEL OBSERVATIONS BY YEAR

Total	Total	Male	Female	Reading	Math	Science	Primary	Secondary
2000	192	157	157	74	75	43	63	129
2001	34	34	34	33	1	0	34	0
2002	4	4	4	2	2	0	4	0
2003	218	218	218	40	89	89	44	174
2004	2	2	2	1	1	0	2	0
2006	277	277	277	123	88	66	107	170
2007	193	193	193	15	96	82	97	96
2008	3	3	3	3	0	0	3	0
2009	225	225	225	79	73	73	6	219
2010	6	5	5	6	0	0	6	0
2011	239	239	239	55	92	92	151	88
2012	202	201	201	74	64	64	10	192
2013	77	54	54	35	27	15	77	0
2014	28	27	27	18	10	0	28	0
2015	324	324	324	75	126	123	98	226
2016	55	54	54	55	0	0	55	0
2017	4	3	3	4	0	0	4	0
Total	2083	2020	2020	692	744	647	789	1294

Notes: This table presents country-subject-level observations by year.

TABLE IV

RELATIONSHIP WITH ALTERNATIVE LEARNING HUMAN CAPITAL MEASURES

	Pearson Coefficient	p-value	Observations
Hanushek and Woessmann (2012)	0.925	<.001	50
Hanushek and Kimko (2000)	0.794	<.001	31
Item Response Theory	0.922	<.001	34

Notes: Comparisons are made using average scores across subjects and schooling levels.

Source: Hanushek and Woessmann (2012a) and Hanushek and Kimko (2000) data are from the respective papers. Item Response Theory data is from the LINCS project (Steinmann, Strietholt**,** and Bos 2014). We compare these data to the learning outcomes in our database.

TABLE V

HUMAN CAPITAL AND GROWTH – COMPARING MEASURES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Human Capital - Learning		1.610***		1.256**	1.164*	1.234**	1.095*
		(0.487)		(0.603)	(0.605)	(0.614)	(0.660)
Human Capital - Penn World Tables			1.292***			0.585	1.514**
			(0.466)			(0.579)	(0.691)
Human Capital - Schooling	0.272***			0.122	0.151		
	(0.100)			(0.122)	(0.122)		
Initial GDP per capita	-0.112***	-0.127***	-0.115***	-0.131***	-0.132***	-0.132***	-1.837***
	(0.0193)	(0.0207)	(0.0197)	(0.0210)	(0.0210)	(0.0211)	(0.383)
Observations	117	117	117	117	117	117	117
R-squared	0.232	0.254	0.234	0.260	0.234	0.260	0.174

Notes: Dependent variable: annual growth rates averaged across 2000-2010. Human Capital - Schooling refers to estimates in 2000, the beginning of the time period. Initial GDP per capita refers to levels at the beginning of the period in the year 2000. Human Capital – Penn World Tables refers to the measure of human capital in the Penn World Tables. Human Capital – Learning refers to the measure of human capital in this database from 2000 onwards. Results exclude countries in civil war, inflation crises and with rents from natural resources above 25 percent. Column (5) includes a measure of Human Capital – Schooling as the average from 2000-2010. Column (7) includes a measure of the log of initial GDP per capita.

Source: Schooling data are from Barro-Lee (2013). GDP and human capital data are from PWT 9.0. Learning estimates are from our database.

TABLE VI

HUMAN CAPITAL AND GROWTH FOR A GLOBAL DISTRIBUTION OF COUNTRIES

	Annual Growth Rates				*GDP Per Capita*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Human Capital - Learning	1.610***	1.256**	1.164*	1.131*	3.729***	3.547***	3.472***	4.871***
	(0.487)	(0.603)	(0.605)	(0.652)	(0.525)	(0.652)	(0.655)	(1.715)
Years of Schooling		Yes	Yes	Yes		Yes	Yes	Yes
Initial GDP Per Capita	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	117	117	117	117	117	117	117	117
R-squared	0.254	0.260	0.264	0.176	0.970	0.970	0.970	0.804

Notes: Dependent variable: annual growth rates and GDP per capita are averaged across 2000-2010. Years of Schooling refers to estimates in 2000, the beginning of the time period. Initial GDP per capita refers to levels at the beginning of the period (2000). These results exclude countries in civil war, inflation crises, and with rents from natural resources above 25 percent. Human capital measured by learning refers to average estimates from 2000 onwards. Column (3) and (7) refers to Years of Schooling as the average over the 2000-2010 time period. Columns (4) and (8) use the log of initial GDP per capita.

Source: Years of schooling are from Barro-Lee (2013). GDP is from PWT 9.0. Learning estimates are from our database.

TABLE VII

HUMAN CAPITAL AND GROWTH BY CURRENT AND INITIAL INCOME STATUS

	Current Income Status				*Initial Income Quartile*			
	Low Income	*Lower Middle Income*	*Upper Middle Income*	*High Income*	*Quartile 1*	*Quartile 2*	*Quartile 3*	*Quartile 4*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Human Capital - Learning	0.709	1.819	3.939***	1.270	2.001	4.002**	0.558	3.135**
	(1.592)	(1.435)	(0.901)	(0.903)	(1.323)	(1.487)	(0.908)	(1.464)
Years of Schooling	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Initial GDP per Capita	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	17	25	29	46	30	29	29	29
R-squared	0.608	0.154	0.635	0.418	0.151	0.373	0.389	0.467

Notes: Dependent variable: annual growth rates are averaged across 2000-2010. Years of Schooling refers to estimates in 2000. Initial GDP per capita refers to levels at the beginning of the period in the year 2000. These results exclude countries in civil war, inflation crises, and with rents from natural resources above 25 percent. Human capital measured by learning refers to average estimates from our database from 2000 onwards.

Source: Years of schooling data are from Barro-Lee (2013). GDP data are from PWT 9.0. Learning data is from our database.

TABLE VIII

BASELINE ACCOUNTING RESULTS AND COMPARISON TO QUANTITY LITERATURE

	This Paper				Literature		
	$w = 0$	$w = .15$	$w = .20$	$w = .25$	Hall and C. Jones (1999)	Hendricks (2002)	Schoellman (2011)
h_{90}/h_{10}	2.24	3.30	3.71	4.20	2.00	2.10	4.70
$\frac{h_{90}/h_{10}}{y_{90}/y_{10}}$	0.11	0.16	0.18	0.21	0.09	0.22	0.21
$\frac{var[log(h)]}{var[log(y)]}$	0.08	0.15	0.18	0.22	0.06	0.07	0.26

Notes: Dependent variable: real output per worker (2000-2010). Schooling and learning estimates are averaged (2000-2010). We convert HLO units into standard deviations based on a cross-country standard deviation of 68 to correspond to the returns to schooling parameter value which is given per standard deviation.

Source: Schooling data are from Barro-Lee (2013). GDP data are from PWT 9.0. Learning estimates are from our database. Literature estimates are derived from the referenced papers.

TABLE IX

BASELINE ACCOUNTING RESULTS AND COMPARISON TO QUALITY LITERATURE

	This Paper				Recent Literature			
	$w = 0$	$w = .15$	$w = .20$	$w = .25$	B. Jones (2014)	Hanushek and Woessmann (2012)	Hendricks and Shoellman (2017)	Caselli and Ciccone (2018)
$\frac{\ln(h_{90})-\ln(h_{10})}{\ln(y_{90})-\ln(y_{10})}$	0.27	0.39	0.43	0.48	*Nearly All*	0.51	0.62	*Potentially None*

Notes: Dependent variable: real output per worker (2000-2010). Schooling and learning estimates are averaged (2000-2010). We convert HLO units into standard deviations based on a cross-country standard deviation of 68 to correspond to the returns to schooling parameter value which is given per standard deviation.

Source: Schooling data are from Barro-Lee (2013). GDP data are from PWT 9.0. Learning estimates are from our database. Literature estimates are derived from the referenced papers. For the Hanushek and Woessmann (2012b) estimate, we use the decomposition comparing Latin American to non-Latin American countries where the *w* parameter value is equal to .20.

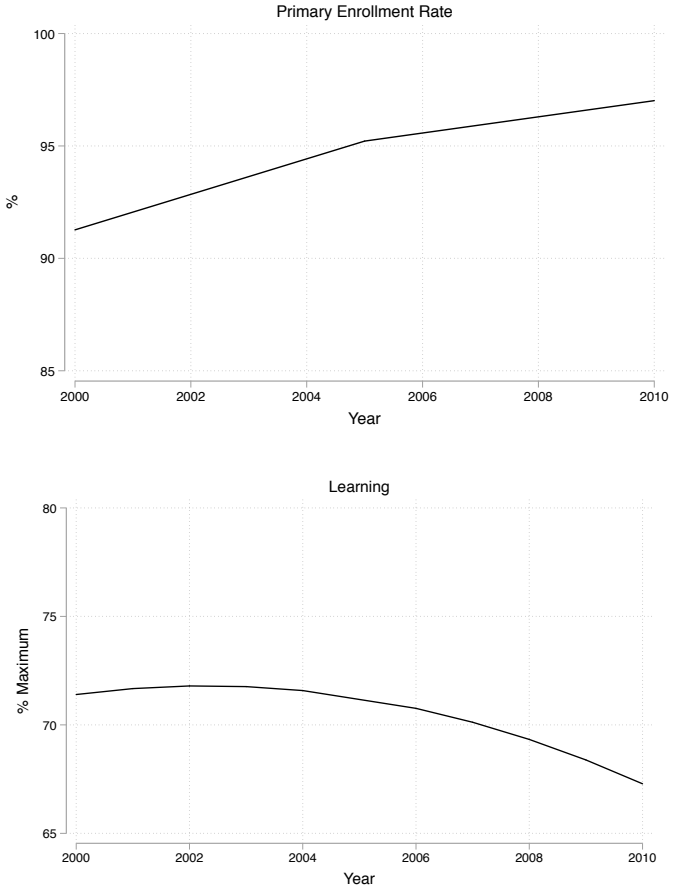


FIGURE I

Enrollment versus Learning (2000-2010)

Notes: Primary enrollment rates are averaged for the cohorts 25 years and older. Both enrollment and learning are averaged across countries. Percent maximum is calculated by dividing scores by the score at the upper end of the scale (625) in line with Kraay (2019).

Source: Primary enrollment rates are from Lee and Lee (2016); Learning outcomes are from our database.

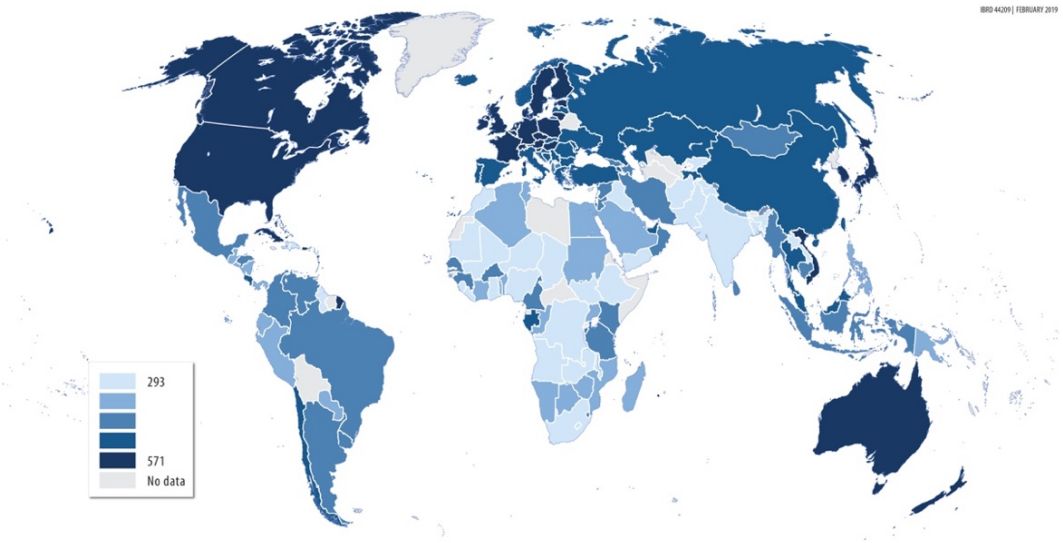


FIGURE II

Average Human Capital Score (2000-2017)

Notes: Average human capital is calculated across subjects and levels over the given period of time.

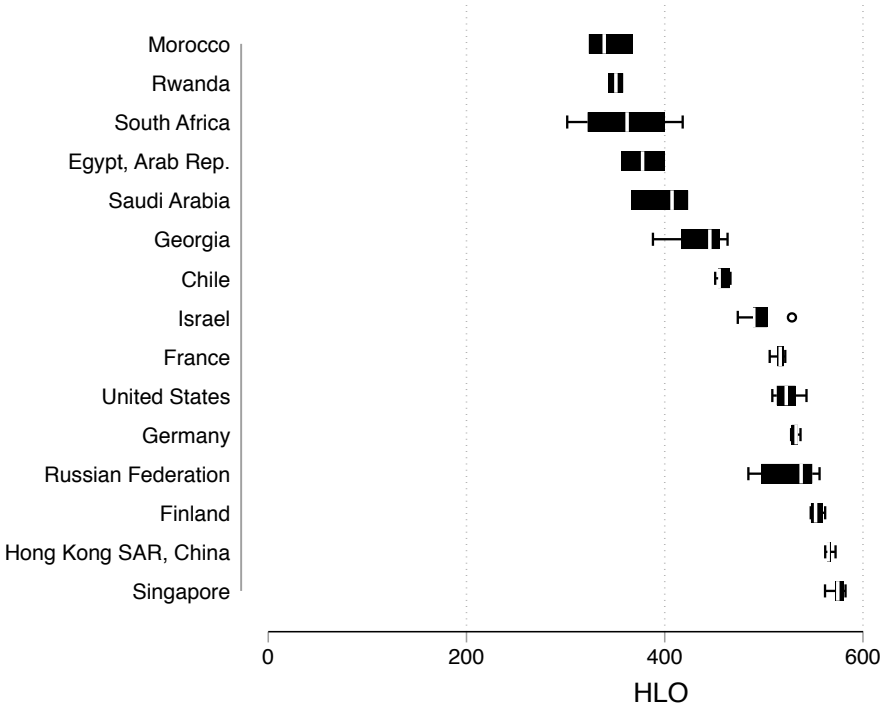


FIGURE III

Average Human Capital (2007-2017) – Selected Countries

Notes: Average human capital is calculated across subjects and schooling levels over the given period of time.

