UNIVERSITY OF CALIFORNIA SAN DIEGO

Three Essays on Improving Learning Outcomes in Africa

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Mauricio Tomas Romero

Committee in charge:

Professor Prashant Bharadwaj, Co-Chair
Professor Karthik Muralidharan, Co-Chair
Professor Jeffrey Clemens
Professor Gordon McCord
Professor Craig McIntosh

2018

The dissertation of Mauricio Tomas Romero is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Co-Chair

University of California San Diego

2018

DEDICATION

To Calde

EPIGRAPH

No se aferren a un único dogma. No sucumban ante las trampas de la ideología. No busquen todas las respuestas en un único libro o un solo predicador. No importa que tan elocuente sea. Esos son con frecuencia los peores.

Los que creen en una sola cosa, los que organizan el mundo con base en parejitas, en narrativas binarias — los civilizados y los barbaros, los explotados y los explotadores, los capitalistas y los proletarios, los buenos y los malos — casi siempre se equivocan, tanto en sus predicciones como en sus prescripciones.

En general desconfíen de los profetas, de los iluminados, de quienes creen en las soluciones totales, de todos aquellos que tienen más discurso que metodología y predican una falsa disyuntiva entre "un sistema injusto y corrupto que no puede mejorarse, y otro racional y armonioso que ya no habría que mejorar". Los profetas casi nunca predicen los desastres, con frecuencia los ocasionan.

El cambio social no es cuestión de todo o nada, es cuestión de más o menos. "En cuestiones prácticas uno no debe aspirar a la perfección".

El conocimiento práctico construye. Poco a poco pero construye. Las ideologías abstractas solo sirven para destruir. En últimas, el reformismo incremental, permanente, basado en la experiencia y el conocimiento de los problemas, es siempre más eficaz que las revoluciones basadas en concepciones ideológicas y visiones grandilocuentes.

Cambiar el mundo es difícil. Las "musculosas capacidades de la política" son una ilusión. Con la excepción, por supuesto, de las "musculosas capacidades" para hacer daño. Ejemplos abundan. No muy lejos de aquí.

Las leyes por sí solas no crean capacidades colectivas. Tampoco cambian la cultura. Ni modifican las normas sociales. No se puede legislar el conocimiento. Tampoco la moral. Las leyes sociales de Noruega y Grecia son las mismas. Los resultados, opuestos. Por algo será.

—Alejandro Gaviria

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This achievement belongs to many people; I could have never done this alone. My wife deserves at least half the credit. Without her, I would have never survived past the first year of grad school. Te amo tal y como estas.

The path that led me here can be traced back to my family. The moral compass my parents instilled in me led me to work on development. They taught me to invest my time and money in "experiences", which led me to a PhD. One of the persons I admire the most, my sister, studied economics. I followed her steps.

I am fortunate to have many great mentors. My pre-PhD mentors: Alejandro Gaviria, Álvaro Riascos, and Diego Jara remain a strong intellectual influence. They taught me most of the economics I know, that the devil is in the details, how to think critically, and how to be a good scholar. Prashant Bharadwaj, Karthik Muralidharan, and Craig McIntosh: Thank you for the training, the lessons, and the friendship. This dissertation is as much yours as mine. I hope I can pay it forward one day, mentoring someone as well as you have mentored me. Justin Sandefur and Isaac Mbiti took on the role of (unofficial) advisers, without me ever asking them. They deserve more than a thank you note in this dissertation, but that's all I have to offer. Thank you. Every single UCSD economics faculty member was instrumental to my education, but Kate Antonovics, Eli Berman, Jeffrey Clemens, Julie Cullen, Graham Elliott, Itzik Fadlon, Roger Gordon, Josh Graff-Zivin, Gaurav Khanna, Gordon McCord, Paul Niehaus, and Kaspar Wuthrich deserve special credit.

I reserve a special thank you for my friends who double as co-authors. Santiago Saavedra, thank you for the unwavering support in all my life adventures, both outside and inside academia. Wayne Sandholtz, thank you for signing up for what turned out to be quite a ride. I could not have finished my job market paper without your help. Justin, getting to know you was the best part of PSL. Thank you for "stirring the pot";

alguien tiene que llevar la contraria.