Source: Our learning outcomes database.

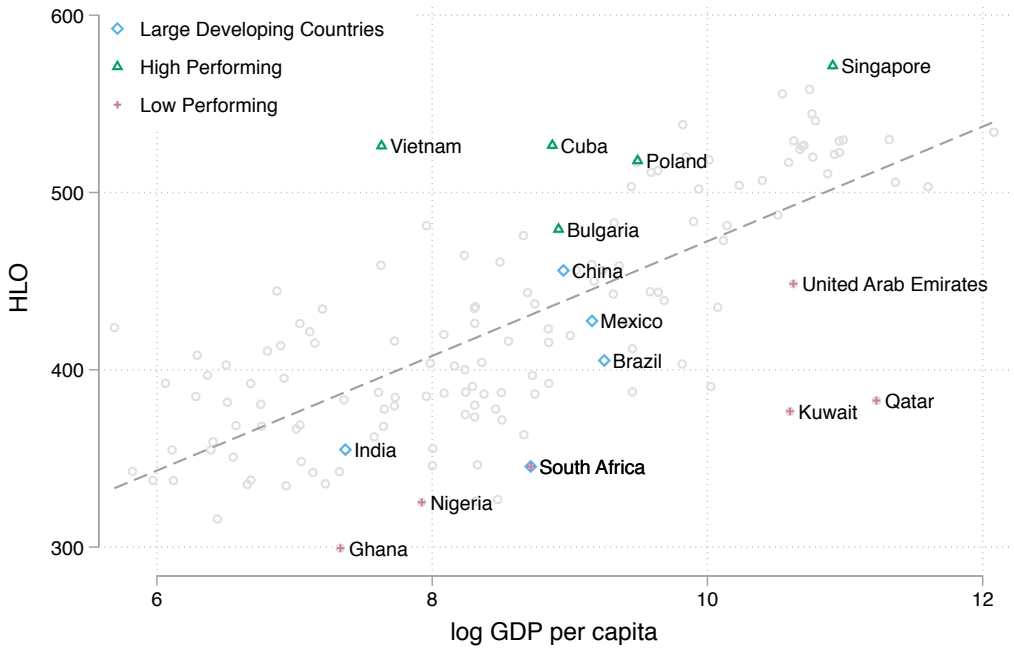


FIGURE IV

Human Capital (2000-2017) versus 2015 GDP per capita

Notes: Average human capital is calculated across subjects and schooling levels over the given time period.

Source: GDP per capita estimates are from World Bank national accounts data; learning outcomes are from our database.

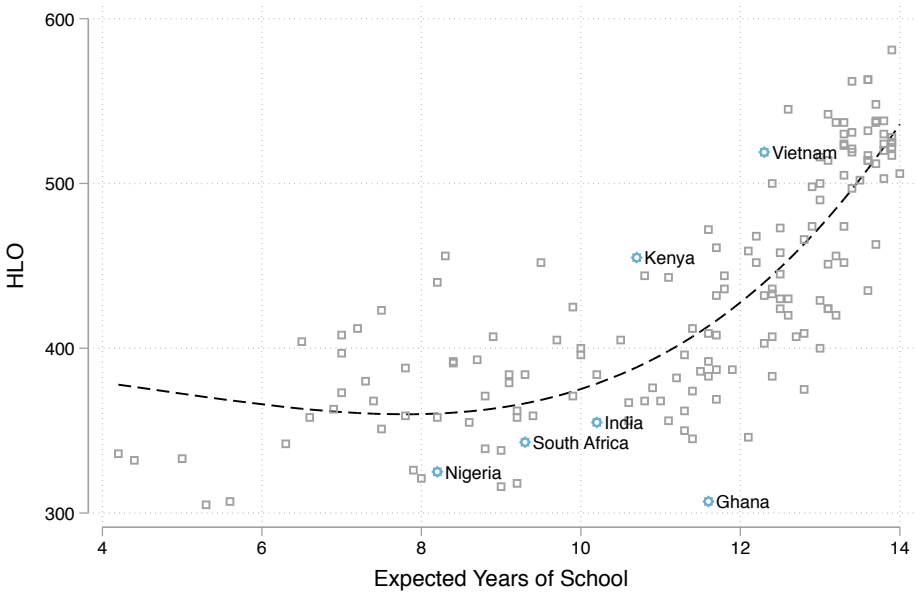


FIGURE V

Schooling versus Learning

Notes: We compare the average country learning in the latest available year from our database to expected years of schooling.

Source: Expected years of schooling are from the World Bank Human Capital Index data; average learning outcomes are from our database. Both measure take the latest data available up until 2018.

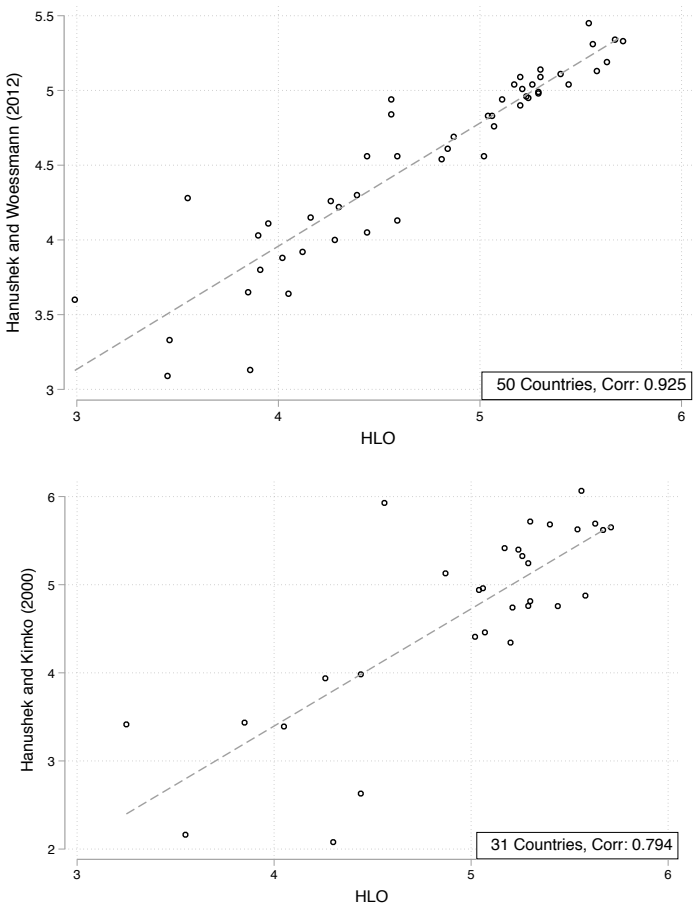


FIGURE VI

Comparison to Hanushek and Kimko (2000) and Hanushek and Woessmann (2012)

Notes: Hanushek and Kimko (2000) provide estimates of comparable learning measures of human capital. Hanushek and Woessmann (2012a) improve and expand these estimates. We compare our HLO measure to both estimates for the set of countries included in their growth regressions.

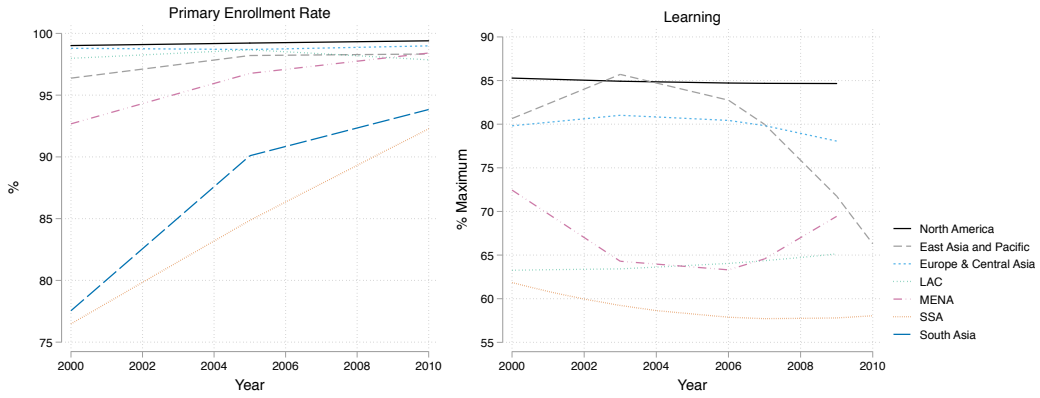


FIGURE VII

Enrollment versus Learning (2000-2010), by Region

Notes: Primary enrollment and learning estimates are averaged within regions. LAC refers to Latin American and the Caribbean; MENA refers to the Middle East and North Africa; and SSA refers to Sub-Saharan Africa. Percent maximum learning is calculated by dividing scores by the maximum score at the upper end of the score scale (625) in line with the methodology used in the Human Capital Index (Kraay 2019).

Source: Primary enrollment rates are from Lee and Lee (2016). Learning estimates are taken from our database.

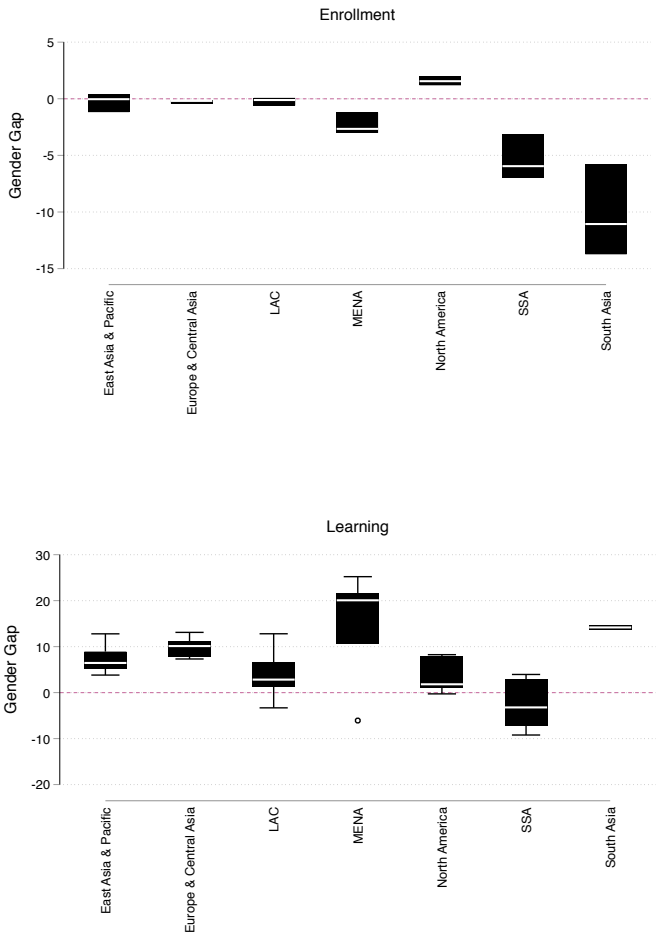


FIGURE VIII

Gender Gap – Enrollment versus Learning (2000-2010), By Region

Notes: LAC refers to Latin American and the Caribbean; MENA refers to the Middle East and North Africa; and SSA refers to Sub-Saharan Africa. The gender gap is calculated by taking the difference of female and male enrollment or learning. A positive gender gap indicates females do better and vice-versa.

Source: Primary enrollment rates are from Lee and Lee (2016). Learning estimates are taken from our database.

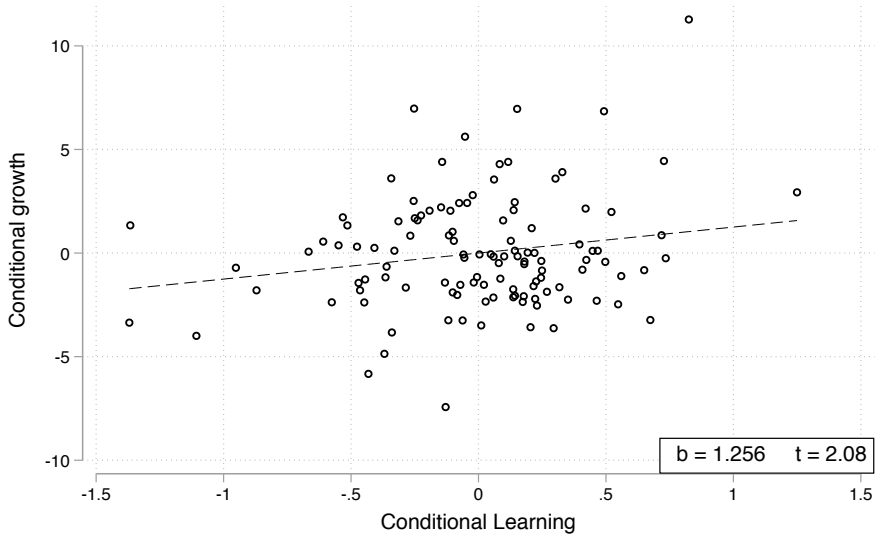


FIGURE IX

Learning and Growth (2000-2010)

Notes: 117 countries are included in the added-variable plot above. This added variable plot is derived from a regression of the average annual rate of growth of real GDP per capita from 2000-2010 on learning outcomes from 2000 onwards conditional on years of schooling in 2000 and initial GDP in 2000.

Source: Years of schooling data are from Barro-Lee (2013). GDP data are from PWT 9.0. Learning data is from our database.

Title: A New Micro Measure for Education Interventions: Learning-Adjusted Years of Schooling (LAYS)

Author: Noam Angrist[1]

Affiliations:

[1]University of Oxford.

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

Acknowledgments: This work builds on a series of collaborative efforts with colleagues Deon Filmer, Dave Evans, Rachel Glennerster, Harry Patrinos, Halsey Rogers, and Shwetlena Sabarwal to produce global public goods in education, including globally comparable test scores (*16*), global databases of education (*20*, *22*), prior work developing the macro Learning-Adjusted Years of Schooling measure (*7*) and policy applications for the World Bank and the Department for International Development (DFID). I am first author and would like to add the mentioned collaborators as coauthors in a future draft for publication.

Abstract: Efforts to improve educational outcomes in low- and middle-income countries abound, but comparing results presents challenges. Access and learning gains are reported separately; standard deviations, the current standard reporting metric, are not an intuitive measure for non-technical audiences. We apply a new measure, Learning-Adjusted Years of Schooling (LAYS), to combine quantity and quality of schooling into a comparable metric of educational progress that uses a single common standard. This paper presents results from hundreds of impact evaluations across 44 countries in LAYS units. The most effective interventions have large effect sizes on the order of global learning gaps between low- and high-performing education systems. LAYS could be a common unit of measurement for education outcomes globally, analogous to Quality-Adjusted Life Years for health.

One-Sentence Summary: Translating education impact evaluation results into a measure that combines access and learning gains shows that effective interventions can dramatically reduce global education inequality.

Main Text:

The typical proxy for education is quantity-based: the number of years of schooling that have been completed, whether by an individual or on average in a country. For many research and policy purposes, however, an education measure that combines quantity and quality of schooling may be preferable. The primary reason for preferring such a combined measure is that the same quantity of schooling can lead to very different levels of learning (*1*). Nor can quantity be effectively proxied by expenditure per head: children in Vietnam, for example, learn twice as much in a year of schooling as those in Peru despite similar levels of expenditure (*2*). There is some evidence that many of the benefits of education, including economic growth, are more closely associated with learning than with years of school (*3*). Therefore, standard education measures based only on years of schooling may be misleading as they miss the crucial dimension of the quality of that schooling. At the same time, access to schooling—even low-quality schooling—has important positive impacts, reducing teenage pregnancy, new HIV infections, and increasing labor market returns (*4–6*). Both schooling and learning are therefore essential measures of education and human capital.

Learning-Adjusted Years of Schooling (LAYS) combine the results of interventions that measure quantity of schooling and those that measure quality into a single easy-to-understand metric (*7*). Moreover, the combination of quantity and quality of schooling gives policy makers incentives to encourage participation of all children in high-quality schooling. A measure of years of schooling alone might crowd out a focus on learning, and a measure of only learning might create incentives to restrict school access to the highest performers to inflate average test scores.

While the methodology for LAYS was developed for summarizing and comparing the aggregate state of education at the country level—a use we refer to here as "macro LAYS"—this approach has not yet been applied to comparing results from specific interventions and policies, a use we refer to here as "micro LAYS." Converting the results of interventions and policies into a comparable measure confers several benefits. First, LAYS units are easy to interpret. Compare them with the standard unit now used to compare learning results from education interventions, which is the "change in standard-deviations units." This technical measure is understandable mainly to education experts, whereas the concept of gaining an additional learning-adjusted (or high-quality) year of schooling is clear, intuitive, economically meaningful and compelling to policy makers and other stakeholders. Second, by converting education intervention results into LAYS, we are seeking to compare results using a common measure expressed as a single learning scale. For example, if students learned to add single digits as a result of an intervention, that would generate the same increase in LAYS in Tanzania and Pakistan. LAYS can also be used to compare studies that have had an effect on participation (e.g., attendance or enrollment) with those whose effects are in terms of learning. Finally, by converting impact evaluation estimates into LAYS, we provide a clear lever for countries to improve their macro LAYS at the country-level if they were to scale a given type of intervention.

In this paper, we outline a methodology to produce micro LAYS, map them to macro LAYS, and convert results from hundreds of impact evaluations across 44 countries into LAYS units, focusing on a subset of studies with high quality cost-effectiveness data. Results reveal that

while many interventions are not effective, the most effective interventions have impacts that are large in comparison to the gap between low- and high-performing education systems globally.

This analysis can be thought of as an education analogue to units used in health, such as the Quality-Adjusted Life Years (QALY) and Disability-Adjusted Life Years (DALY). QALYs were first introduced in the 1970s and early 1980s (*8–10*), and today are accepted widely as the reference standard in cost-effectiveness analysis (*11*, *12*). Although only recently introduced, LAYS has already been incorporated into several large-scale policy efforts to improve education. It is a core component of the World Bank's recently launched Human Capital Index (*13*), it has been used internally by the United Kingdom's Department for International Development (DFID) to prioritize interventions, and it features in the United States Agency for International Development (USAID) Country Roadmaps to Self-Reliance. The goal of this paper is to demonstrate the promise of the LAYS measure, lay out its key analytical underpinnings and assumptions, generate momentum to continually improve the measure over time, and hopefully use this simple yet comprehensive measure to guide the building of high-quality education systems globally.

WHAT ARE LEARNING-ADJUSTED YEARS OF SCHOOLING?

LAYS at the macro level

Learning-adjusted years of schooling (LAYS) for a given country are the product of years of schooling and a measure of schooling quality. Specifically, they are produced by multiplying the country's average schooling by its test-score performance relative to a global high-performance benchmark (*7*). Figure 1 shows results using data from the World Bank Human Capital Index (*14*). For example, Singapore's average student test scores are closer than any other country's to the high-performance benchmark. As a result, its 14 average years of schooling become 13 learning-adjusted years of schooling. In contrast, South Africa has 10 years of schooling and only about 5 LAYS, because its test scores are only about half of Singapore's.

The high-performance benchmark used in the World Bank Human Capital Index is an artificial benchmark of high performance of 625 as defined by the international assessment Trends in Mathematics and Science Study (TIMSS), which was chosen because that benchmark is stable over time and is apolitical (*14*). 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. Alternatively, we could use the average of a basket of top-performing countries rather than just one. LAYS at the macro level are the product of the years of schooling of one cohort of the population and the learning for that cohort, relative to learning in a benchmark population. [Technical details are in the supplementary materials (SM).]

LAYS can be produced using any data on test scores and years of schooling. For example, the World Development Report 2018 uses test score data from TIMSS 2015, which covers 57 countries, and average years of schooling data from the Barro and Lee education

database (*15*). The version of LAYS embedded in the World Bank Human Capital Index uses different inputs that have wider global coverage. For test scores, this includes a database of globally comparable learning outcomes called Harmonized Learning Outcomes (HLO) that covers 164 countries and territories (*16*, *17*). For years of schooling, a measure of expected years of schooling is derived from UNESCO estimates (*14*). As these examples show, macro LAYS is a versatile unit of measurement, and can be used with a variety of data inputs depending on the distribution of countries covered and other relevant data features.

In summary, 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 LAYS provide a unified and user-friendly measure for a variety of education outcomes.

LAYS at the micro level

The number of rigorous studies showing improvement in educational outcomes is growing, with nearly 300 impact evaluations focused on learning outcomes by 2016 in developing countries (*1*). A common metric by which to compare them is needed to enable evidence synthesis and clear policy recommendations. Moreover, the current methods of reporting the results, such as in standard deviation improvements of access or learning, present several challenges: (1) standard deviations are not an intuitive outcome measure for non-technical audiences; (2) standard deviations measure outcomes relative to local variation rather than against an absolute level of learning; (3) access and learning are separated, forcing policy makers to consider each challenge separately; and (4) individual study results do not translate into how the size of the impact of interventions compares to the size of educational gaps across countries.

We address these challenges with a methodology that expresses education outcomes from interventions and policies in terms of LAYS units, combining the results of interventions that measure quantity of schooling and those that measure quality into a single, internationally comparable, and policy-salient metric. We refer to LAYS gained from an intervention or policy as micro LAYS. If micro studies tested students, and reported results, against internationally agreed test scores such as TIMMS, PISA or EGRA, the translation into LAYS would be straightforward. Our hope is that this will increasingly become the norm in education studies. However, this is currently not the norm, and therefore a number of assumptions are needed to translate existing studies into LAYS. In this section, we outline the approach to producing micro LAYS for studies that report effects on schooling participation, such as attendance or years of school gained, as well as studies which report effects on learning outcomes.

When studies report effects on schooling participation, micro LAYS are the product of the learning gains as a result of the intervention, multiplied by schooling quality in the country where the intervention took place relative to a global benchmark of high performance. We then multiply these gains by the duration the effects of the intervention are expected to persist. For example, 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, then we would simply multiply the additional year of high-quality schooling by four. Thus, the intervention produced four years of high-quality schooling. The decision of the time over which the intervention is expected to be effective has received much attention in the public health literature in the context of QALYs and DALYs. In our analysis, we explore various options, as discussed in the supplementary materials.

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 .25 standard deviations per year as a result of an intervention in South Africa, and students typically learn .25 standard deviations per year in a given year in South Africa, students will have learned a year's worth of South African schooling. Second, we then apply a global quality-adjustment factor to derive learning-adjusted years of schooling (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 multiply by the period of time over which effects are expected to last. As an example, if students had fallen behind grade level and an intervention enables students to catch up to grade level, they might now benefit from day-to-day schooling for the remaining school life expectancy. A detailed methodology and set of assumptions is provided in the supplementary materials. We find that micro LAYS are robust to a range of sensitivity and robustness tests, also outlined in the supplementary materials.

In summary, both participation- and learning-based LAYS yield an interpretation analogous to macro LAYS: they tell us that a given intervention in a given country produces a certain number of years' worth of globally benchmarked high-quality learning. Thanks to this common unit, the impacts of studies that measure these two different types of outcomes can be directly compared, as we will illustrate below. One challenge in assembling these micro LAYS estimates is how to handle a study that reports impacts on both participation and learning. One option would be to combine the two, but there are trade-offs involved in treating participation and quality impacts as complements rather than substitutes. If we sum the estimates, we will double- count in cases where gains in learning were a direct, even partial result of 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 both estimates, we could use only estimates from either participation or learning. However, under this approach we are assuming that one is the central output, and that the other outcome dimension is largely captured within that central output. For the purposes of this paper, we take the estimate that is highest between schooling or learning in each case. This approach lets the data speak, placing *a priori* equal weight on schooling and learning, introduces no new assumptions, and avoids double counting.

Applying LAYS to over 100 education studies

To show concretely how the micro LAYS measure allows comparisons of impacts, in this section we apply it to a sample of over a hundred of studies exploring various types of education interventions covering 44 countries. We further examine a subset of studies with comparable cost data, to analyze cost-effectiveness. The subset of cost-effectiveness studies also provides sufficient granularity to explore specific mechanisms which we relate to insights generated from the global database. We examine how many LAYS each policy or intervention delivers; how cost-effective those gains are; and how much of the cross-country education gap (expressed in LAYS) that intervention would close if it were scaled up, assuming that the effectiveness of the intervention would be maintained. (This assumption is not trivial, and we carry out this calculation as a calibration exercise rather than a simulation exercise. Effectiveness at large scale is often substantially lower than effectiveness at the scale of a pilot study.) We start with over 300 studies compiled from a database produced by Evans and Yuan (*18*), which draws studies from a wide range of reviews (*19–25*). We include three additional recent studies (*26–28*) with rigorous evaluation methodologies as well as high quality impact and cost data.

Our inclusion criteria are that the study should: (1) provide impact estimates in terms of standard deviations and (b) be based on a credible causal inference strategy, either using randomized controlled trials or quasi-experimental methods, such as differences-in-differences, instrumental variables, regression discontinuity, fixed effects, or propensity score matching. (We ensure all outcomes are coded to be directionally positive to aggregate across outcomes. For example, a reduction in absenteeism is coded as an increase in attendance.) The list of studies included in our analysis is illustrative rather than exhaustive, and in the future, we aim to include more studies and build a comprehensive database of education interventions. Figure 2 below summarizes the country coverage of the database and number of impact evaluations per country. In total, after accounting for our inclusion criteria, we analyze data from over 150 impact estimates across 44 developing countries.

This methodology has uses beyond comparison of effects of impact evaluations; LAYS is a unit of measurement and therefore any result, including descriptive and observational results, can be expressed in terms of LAYS. Our emphasis in this paper is to show how LAYS allows us to compare impact-evaluation impacts.

In this illustrative analysis, we make several choices for parameters and data inputs. First, we assume that the intervention's effects last only for the duration of the evaluation, since this is a known quantity and requires no further assumptions. In Figure S1 we explore the alternative assumption that the impacts last just one year. A second choice that we make is to set the high-quality benchmark learning rate equal to 0.8 standard deviations per year, a conservative estimate for high rates of learning based on the literature, standard benchmarks, and policy-relevant changes, explained in detail in the supplementary materials. We choose an artificial benchmark in this analysis, which has the advantage of being stable over time, unlike the actual learning rates of high performers, such as Finland or Singapore, and of being apolitical. This approach to defining high-quality learning *rates* is similar to the one used to define the high-performance benchmark learning *level* in the World Bank Human Capital Index, which sets the benchmark at 625 on the scale of TIMSS and the Programme for International Student

Assessment (PISA). In the supplementary materials, we explore four plausible approaches to validate this 0.8-standard-deviation high-performance benchmark.

We calculate how much to adjust improvements in access for the level of learning (i.e., the learning adjustment rate) using Harmonized Learning Outcomes (HLO), which are globally comparable measures of learning introduced by Angrist et al. (*16*) and Patrinos and Angrist (*17*). Alternative test score data could be chosen. We choose HLO data for various reasons. First, these data enable us to use the same learning scale for interventions from 164 countries across the world, a wide range of countries from which we also draw impact evaluation education estimates. Second, since these data are used in the World Bank Human Capital Index (HCI), this enables us to produce micro LAYS that map directly to the macro LAYS in the HCI. The international tests used to measure learning and 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 new 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 Figure S2 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 the supplementary materials, we further explore the sensitivity of results to the score scale.]

WHAT DOES LAYS TEACH US ABOUT IMPROVING EDUCATION?

Effectiveness

We start by analyzing the LAYS gained under each intervention, using our global database of impact evaluations. Figure 3 summarizes results by category of education intervention, such as Early Childhood Development (ECD) or instruction targeted to the child's level of learning. Intervention categories are based on original study designations, with a few adjustments. We recategorize technology interventions into either "targeted instruction" or "school resources and inputs" based on whether they involved adaptive software or were largely a hardware-based intervention. We also define teacher training interventions narrowly. Many interventions include training of teachers; for this analysis, we would categorize a program that provides materials to help teachers target instruction to the level of the child and also provides training to those teachers as targeted instruction. "Teacher training only" captures only general teacher training programs without other major elements. For interventions with multiple components, we selected the central component and used that. Later in the paper, we examine individual studies and so can characterize them more precisely.