My friends made the past years an unforgettable experience. The stimulating environment and the friendship UCSD graduate students from all cohorts provided made UCSD a happy and productive place. Mitch Downey, John Rehbeck, and Zach Breig: Thank you for pushing me through quals. Mitch and Diego Vera-Cossio: Thank you all the insightful discussions. I found two adoptive families in San Diego: The Downeys and The Schmottos. Thank you for making San Diego feel like home. I have many happy memories cycling (with Mike, Ajay, and Patrick), climbing (with Rico, Yvonne, Ajay, Martin, and Becky), and drinking beer (with pretty much everyone I know, but specially Diego, Bruno, and Patrick) in San Diego.

I finish this journey with more happy memories than I ever imagined. I moved to San Diego, I settled (with Mitch, and then Olivia, and then Calde). I traveled around the west coast with friends (Maria, Santiago, Calde, and David) and family (Nena, Jaime, Lauris, Paulis, Jota, Libia, Maita, Patricia, and Daniel). I camped, I biked, I climbed, I had beers. I learned how to cook. I studied. I started doing field work in Tanzania. I started my own research projects. Many friends and family visited. I biked from the lowest to the highest point in the continental U.S., in a single push. I proposed to the love of my life at the top of Mt. Whitney. We had a (temporary) adoptive daughter. We got married. We drove from the West to the East Coast. We moved to New Haven. We meet new friends from around the globe. I was a stay-home husband for a while. We lived through seasons. We canoed, biked, hiked, and camped in the North East. We honeymooned bike-touring in Cuba. I did El Camino de Santiago with my parents. I started a research project in Liberia that became my JMP. I traveled more in two years than in the previous 28. I spent no more than 30 continuous days in any given place in that time period. We moved to Washington, DC. We met new friends. We decided to move to Mexico City. I worked hard, but above all I enjoyed life.

Chapter 1, in full, is currently being prepared for submission for publication of the material. Romero, Mauricio; Sandefur, Justin; Sandholtz, Wayne Aaron. “Outsourcing Service Delivery in a Fragile State: Experimental Evidence from Liberia”. The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Romero, Mauricio; Chen,Lisa; Magari, Noriko. “Cross-Age Tutoring: Experimental Evidence from Kenya”. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Mbiti, Isaac; Muralidharan, Karthik; Romero, Mauricio; Schipper,Youdi; Manda, Constantine; Rajani, Rakesh. “Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania”. The dissertation author was a primary investigator and author of this material.

VITA

Employment	2018-	Assistant Professor Centro de Investigación Económica Instituto Tecnológico Autónomo de México
Education	2012-2018	Ph.D. Economics, University of California San Diego
	2012-2013	M.A. Economics, University of California San Diego
	2004-2010	B.A. Economics (*summa cum laude*) B.A. Mathematics (*cum laude*) Universidad de los Andes, Bogotá, Colombia
Past positions	2009-2012	Junior Researcher Quantil | Matematicas Aplicadas
Awards	2017:	Clive Granger Fellowship University of California San Diego
	2015:	Benjamin C. Horne Memorial Prize University of California San Diego
	2012-2016	Lauchin Currie Scholarship Central Bank of Colombia

ABSTRACT OF THE DISSERTATION

Three Essays on Improving Learning Outcomes in Africa

by

Mauricio Tomas Romero

Doctor of Philosophy in Economics

University of California San Diego, 2018

Professor Prashant Bharadwaj, Co-Chair
Professor Karthik Muralidharan, Co-Chair

Too often governments fail to provide access to quality public services to the poor. My work focuses on this issue and the bottlenecks that impede high-quality government provision of education in sub-Saharan African countries.

Chapter 1 studies whether outsourcing public services to private entities improves service delivery in fragile states. It provides experimental evidence from the Partnership Schools for Liberia (PSL) program, which delegated management of 93 public schools to eight different private organizations. Within one academic year, outsourcing increased students scores in English and math by .18σ, relative to control schools. While the highest-performing providers generated increases in learning of 0.26σ, the lowest-performing providers had no impact on learning. Consistent with the rules of provider contracts, we find no evidence that providers engaged in student selection. However, providers were allowed to shift pupils from oversubscribed schools and underperforming teachers to other government schools. These results suggest that leveraging the private sector to improve service delivery in fragile states is promising, but they also highlight the importance of procurement rules and contracting details to aligning public and private interests.