The most effective interventions, in terms of median LAYS gained, are (in order): early childhood development (ECD) programs, school choice and public private partnerships (such as voucher schemes), structured pedagogy, targeted instruction (through technology or level-appropriate activities and materials, rather than instruction pitched at grade level competencies few children have attained), scholarships, targeted school construction (for example, in remote underserved areas), cash transfers, incentives (such as merit based pay), nutrition (such as school feeding), information on the returns to schooling or feedback on student performance,

management and accountability schemes (such as report cards and training for community members), teacher training (when the central and only input), health products (such as sanitary pads, deworming pills, malaria prevention products), schooling resources and inputs (such as technology hardware, uniforms, school grants), and reducing class size.

Two patterns emerge. First, some interventions (such as the most effective scholarship schemes) provide up to a full additional year of globally benchmarked high-quality schooling, while other interventions and policies (such as reductions in class size) yield no gains. Second, we observe some categories have low variance—as in the case of class-size reductions and school resources and inputs, which are tightly concentrated around zero—while other categories have high variance. An example of the high-variance group is targeted school construction: some of the impact estimates for this category are around zero, while others are at the highest end of the spectrum, with over half a high-quality year of schooling gained. This indicates that when considering interventions, we should consider not only the average effect but also the variance. For example, targeted instruction produces large gains with relatively low variation, whereas targeted school construction has a low average effect but high variation. This points to the importance of contextual relevance: some interventions have similar effects across contexts, while others work extremely well in one context but not in others.

Cost-Effectiveness

Comparative information on effectiveness will be most useful to policymakers when it incorporates information about cost. Therefore, in addition to the general global trends, in Figure 3 we further denote the subset of impact evaluations with cost-effectiveness data. By and large, the subset of interventions with cost-effectiveness data are not systematically biased towards high or low impacts, although within each category the studies with cost-effectiveness data may skew one way or the other. We next analyze the studies with cost-effectiveness data in more detail, with the understanding that these results should be interpreted alongside the broader set of studies. In this analysis, we also focus in on specific intervention designs, which allows more granular insights.

Figure 4 shows results for LAYS gained by intervention and country for the studies that include cost-effectiveness data. The top ten performers are: the Campaign for Female Education (Camfed) program in Tanzania—a program that provides scholarships for girls along with school materials and training for teachers and parents (1.1 LAYS); Tusome (Swahili for "Let's Read") in Kenya—a program that provides textbooks, teacher coaching, and teacher training (1.04 LAYS); building village-based schools in Afghanistan (.74 LAYS); computer-assisted learning (CAL) in India (.43 LAYS); vouchers for mind-spark adaptive learning software in India (.36 LAYS); merit scholarships for girls in Kenya (.34 LAYS); ability grouping using extra teachers in Kenya (.29 LAYS); textbooks for the top of the class in Kenya (.28 LAYS); providing information on the returns to schooling in Madagascar (.25 LAYS); school elections and village council engagement in Indonesia (.25 LAYS); and public-private partnerships in Liberia (.24 LAYS). We also see that about half of the interventions produce no significant effects.

These findings point to a few overall lessons. Among the most effective programs—in this illustrative sample of studies—are: multidimensional programs (both Camfed in Tanzania

and Tusome in Kenya) as well as programs that are carefully targeted to a local need, such as scholarships (for girls), information (when returns to schooling are unknown), and school construction in under-serviced remote areas; pedagogical instruction that is pitched to students' levels of learning, not based on a rote syllabus or an over-ambitious curriculum; and programs to facilitate competition and accountability in the schooling market and to build community support for learning. In addition to producing substantive insights about types of effective programs, these results underscore the importance of contextual relevance. For example, 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; similarly, a deworming program is unlikely to be cost-effective in a place with low levels of intestinal worms.

Figure 5 shows cost-effectiveness estimates for these interventions, expressed in LAYS per $100. When we take cost into account, several new interventions join the list of top performers: school links to village councils in Indonesia, remedial education in India, camera-based monitoring of teachers and incentives in India, and deworming in Kenya. By contrast, public-private partnerships, scholarship programs, and technology-assisted learning programs drop down the list because of their higher costs, although these programs continue to be cost-effective, with LAYS per $100 gained ranging between .24 to .36.

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

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

What could the macro, systemwide effects of these highly effective interventions be? To answer this, we explore first how many LAYS a given intervention would contribute toward closing the gap to achieve a full and globally benchmarked quality education at scale in a given country—assuming, as mentioned before, that the nationally scaled-up version of the program were as effective as the evaluated version. (This is not always the case (*29*). An alternative approach would be to apply a "discount rate" to interventions for this exercise mapping micro to macro LAYS.) In essence, we map micro LAYS onto macro LAYS. Figure 7 shows that certain programs, such as Tusome in Kenya, could close a quarter of the gap between Kenya's own years of schooling and learning-adjusted years of schooling, vaulting Kenya ahead of Mexico and the Philippines in globally benchmarked high-quality schooling. Even small absolute gains can make a meaningful difference. Consider the case of camera monitoring in India. The average expected years of schooling in India is 10.2, but because learning levels are low, India averages just 5.8 Learning-Adjusted Years of Schooling (LAYS), according to the World Bank Human

Capital Index. If the camera monitoring program were made universal with no loss of effectiveness, India would improve to around 6 LAYS, closing nearly 5% of its LAYS gap.

A second macro question is how effective each intervention could be at closing gaps at a given level of spending. Figure 8 takes cost-effectiveness into account, making the interventions comparable by showing the gap that each could close at a cost of $100 per child. It shows that highly cost-effective programs like Tusome and ability grouping could sharply narrow the learning gap separating the children in Kenya from their peers in high-quality systems.

CONCLUSION

In this paper, we propose a new measure—micro LAYS—to make the impact of education interventions more easily comparable and interpretable, and we use that measure to translate hundreds of education impact evaluations across 44 countries into micro LAYS units. Using micro LAYS to express impact sizes achieves three goals: (a) it places attainment and learning outcomes on a unified scale, allowing interventions to be compared directly; (b) it expresses educational outcomes using an easy-to-interpret measure that improves incentives for policy makers to promote both quantity and quality of schooling; and (c) it identifies levers for countries to close gaps between high and low-performing education systems globally. This analysis provides guidance into which interventions might have potential in a given context.

We convert results from hundreds of interventions into LAYS units. We find that while some interventions are not effective, the most effective interventions can close a substantial portion of the gap between high and low-preforming education systems globally. This analysis highlights what types of programs and policies are most effective and cost-effective from an illustrative sample and builds the foundation for the use of LAYS as a common metric for economic evaluation of education interventions. These results should be interpreted with context in mind: challenges should be identified locally, and the global evidence base should then be used to identify possible cost-effective solutions, which should then be carefully adapted solutions to the local context.

Through this first analysis, we provide the foundation for LAYS to become a unit of measurement for education outcomes globally. Over time, we plan to expand the set of impact evaluations included in the analysis, which will enable us to compile a global database of education outcomes expressed in terms of LAYS. The objective is to use common and comprehensive policy-relevant education measures to galvanize progress across education systems worldwide. We hope that the micro LAYS measure proves immediately useful for that purpose, but also that it continues to improve as data and measurement improve, much as QALYs and DALYs in the health sector have evolved over the past few decades.

References and Notes:

1. World Bank, “World Development Report 2018: Learning to Realize Education’s Promise” (Washington D.C., 2018).

2. A. Singh, Learning More with Every Year: School Year Productivity and International Learning Divergence. *Journal of the European Economic Association* (2019), doi:10.1093/jeea/jvz033.

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

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

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

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

7. D. P. Filmer, F. H. Rogers, N. Angrist, S. Sabarwal, "Learning-Adjusted Years of Schooling (LAYS) : Defining A New Macro Measure of Education" (WPS8591, The World Bank, 2018), pp. 1–61.

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

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

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

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

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

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

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

15. R. J. Barro, J. W. Lee, A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics*. **104**, 184–198 (2013).

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

17. H. A. Patrinos, N. Angrist, “Global Dataset on Education Quality : A Review and Update (2000-2017)” (WPS8592, The World Bank, 2018), pp. 1–37.

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

19. K. M. Conn, Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations. *Review of Educational Research*. **87**, 863–898 (2017).

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

21. P. Glewwe, E. A. Hanushek, S. Humpage, R. Ravina, in *Education Policy in Developing Countries* (University of Chicago Press, 2013; https://www.universitypressscholarship.com/view/10.7208/chicago/9780226078854.001.0001/upso-9780226078687-chapter-2).

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

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

24. P. J. McEwan, Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research*. **85**, 353–394 (2015).

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

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

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

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

29. M. A. Kraft, D. Blazar, D. Hogan, The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*. **88**, 547–588 (2018).

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

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

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

33. M. A. Kraft, Interpreting Effect Sizes of Education Interventions. *Educational Researcher* (forthcoming).

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

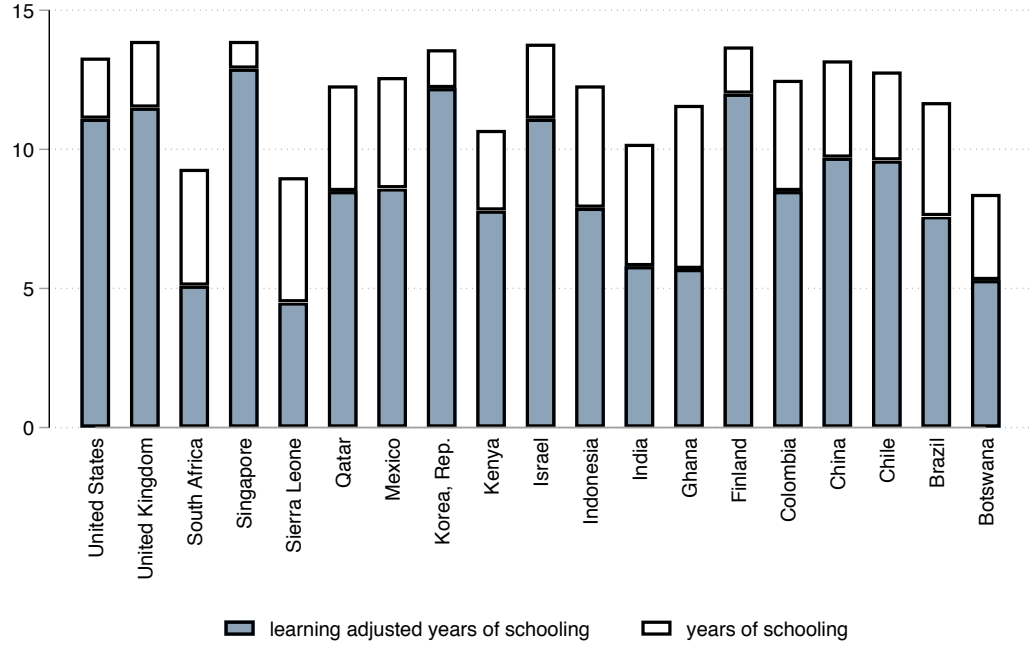
35. T. J. Kane, "The Impact of After-School Programs; Interpreting the Results of Four Recent Evaluations" (William T. Grant Foundation, 2004), (available at https://rhyclearinghouse.acf.hhs.gov/library/2004/impact-after-school-programs-interpreting-results-four-recent-evaluations).

Supplementary Materials:

Supplementary Text

Figures S1-S9

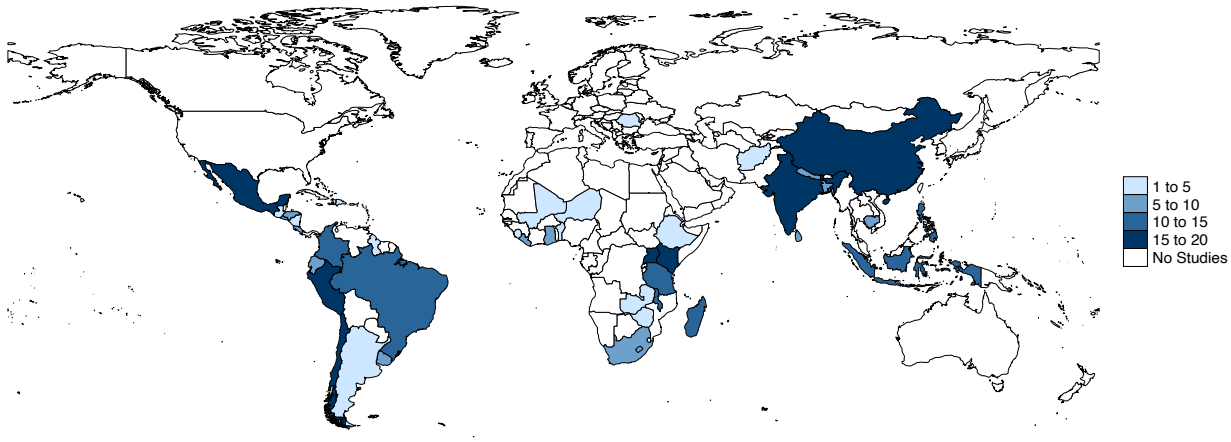
Figure 1: Years of Schooling and Learning-Adjusted Years of Schooling (Macro LAYS)



*Note: Schooling data is based on UNESCO expected years of schooling and learning data is based on Harmonized Learning Outcomes (HLO) (*14*).*

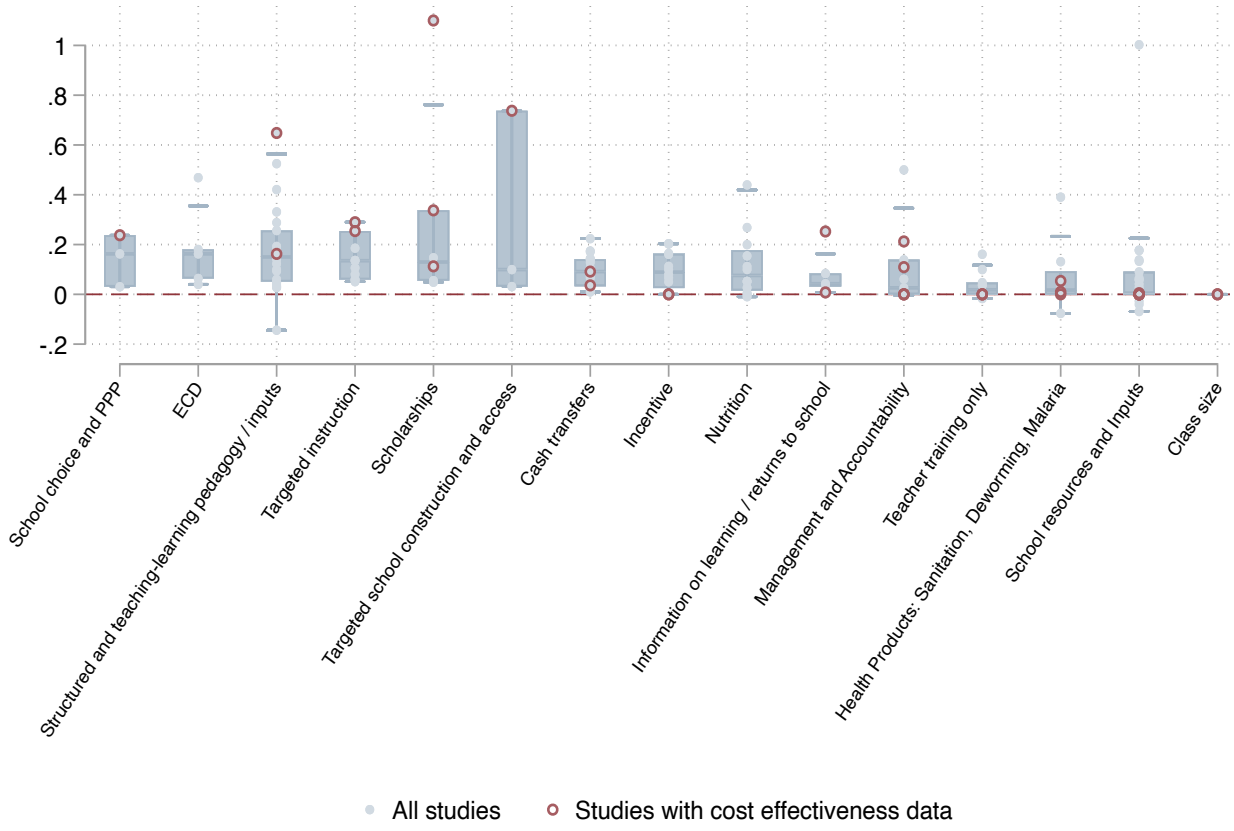
*Source: The Human Capital Index is described in Kraay(*14*) and is based on Angrist et al. (*16*) learning data and UIS enrollment data.*

Figure 2: Number of Impact Evaluations by Country



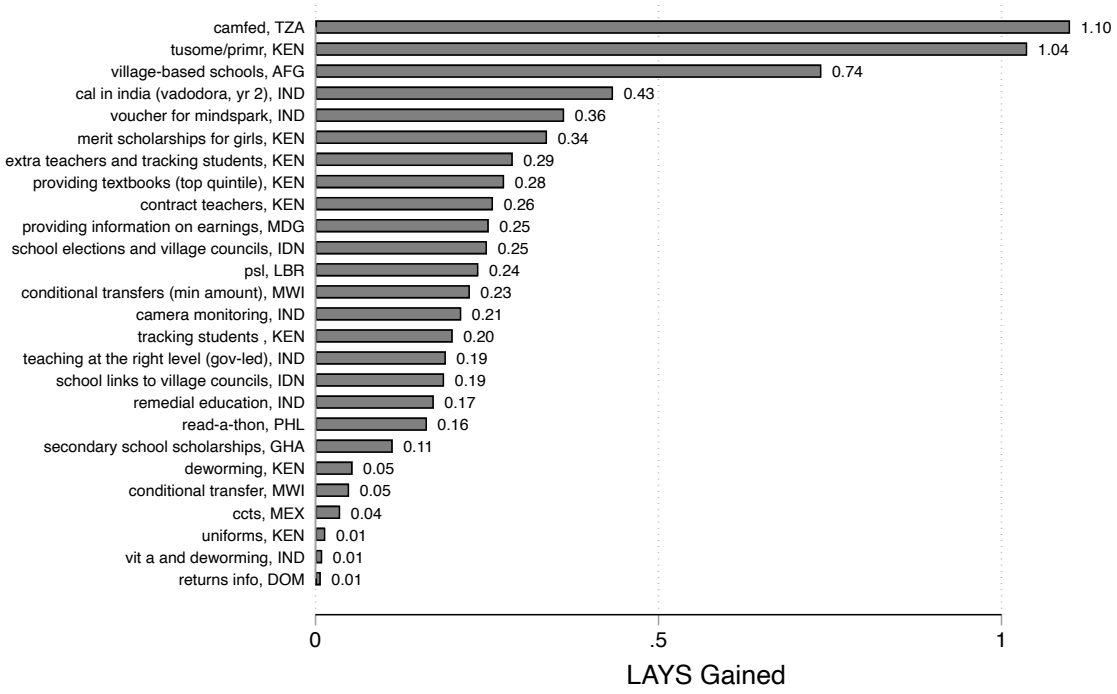
Notes: 44 developing countries have at least one impact evaluation included in this database.

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



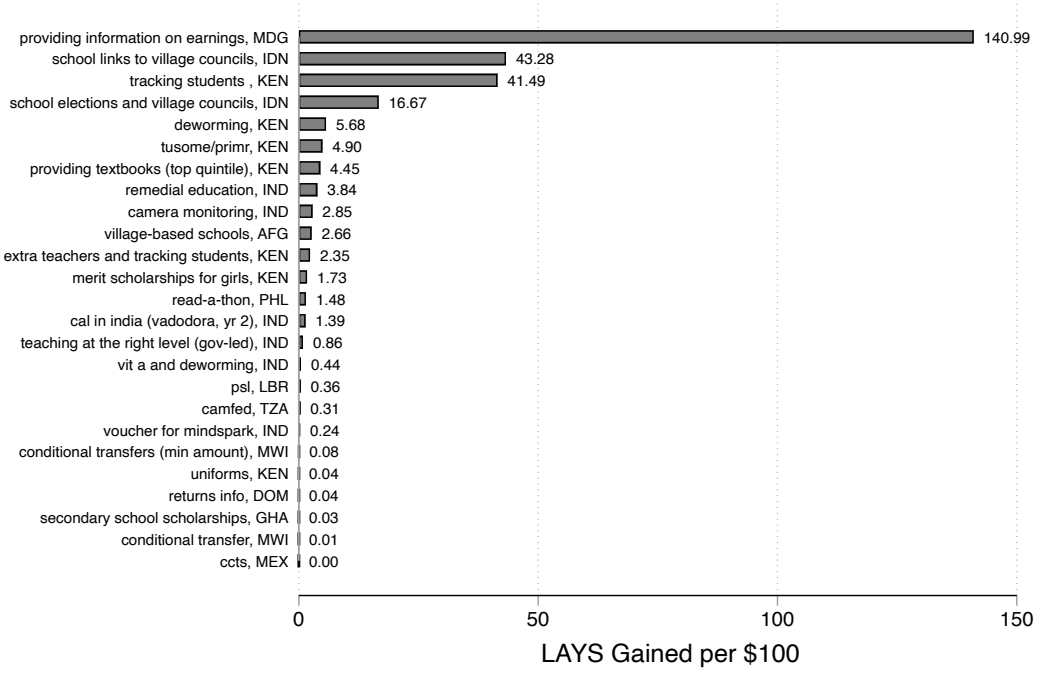
Notes: Each category of education intervention shows the learning-adjusted years of school (LAYS) gained from a given intervention or policy across over 150 impact estimates in 44 countries. The boxplot is ordered from largest to smallest median effects and the shaded boxplot describes the 25th and 75th percentile, with whiskers at upper and lower fences at a distance of 1.5 times the interquartile range above and below the nearest quartile.

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



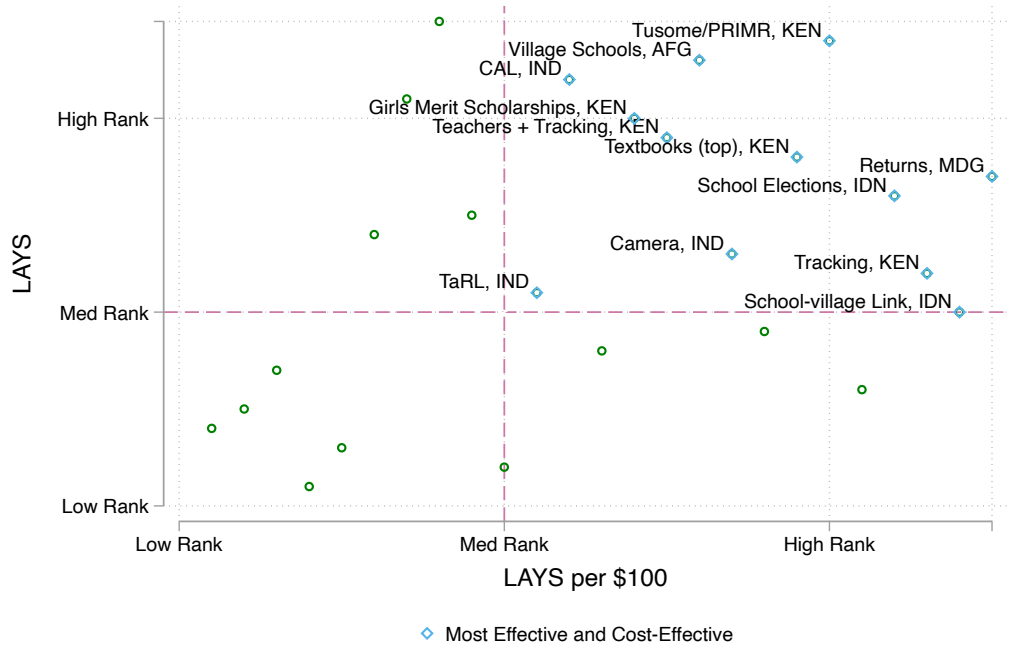
Notes: We do not include null impacts, which are not cost-effective by definition.

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



Notes: We omit interventions with a null effect.

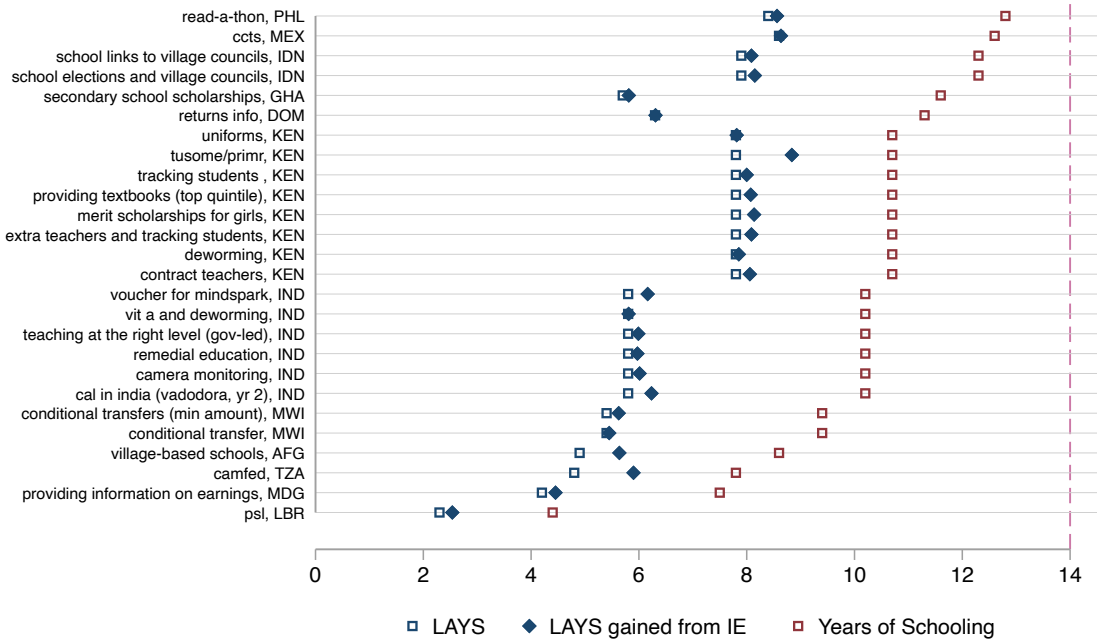
Figure 6: Effective and Cost-Effective Interventions



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

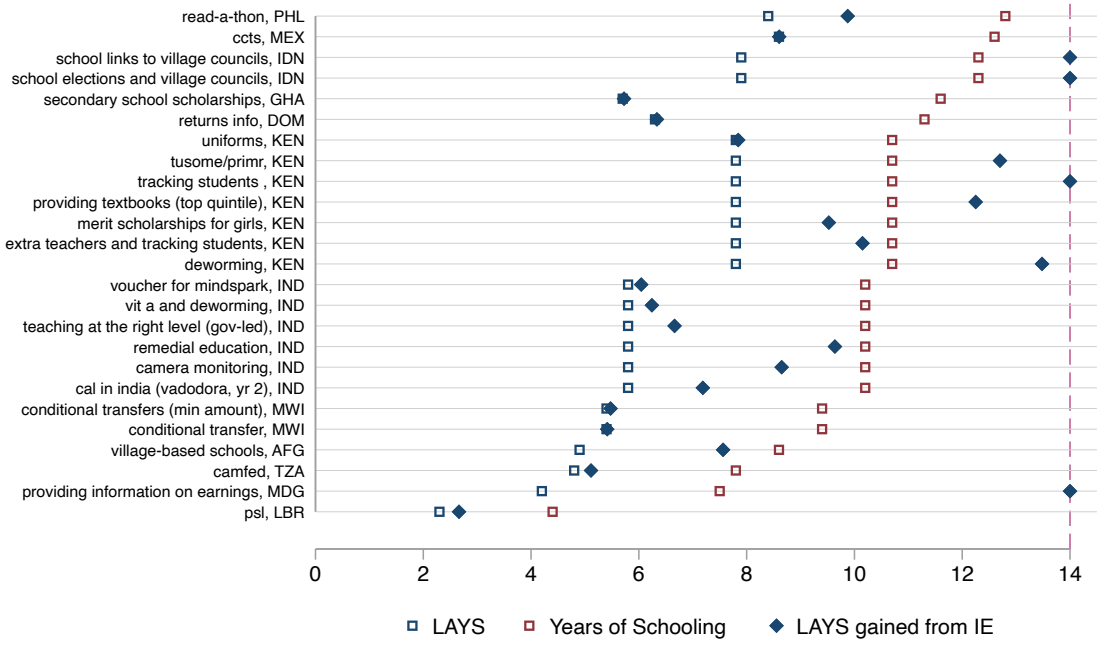
5

Figure 7: Closing the Gap to Reach Full and Quality Learning, by Country



Notes: 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 (HCI). 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.