Chapter 2 studies cross-age tutoring — in which older students tutor younger students — as an inexpensive alternative for providing personalized instruction. Tutoring in math has a small positive effect on math test scores. The effect is concentrated among middle-ability students, suggesting that tutors are not able to help advanced learners and those lagging behind grade-level competencies.

Chapter 3 studies complementarities across policies in education. While the idea that complementarities across policies can lead to increasing returns has a long tradition in economics, there is limited evidence that clearly identifies such complementarities. It presents evidence of the impact of providing schools with (a) unconditional capitation grants, (b) bonus payments to teachers based on student performance, and (c) both of the above. We find no impact on student learning from providing either the grants or teacher incentives but significant positive effects from providing both. We find strong evidence of complementarities between improving school inputs and teacher incentives, with the combined effect being greater than the sum of the individual effects.

Chapter 1

Outsourcing Service Delivery in a Fragile State: Experimental Evidence from Liberia

(Co-authors: Justin Sandefur and Wayne Aaron Sandholtz)

1.1 Introduction

Fragile states are often unable to deliver basic services to their citizens. Building state capacity is difficult and takes time. Outside efforts to promote stronger institutions often fail (Pritchett and Woolcock 2004). Influential studies in the 1990s concluded that development aid was least effective in poorly governed states, and advocated directing aid elsewhere (Burnside and Dollar 2000; Collier and Dollar 2002). An alternative strategy is to sidestep the bottleneck of weak state capacity in fragile states by outsourcing the provision of public services to private providers (Krasner and Risse 2014; Collier 2016b). This paper tests the latter approach.

Both theoretical and empirical analyses of outsourcing suggest a need for caution. Theoretically, contracting out the provision of a public good may worsen its quality if contracts are incomplete (Hart, Shleifer, and Vishny 1997). While contractors have incentives to increase cost-efficiency to maximize profits, they may cut costs legally, through actions that are not in the public's best interest but still within the letter of the contract. Empirically, while outsourcing has delivered better outcomes in some settings (e.g., water services in Argentina (Galiani, Gertler, and Schargrodsky 2005) and food distribution in Indonesia (Banerjee et al. 2015b)), it has failed to do so in others (e.g., prisons in the U.S. (Useem and Goldstone 2002) and in Brazil (Cabral, Lazzarini, and Azevedo 2013)).

In the case of education, proponents argue that combining public finance with private management has the potential to overcome a trade-off between efficiency and equity (Patrinos, Osorio, and Guáqueta 2009). On the efficiency side, evidence suggest that private firms (Bloom and Van Reenen 2010; Bloom, Sadun, and Van Reenen 2015) and schools (Bloom et al. 2015; Muralidharan and Sundararaman 2015) tend to be better managed than their public counterparts. However, fee-charging private schools may increase inequality and induce sorting (Hsieh and Urquiola 2006; Lucas and Mbiti 2012; Zhang 2014). Most of the empirical evidence on outsourcing education to overcome this trade-off comes from the U.S., where charter schools appear to improve learning outcomes when held accountable by a strong commissioning body (Cremata et al. 2013; Woodworth et al. 2017). But there is limited evidence on whether outsourcing education can improve learning levels in developing countries, and particularly in fragile states, where governments have limited capacity to enforce top-down accountability.

In this paper we provide experimental evidence on outsourcing education in Liberia, a low-income country with limited state capacity. The Liberian government is unable to deliver most public goods and services, including universal, high-quality primary education to all children. Net primary enrollment stood at 38% in 2014, compared to 80% across all low-income countries (WB 2014). We study the Partnership Schools for Liberia (PSL) program, which delegated *management* of 93 public schools (3.4% of all public primary schools, serving 8.6% of students enrolled in public early childhood and primary) to eight different private organizations. Providers received funding on a per-pupil basis. In exchange, they were responsible for the daily management of the schools. These schools were to remain free and non-selective (i.e., providers were not allowed to charge fees or screen students based on ability or other characteristics). PSL school buildings remained under the ownership of the government. Teachers in PSL schools were civil servants, drawn from the existing pool of government teachers.