Figure 8: Closing the Gap to Reach Full and Quality Learning per $100, by Country



Notes: 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.

Supplementary Materials for

A New Micro Measure for Education Interventions: Learning-Adjusted Years of Schooling (LAYS)

Supplementary Text

Macro LAYS

Technically, LAYS at the macro level can be defined as follows:[1]

$$LAYS_i = S_i \times L_i^h \quad (1)$$

where S_i is a measure of the average years of schooling acquired by a relevant cohort of the population of country *i*, and L_i^h is a measure of learning for a relevant cohort of students in country *i*, relative to a high-performance benchmark numeraire *h* such that:

$$\mathrm{L_i^h} = \frac{\mathrm{L_i}}{\mathrm{L_h}} \quad (2)$$

where L_i and L_h are the measures of average learning per year in country *i* and in numeraire (or benchmark) country *h*.

The assumptions invoked in the construction of macro LAYS are explored in depth in Filmer et al. *(7)*. Here, we highlight one important assumption: constant average learning trajectories, or the idea that students learn the same amount with each additional year of schooling. Figure S3 demonstrates the utility of this assumption using a hypothetical example. Assume that we observe Grade 8 test scores of 600 for Country A and 400 for Country B and that individuals in Country B average 9 years of schooling. LAYS allows 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. *(7)* explore this question with a series of empirical tests on whether learning trajectories are constant on a locally defined interval. Figure S4 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 *(30)*. 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.[2] In the case of a mathematical skill, division, Figure S4 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

[1] This exposition of the macro LAYS draws from Filmer et al. *(7)*.

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

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. *(7)* also yield results consistent with the linearity assumption (at least over a significant local interval).

Methodology for Construction of Micro LAYS

The number of rigorous studies showing improvement in educational outcomes is growing, with nearly 300 impact evaluations focused on learning outcomes by 2016 in developing countries *(1)*. If only a few dozen studies existed, we could review each individually. However, since the studies number in the hundreds and are accumulating rapidly, a common metric by which to compare them is needed to enable evidence synthesis and clear policy recommendations. Moreover, the current methods of reporting the results, such as in standard deviation improvements of access or learning, present several challenges: (1) access and learning are separated, forcing policy makers to consider each challenge separately; (2) standard deviations are not an intuitive outcome measure for non-technical audiences; and (3) individual study results do not translate into how the size of the impact of interventions compares to the size of educational gaps across countries.

We address these challenges with a methodology that expresses education outcomes from interventions and policies in terms of LAYS units, combining the quantity and quality of schooling into a single, internationally comparable, and policy-salient metric. We refer to LAYS gained from an intervention or policy as micro LAYS. To ensure a coherent unifying approach, the micro-LAYS methodology invokes assumptions similar to those used in constructing country-level macro LAYS estimates. In this section, we outline the approach to producing micro LAYS for studies that report effects on schooling participation, such as attendance or years of school gained, as well as studies which report effects on learning outcomes.

Micro LAYS using Participation Estimates

The construction of micro LAYS derived from impacts on schooling participation, denoted by subscript *p,* can be expressed as follows:

$$LAYS_p \;=\; \gamma_i * t * \; L_i^h$$

where L_i^h is a measure of learning for a cohort of students in country *i* relative to a high-performance learning benchmark *h*, such that $L_i^h = \frac{L_i}{L_h}$. This adjustment is analogous to the learning adjustment used for macro LAYS, except that instead of adjusting *total* schooling *S* in country *i* as we do for macro LAYS*,* we now adjust the years of school gained per year *as a result of the intervention* γ in country *i.* Because interventions differ in the duration of their impacts, we add a multiplicative factor *t* that represents how long intervention γ's effects are expected to persist. For example, 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 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, then we would simply multiply the additional year of high-quality schooling by four. Thus, the intervention produced four years of high-quality schooling.
The decision of the time *t* over which the intervention is expected to be effective has received much attention in the public health literature in the context of QALYs and DALYs. In our analysis, we explore various options. These include: per single year (*t=1)*; the length of the evaluation (*e)*; and the remaining school life expectancy (*s)*.

Micro LAYS using Learning Estimates

To convert estimates of learning gains into micro LAYS, which we denote with the subscript *l*, we take three steps. First, we 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. Second, we then apply a quality-adjustment factor to derive learning-adjusted years of schooling (LAYS) that are comparable on a global scale. In the third and final step, we multiply by the period of time *t* over which effects are expected to last. We detail each step below.

First, we express the intervention's learning impacts in terms of equivalent years of school gained. Drawing on the methodology used by Evans and Yuan *(31)*, we derive equivalent years of schooling by expressing learning gained relative to learning in the status quo. For example, if students learned .25 standard deviations (which we will now refer to as .25σ) per year as a result of an intervention in South Africa, and students typically learn .25σ in a given year in South Africa, students will have learned a year's worth of South African schooling. This formulation can be expressed as follows:

$$\text{Equivalent Years of School} = \frac{\beta_i^{\sigma,test}}{\delta_{i,x}^{\sigma,test}}$$

where β is the learning gain produced by the intervention expressed in standard deviations (σ) per year in country *i; test* denotes the test used to measure learning; δ is the status-quo learning rate in standard deviations (σ) per year in country *i;* and *x* denotes the population for which this status-quo learning rate is calculated. For example, *x* could be the control group of the same study from which the β estimates are drawn, or *x* could be the student population in country *i,* so that δ becomes the average learning trajectory for the country as a whole. The choice of *x* will affect our interpretation. If we choose the control group, then the resulting value for equivalent years of schooling gained is relevant to the study sample only. If we choose national-average learning trajectories, we can interpret the value as the equivalent years of schooling gained at the national level. In this paper's main calculations, we use national level learning trajectories *n*, and in Section V, we explore the trade-offs in choices of what should be used as the measure of status quo learning.

Next, we adjust equivalent years of schooling gained in country *i* by the quality of learning L_i^h in that country relative to learning in a high-performance global benchmark country *h*:

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

We substitute in terms for $L_i^h = \frac{L_i}{L_h}$. This is analogous to the quality adjustment used in macro LAYS.

$$LAYS_l = \overbrace{\frac{\beta_i^{\sigma,test}}{\delta_{i,n}^{\sigma,test}}}^{equivalent\ years\ of\ school} * \overbrace{\frac{L_i}{L_h}}^{learning\ adjustment}$$

Next, we specify that both the numerator and denominator of the learning-adjustment term are expressed in terms of standard deviations (σ) on a *test* that is representative at national level *n* for each country, such that:

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

For the next step, we invoke two assumptions. First, we assume that learning is constant along a local trajectory. This assumption, explored in depth in Filmer et al. *(7)*, 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, as follows:

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

Second, we assume that standard deviations across tests and samples are comparable. This assumption, though not trivial, is not novel: it is implicitly invoked any time standard deviation effect sizes are compared across studies, which is the dominant practice in the literature on education interventions. We note that this assumption is most robust when learning gains in a given study are based on similar tests to the ones used in computing the learning-adjustment factor. This assumption simplifies our equation to:

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

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

$$LAYS_l = \frac{\beta_i}{\delta_{h,n}}$$

This expression produces an intuitive metric: the years of *h*-quality learning from the intervention. For example, assume that an intervention in South Africa yields .25σ per year of learning ($\beta_{South\ Africa} = .25$), and that in Singapore, a high-performance benchmark on international learning assessments, students learn 1σ over the course of a given year ($\delta_{Singapore} = 1$). Then we have .25 $LAYS_l$, in other words, the intervention enabled South African students to gain a quarter of a years' 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:

$$LAYS_l = \frac{\beta_i}{\delta_{h,n}} * t$$

As with micro LAYS based on participation estimates, *t* can take on a few values: a single year, such that *t=1*; the length of the evaluation, *e*; 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*.[3]

Because studies that have the statistical power to measure gains in access may not have the statistical power to measure gains in learning, we assume that interventions that increase years of schooling but for which learning gains are not well measured, lead to learning at the same rate as the average annual progression in the absence of the intervention. For those studies that include both learning and participation estimates, we use only the learning estimates.

SENSITIVITY AND 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.

High quality learning benchmark – We use .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 *(14)*. We explore three approaches to validating this high-performance benchmark: (a) rules of thumb and a range of effect sizes in reviews of multiple studies; (b) average annual learning trajectories in high-performance cases; and (c) policy-relevant learning changes.

First, rules of thumb: Cohen *(32)* proposed the following standardized effect-size benchmarks: at least .20 for "small" effects, .50 as "medium" effects, and .80 for "large" effects.

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

This framework has been broadly applied across interventions and contexts for decades. There is debate about the relevance of these indicators to education interventions: almost all U.S. based interventions have much smaller impacts *(33)*. But a recent review *(31)* finds similar upper benchmarks of learning gains in developing countries, with the largest effect size in their review reaching .81σ.

Second, average 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 *(2)* finds that the causal effect of an additional year of primary school in Vietnam is .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 *(34)*, building on methods used by Kane *(35)*. The largest year-on-year learning gains are between grade 1 and 2 and range from .97σ in reading to 1.03σ in math. Finally, we can derive approximate year-on-year changes for global high performers. We assume that the appropriate high-performance rescaled HLO benchmark is a score of 325 at the primary level. This score is assumed to be obtained over four years, since most primary international assessments occur in grade 4; average high-performance learning per year is thus 81.25 points. We then assume a within-country standard deviation of 85 points, based on the values for the five highest-performing countries using 2006 PISA microdata. Taking the ratio of these two values yields a year-on-year gain of .96σ.

Third, 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. A gain of .8σ would enable the United Kingdom or Vietnam to catch up to Singaporean learning levels: because the cross-country standard deviation is equivalent to 70 HLO points, so a .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 .99σ to 1.04σ in grades 4 and 8 *(34)*. A gain large enough to nearly close either of these gaps would be highly meaningful in policy terms.

In summary, these various approaches all yield high-performance benchmark learning rates ranging from around .8σ to 1.σ. In this paper, we use an artificial benchmark of .8σ for learning gains, which is a conservative high-performance benchmark consistent with this range.

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 Figure S2, we corroborate this de facto floor with data from EGRA, which in shows that an HLO score of 300 corresponds roughly to zero percent reading comprehension.

Figure S5 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 .2; by contrast, under the re-scaled version (x-axis), the distance is closer to .5.

Figure S6 compares results using the two score scales. In Panel A, we see that the main effect of the rescaling is to reduce the micro LAYS values that are based on participation impacts—for example, conditional cash transfers in Malawi. This is because under the original scale the maximum learning adjustment discounted a year of school by about half, since the *de facto* floor of the HLO scale was 300, which produced a learning-adjustment factor, L_i^h, of .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 .73 to .48, while countries on the lowest tail of distribution, such as Malawi, shift from a learning-adjustment of .57 to .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 $.8\sigma$. However, as an added sensitivity test, we can use the old scale to derive a new corresponding high-performance benchmark of 1.6σ. Panel B plots LAYS using this new benchmark.

These sensitivity tests show that the unscaled scores yield higher values for micro-LAYS based on participation estimates. Also, although unscaled scores largely preserve rank, they understate the degree of learning adjustment, especially for countries on the lower tail of the distribution. For these reasons, in the main results presented in this paper we use micro LAYS based on re-scaled scores. Since the lowest-performing countries are already far behind, re-scaling of scores is unlikely to yield major new insights and will not change ranks; however, we think re-scaling is important for capturing the full degree of the learning gap.

Standard deviations across tests and samples – We also test sensitivity of our results to differences in standard deviations across tests and samples. This is relevant because, for the interventions included in our comparisons, we use effect sizes expressed in standard deviations *of each intervention's test and sample* as inputs into LAYS. Note that this issue is not inherent to LAYS, which are just a unit of measurement that can be applied to any assessment results. In an ideal world (ideal from a comparability perspective, that is), all interventions and national-level assessments would use the exact same test to measure learning progress—and so there would be no issue with LAYS conversion.[4] However, in practice studies and interventions vary widely in the test used, making comparability an issue. LAYS can only be currently built with the data as they are, so a separate effort is needed to produce comparable underlying learning assessments.[5] Once that effort bears fruit, it will enhance the comparability of LAYS.

[4] Of course, that ideal test would need to be comprehensive and adaptive, so that it could be equally informative across all parts of the learning distribution.

[5] Short of a universal test, a number of efforts currently underway may improve comparability. One approach focuses on a set of common test items that can be inserted into assessments and linked to regional and international assessments, such as TIMSS and PISA. The Jameel Poverty Action Lab has launched an initiative to produce an "item bank" of questions for inclusion in education interventions to produce comparable assessment. The Rosetta

For now, therefore, we rely on standardized effect sizes. Standardized effect sizes are used to account for differences across measurement scales *(32)* 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" such that there are 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. Figure S7 compares standard deviations for Tanzania, Malawi, and Indonesia using various assessments: HLO scores derived from EGRA, Progress in International Reading Literacy Study (PIRLS), raw EGRA or Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) tests. We find only small differences of a few points, and as a result, the estimated learning rates per year across assessments are quite similar, as shown in Figure S8. As a robustness check, Figure S8 also shows learning rates per year using raw data from the assessments before they are converted to HLO scores.

In terms of standard deviations across samples, it is well known there is high variation across sample populations. As a result, when students answer five questions correctly in a country with less variation, this will appear as a larger standardized effect size than if they had gotten the same questions correct in a country with more variation. This is a feature of standard deviation comparisons rather than a bug, since standardized effect sizes produce *relative* comparisons to derive a sense of magnitude. Therefore, LAYS that use standardized effect sizes as inputs should prioritize comparisons of relative performance and rank orders over precise magnitudes and absolute performance.

Status-Quo Learning. When we compute LAYS from learning estimates, we first convert learning into equivalent years of school gained. To do this, we express learning relative to status-quo learning trajectories. For example, if students learned .25σ per year in an intervention in South Africa, and students typically learn .25σ in a given year in South Africa, students will have learned a year's worth of schooling in South Africa. A few options exist for possible status-quo learning trajectories: for example, we might use the national average learning trajectory, or the learning gains in the control group of the same evaluation from which effect sizes are drawn. Alternatively, we could also 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, with the embedded assumption that the standard deviations are comparable. In the main analysis reported above, we use national-level learning trajectories; in this section, we explore the alternative of using the study's control group to measure status-quo learning.

Stone initiative of the International Association for the Evaluation of Educational Achievement (IEA), the UNESCO Institute for Statistics (UIS), and other organizations are also working to improve comparability across regional and global assessments by having the same samples of students take both types of assessments.

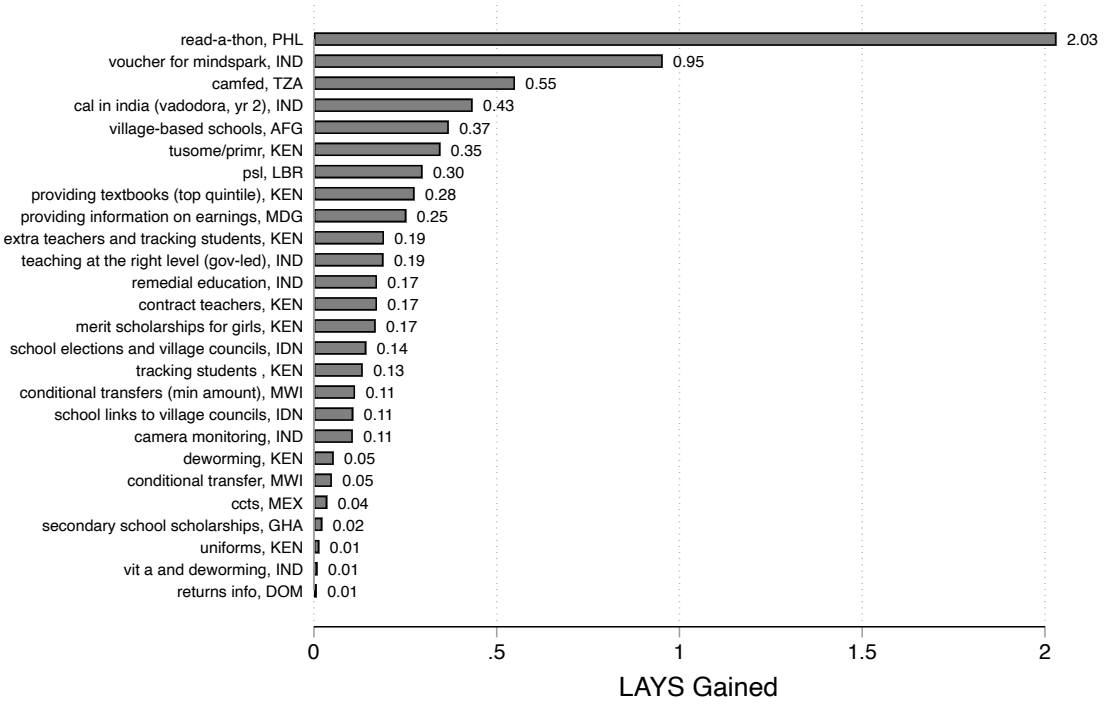
An advantage of using control-group status-quo learning is that this estimate is drawn from the same sample as the learning gains from the intervention are, so the two are directly comparable. If the study sample is not representative of the nation, however, then when we later apply a national-level 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 Figure S9, 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 the sample represents the nation.

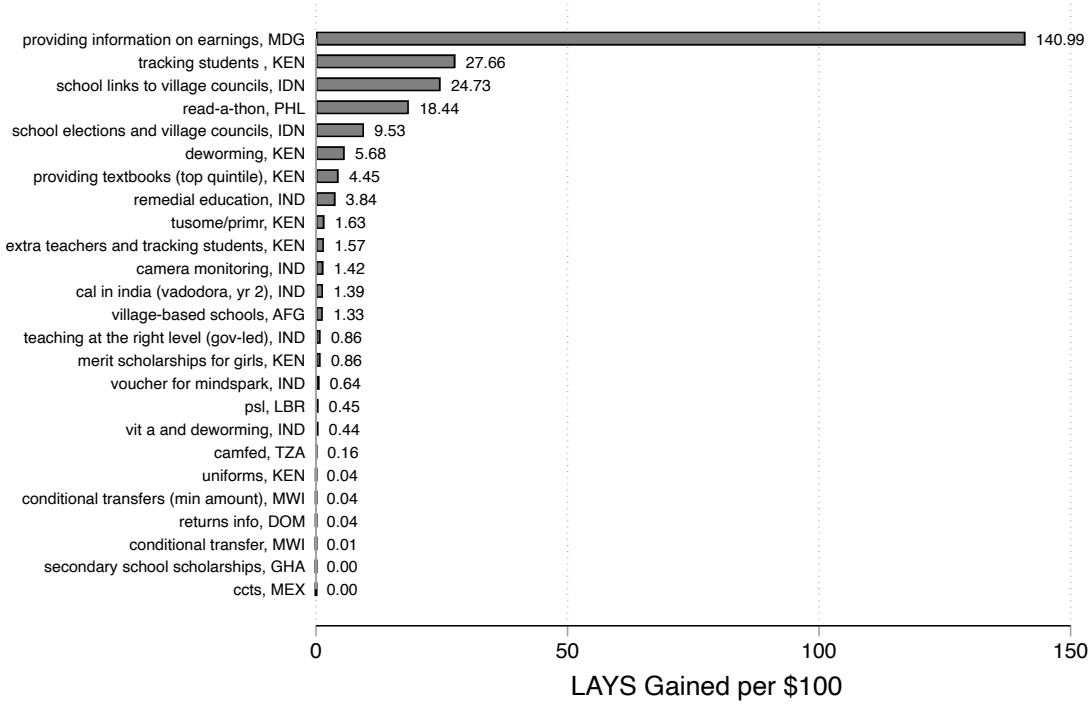
As a specific example, control-group status-quo gains in India in our sample of studies range from .5σ to .76σ, implying high variance from study to study. As another example, the status-quo control-group gain in Kenya in one study sample is just .045σ—ten times smaller than the smallest status-quo gains India (.5σ)—even though the annual learning rate in Kenya is much higher on average than in India, as suggested by the international rankings. On the World Bank Harmonized Learning Outcomes ranking, Kenya (ranked 57th) dramatically outperforms India (139th). These two examples imply that learning gains in study-sample control groups vary markedly from national average learning trajectories.

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

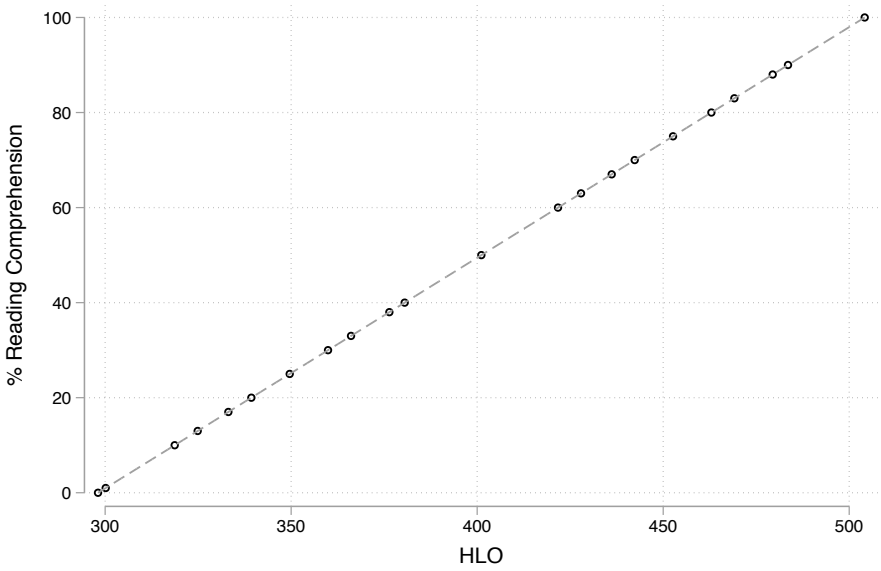
Supplementary Figures

Figure S1: Expressing LAYS gained per year ($t = 1$)





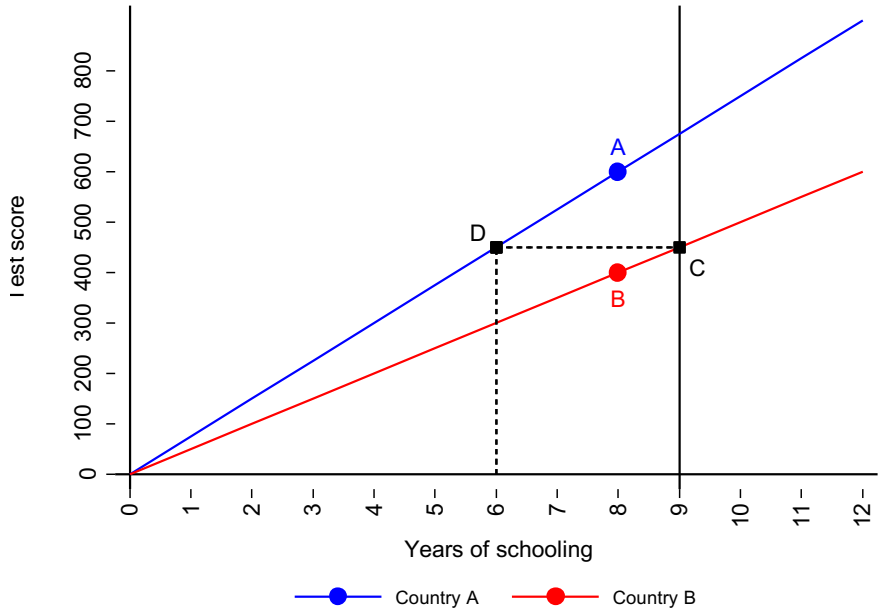
5



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.

5

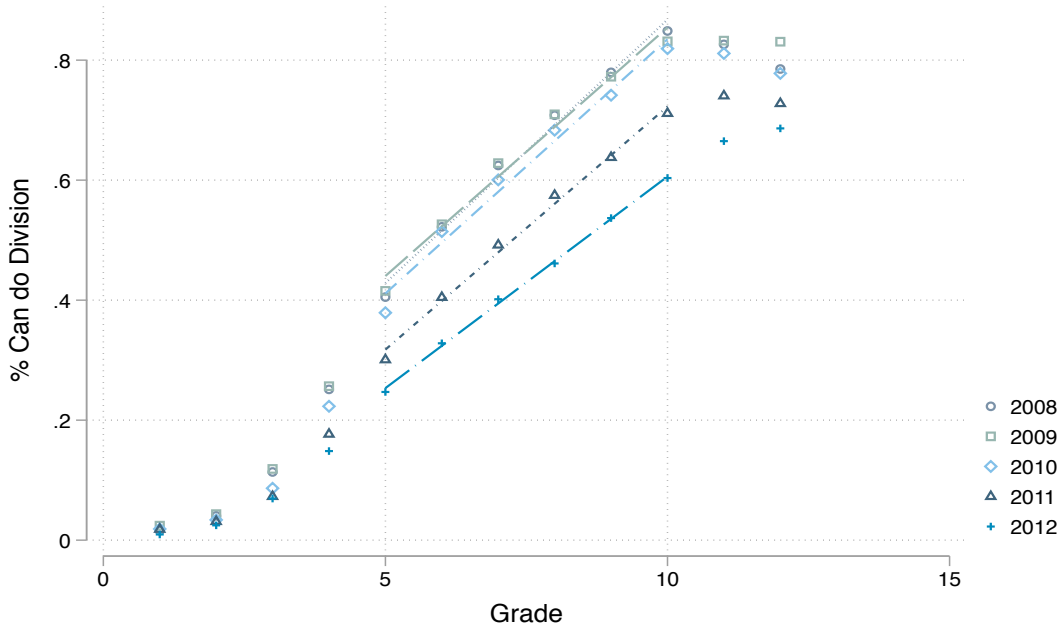
Figure S3: Constant Learning Trajectories



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

5

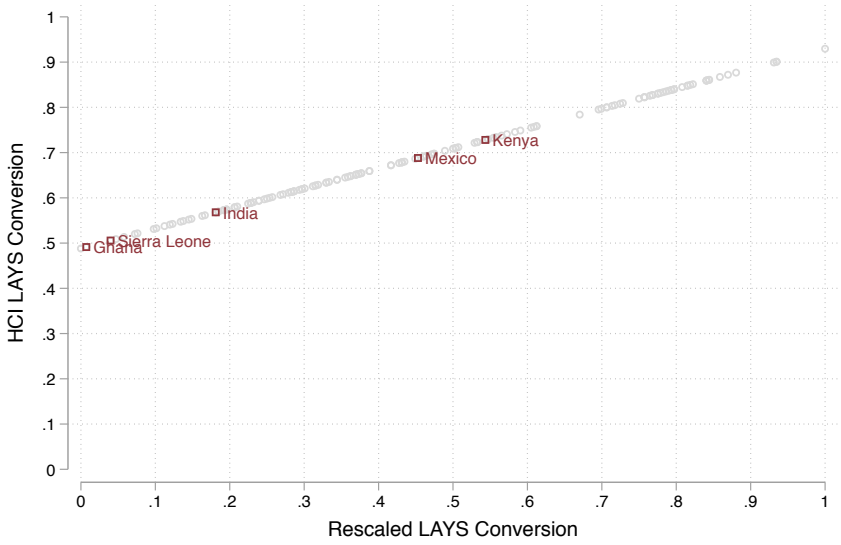
Figure S4: Learning Trajectories in India



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

Source: ASER India data from 2008 to 2012 as analyzed by Filmer et al. *(7)*.