We study this public-private partnership by randomly assigning existing public schools to be managed by one of several private operators. We randomized treatment within matched pairs of schools (based on infrastructure and geography), which allows us to estimate treatment effects across providers. Since treatment assignment may change the student composition across schools, we sampled students from pre-treatment enrollment records. We associate each student with her "original" school, regardless of what school (if any) she attends in later years. The combination of random assignment of treatment at the school level with sampling from a fixed and comparable pool of students allows us to provide clean estimates of the program's intention-to-treat (ITT) effect on test scores, uncontaminated by selection. Program effects could arise from improved teaching, better resources, or peer effects through selection of other students.[1]

The ITT effect on test scores after one year of the program is 0.18σ for English and 0.18σ for mathematics. These gains do not reflect teaching to the test, as they are also seen in new questions administered only at the end of the school year and in conceptual questions with a new format. The average increase in test scores for each extra year of schooling is relatively low in the control group and equal to 0.31σ in English and 0.28σ in mathematics. Thus, the treatment effect is equivalent to 0.56 and 0.65 additional years of schooling for English and mathematics. Consistent with the promise that publicly financed, but privately managed schools would improve efficiency without compromising equity, we find no evidence of heterogeneity by students' socio-economic status, gender, or grade. While the experiment was designed to overcome this bias if it occurred, there is also no evidence that providers engaged in student selection: the probability of remaining in a treatment school is unrelated to age, gender, household wealth, or disability.

These gains in test scores reflect a combination of additional inputs and improved management. There is some evidence that both mattered. PSL doubled yearly per-student expenditure (relative to a mean of $\sim$\$50 in the control group) as part of the program, and some providers independently raised and spent far more.[2] In addition,

[1] We focus on the ITT effect, but the treatment-on-the-treated (ToT) effect (i.e., the treatment effect only for students that actually attended a PSL school in 2016/2017) can be computed, under standard assumptions, using the fraction of students originally assigned to treatment schools who are actually in treatment schools at the end of the 2016/2017 schools year (77%) and the fraction of students assigned to control schools who are in treatment schools at the end of the 2016/2017 schools year (0%).

[2] This increase is unprecedented in the development literature. Two school grant programs that

PSL schools had an average of one teacher per grade compared to 0.78 per grade in traditional public schools. The program also increased management quality, as proxied by time on task. Teachers in PSL schools were 50% more likely to be in school during a spot check (20-percentage-point increase, from a base of 40%) and 43% more likely to be engaged in instruction during class time (15-percentage point increase, from a base of 35%). Non-experimental mediation analysis using observational variation in management, inputs, and teachers suggests at least half of PSL's learning impacts can be explained by better management. Teacher attendance and time on task improved for incumbent teachers, which we interpret as evidence of better management.

While average scores in PSL schools were higher, there is significant heterogeneity across providers. Since each provider was randomly assigned schools in a matched-pair design, we are able to estimate (internally valid) treatment effects for each provider. To account for differences in the specific contexts where each provider operated, we adjust for observed pre-treatment characteristics in a regression framework. To account for the small number of schools run by some providers (and thus noisy estimates), we estimate provider-specific effects using a Bayesian hierarchical model along the lines proposed by Rubin (1981). While the highest-performing providers generated increases in learning of 0.26σ, the lowest-performing providers had no impact on learning.

One worry is that improved performance in PSL schools might come at the expense of traditional public schools. Unenrolling students and dismissing teachers may have allowed contractors to boost learning outcomes in their own schools, while imposing negative externalities on the broader school system. In principle, removing under-performing teachers need not have negative spillovers. In practice, dismissed

doubled per-school expenditure (excluding teacher salaries) in India and Tanzania increased per-student expenditure on the order of $ 3-10 per student (Das et al. 2013; Mbiti et al. 2017). Of 14 programs reviewed by JPAL, no program spent more than $30 per student (inclusive of all implementation costs). See https://www.povertyactionlab.org/policy-lessons/education/increasing-test-score-performance for details.

teachers ended up either teaching at other public schools or receiving pay without work (as firing public teachers was almost impossible). Reshuffling teachers is unlikely to raise average performance in the system as a whole, and Liberia already has a tight budget and short supply of teachers (the literacy rate is below 50%). Hence, large dismissal of teachers is unsustainable if the program expands. Similarly, reducing class sizes may be good policy, but shifting students from PSL schools to other schools is unsustainable and may lead us to overstate the scalable impact of the program.