Figure S5: Learning Adjustment Rates

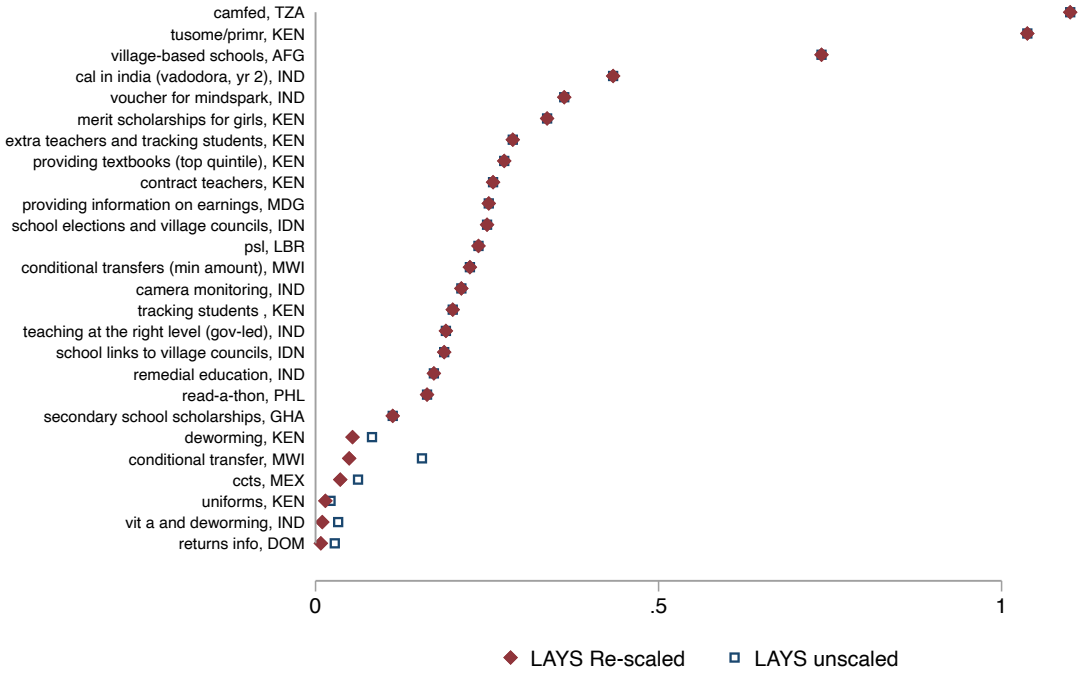


Notes: We rescale LAYS conversion rates. Initial conversion rates are based on scores which often floor around 300 due to underling test scores scales. Since the LAYS conversion rate is calculated out of 625, this produces a floor conversion rate of .48. However, when learning levels are very low this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

Figure S6: Comparing LAYS using scaled and unscaled test scores

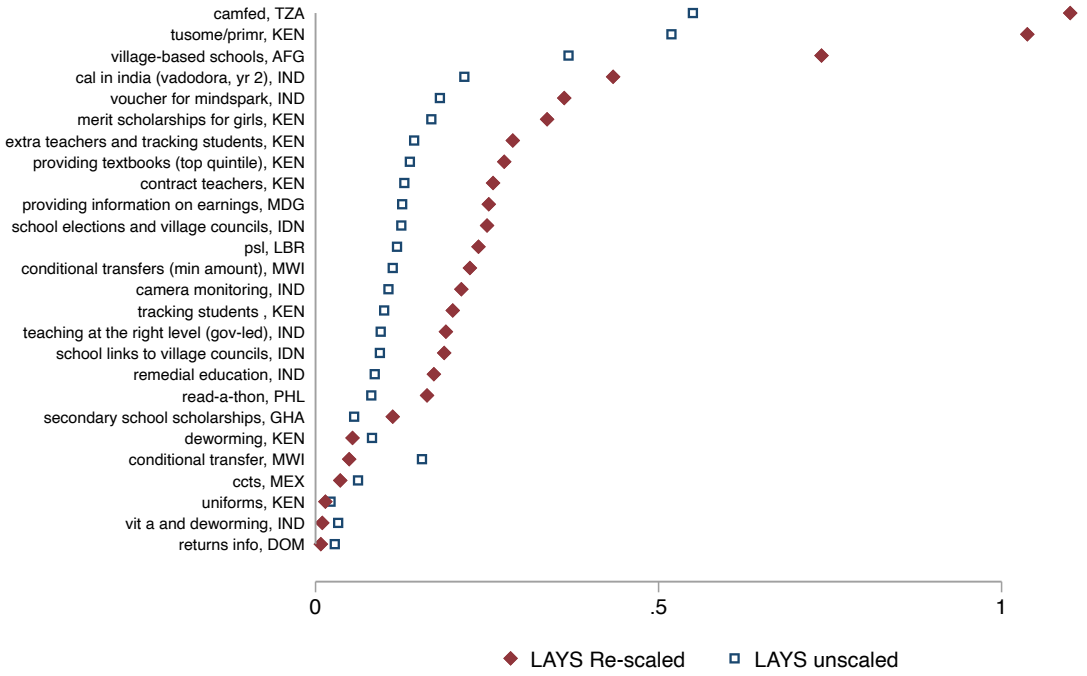
5

Panel A



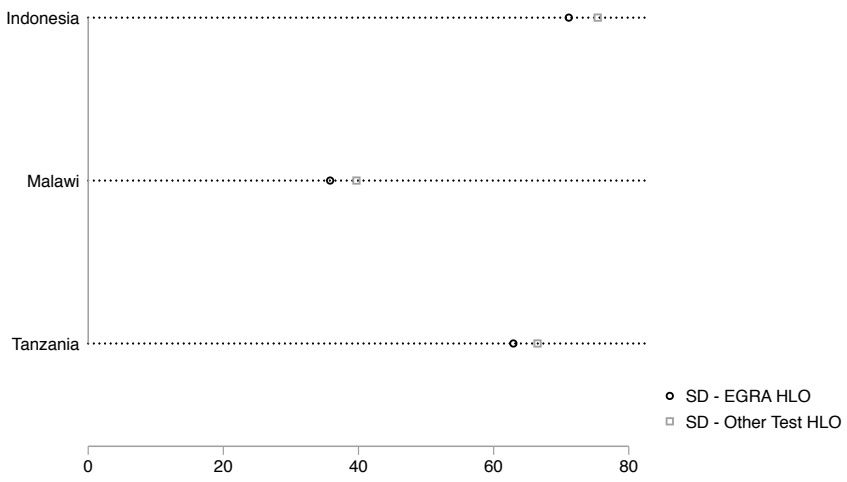
Notes: We rescale LAYS conversion rates for participation-based estimates to adjust for a quality factor that accounts for the new HLO scale. We do not adjust learning-based estimates in this figure since they are not adjusted by a learning-factor. Rather, they are based on an assumption of high-benchmark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

Panel B



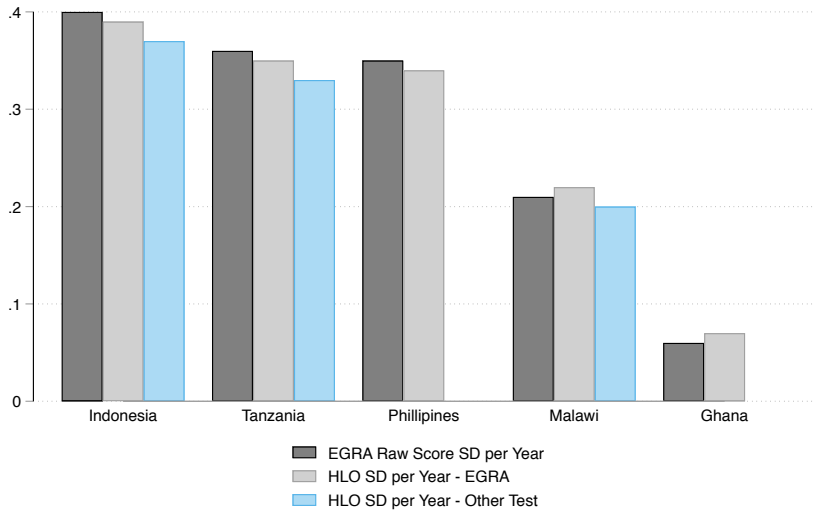
Notes: We rescale LAYS conversion rates and apply this rescaling to participation-based estimates to adjust for a quality-adjustment factor that accounts for the new HLO scale. In this figure we adjust learning-based estimates by deriving a new high-benchmark learning trajectory based on a new scale which yields 1.6σ. as a robustness test.

Figure S7: SD Comparisons, by Source Test



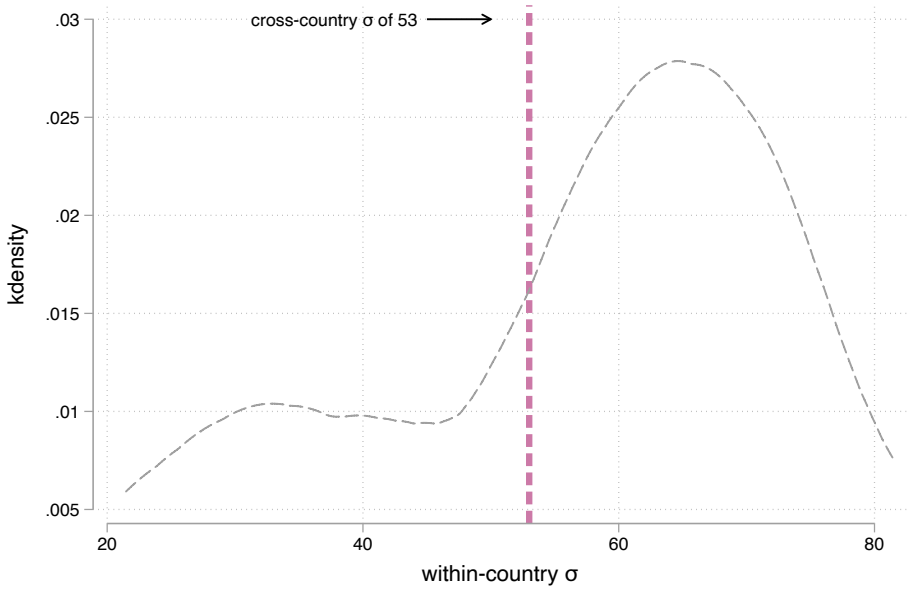
Notes: The ‘other test’ for Indonesia is PIRLS 2011; for Tanzania and Malawi it is SACMEQ 2007.

Figure S8: Learning Per Year (in SD), by Source Test



Notes: The 'other test' for Indonesia 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 S9: Within- vs. Cross-Country Variation in Test Scores



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

Conclusion

This thesis comprises three papers under the theme of external validity. External validity explores how casual effects of interventions or policies translate across contexts. I focus on human capital in developing countries. To address external validity questions well, we need comparable outcomes, comparable interventions, and identification of causal mechanisms between interventions and outcomes across contexts. My papers contribute to each aspect. Two of my papers focus on globally comparable outcomes to measure human capital. In my main paper, I conduct a large-scale randomized controlled trial of a sex education intervention in Botswana. This study builds on a trial showing a similar intervention reduced risky sex in Kenya ten years prior. My trial both provides replicable evidence across contexts as well as deepens understanding of the underlying mechanism driving behavior by testing novel dimensions, varying the message and messenger delivering sex education.

Each article in this thesis has significant policy ramifications. Two papers produce global public goods in the form of comparable educations outcomes and have been adopted by the World Bank, DFID and USAID in large-scale efforts. "Measuring Human Capital" produces a database of comparable learning outcomes for 164 countries from 2000-2017 and demonstrates for the first time on a global scale the importance of learning for economic development and growth. In "A New Micro Measure of Education Interventions: Learning-Adjusted Years of Schooling (LAYS)", I create a unified outcome metric of human capital to express impacts of interventions and policies across a broad array of education outcomes. This paper presents results from hundreds of impact evaluations across 44 countries in LAYS units. The most effective interventions have large effect sizes on the order of global learning gaps between low- and high-performing education systems This paper demonstrates the potential for LAYS to become a standard unit of economic analysis in education analogous to two standard outcome units in health: Quality-Adjusted Life Years (QALYs) and Disability-Adjusted Life Years (DALYs).

The "Messenger Matters" paper has far-reaching policy ramifications for the delivery of sex education globally. Teachers are often the default messenger for sex education. Teachers are respected figures in the community and a fixture of the education system. However, viable alternatives exist to deliver sex education. For example, many countries have national service programs which deploy young people in schools. AmeriCorps in the United States places young volunteers in over 12,000 schools throughout the country. In sub-Saharan Africa, over fifteen countries have national service programs with cumulative government investments of over $240 million per year. This presents a scalable model for near-peers to deliver sex

education in schools. Our results suggest that while the current default messenger for sex education, teachers, can be effective for transferring information, they might not reduce risky sex. In contrast, near-peers can reduce pregnancy by over 40 percent. Leveraging near-peers to deliver sex education could thus yield large human capital and welfare gains.

The main and fundamental insight into questions of external validity comes from the "Messenger Matters" paper. The effectiveness of the sex education message across both Botswana and Kenya, with up to 40 percent and 28 percent reductions in pregnancy, respectively, is striking since the trials took place across highly diverse contexts. The Kenya trial took place from 2003-2005, in a low-income context and when HIV medication availability was limited. The Botswana trial took place from 2014-2015 in an upper-middle income setting with widely available HIV medication. The results suggest that despite significant scope for replication failure, the trial replicated. These results reinforce the notion that information matters and can reduce risky sex, despite numerous other highly contextual factors driving risky sex, such as credit constraints. However, our finding that the messenger determines the direction and magnitude of effects within Botswana suggests that whether results translate across contexts might turn less on geography and associated covariates, which often receive the most attention in the literature on external validity, and more on the implementor or sender of information.

Future areas for research include continued generation of comparable outcomes, in education and beyond, enabling ingredients for replicable scientific inquiry and policy exchange across contexts. In addition future research might accelerate emphasis on rigorous replication trials across contexts, with a focus on previously underexplored dimensions on which interventions or policies might turn, such as the messenger, or other dimensions yet to be discovered.