Some providers do engage in behavior that could create these sorts of negative spillovers, and some of this behavior can be explained by differences in contract terms. The largest provider bypassed the competitive procurement process to negotiate a bilateral agreement with the government, and thus was not covered by the same contract as other providers. While other providers were reimbursed on a per pupil basis from a pooled fund, the largest provider was funded by lump-sum grants, and limitations on removing government teachers were stipulated only verbally (every other provider had written limitations in the contract).[3] This provider unenrolled pupils after taking control of schools with large class sizes, and removed 74% of incumbent teachers from its schools.[4]

However, contract differences cannot easily explain all differences in provider behavior. All providers were authorized to cap class sizes, and no provider received payment for enrolling students beyond sixty-five pupils per class. Yet several providers enrolled more students than they were paid for. The Ministry allowed all providers to replace up to 40% of under-performing teachers, yet our results show no discernible effect on teacher exit rates for other providers. Differences in behavior with uniform contracts suggest differences in mission alignment, à la Besley and Ghatak (2005) or

[3] Contract differences are endogenous. Thus, we cannot identify whether behavior is different because of unobservable differences in providers' characteristics or differences in contracts.

[4] As mentioned above, there is no evidence of selective unenrollment based on observable characteristics.

Akerlof and Kranton (2005), that may be important when outsourcing public services.

Turning to whether PSL is a good use of scarce funds, we compare the effect of the program to other successful interventions studied in the literature. However, many education interventions have either zero effect or provide no cost data for cost-effectiveness calculations (Evans and Popova 2016). At present, providers have expressed interest in the program with an offer of a $50 subsidy per pupil, over and above the Ministry of Education's $50 expenditure per pupil in all schools.[5] Using this long-term cost target of $50, learning gains of $.18\sigma$ on average and even 0.26σ for the best-performing providers represent low cost-effectiveness relative to many alternative interventions in the literature (Kremer, Brannen, and Glennerster 2013). However, Liberia is a challenging environment and cost-effectiveness calculations from other contexts are far from perfect comparisons for this fragile state. Furthermore, it is not clear that traditional schools would have been capable of using additional resources allocated through a different intervention to improve performance.

Managing private providers requires some state capacity, but it may be more feasible to augment the capacity to procure, contract, and manage private providers, than to augment the capacity to provide services directly.[6] Hart, Shleifer, and Vishny (1997) argue that the bigger the adverse consequences of non-contractible quality shading, the stronger the case for governments to provide services directly.[7] Some

[5] In the first year, providers spent far more than this amount. But if the providers are willing to enter into agreements in which the government pays $50 per pupil, providers' losses are inconsequential to the government, unless the providers spend more in the first years of the program to prove effectiveness but plan to reduce expenditures once they sign long-term contracts.

[6] In the particular case of PSL, the government received support from the Ark Education Partnerships Group for the procurement and contracting process.

[7] Empirically, in cases where quality is easy to measure and to enforce, such as water services (Galiani, Gertler, and Schargrodsky 2005), outsourcing seems to work. Similarly, for primary health care, where quality is measurable (e.g., immunization and antenatal care coverage), outsourcing improve outcomes in general (Loevinsohn and Harding 2005). In contrast, for services for which quality is difficult to measure, such as prisons (Useem and Goldstone 2002; Cabral, Lazzarini, and Azevedo 2013), outsourcing seems to be detrimental. Contrary to primary health care, there is some evidence that contracting out advanced care (where quality is harder to measure) increases expenditure without increasing quality (Duggan 2004).

quality aspects of education are easy to measure (e.g., enrollment and basic learning metrics), but other are harder (e.g., socialization and selection). We provide the first experimental estimates on contracting out *management* of existing public schools in a developing country (for a review on the few existing non-experimental studies see Aslam, Rawal, and Saeed (2017)).[8] While outsourcing management works on average, we find heterogeneity in learning outcomes across providers and that limited state capacity to monitor contractors led to actions that might generate negative spillovers for the broader education system.

Previous studies on public-private partnerships in education have focused on charter schools in the United States, using admission lotteries to overcome endogeneity issues (for a review see Chabrier, Cohodes, and Oreopoulos (2016) and Betts and Tang (2014)). But oversubscribed charter schools are different (and likely better) than undersubscribed ones, truncating the distribution of estimated treatment effects (Tuttle, Gleason, and Clark 2012). We provide treatment effects from across the distribution of outsourced schools in this setting. Relatedly, relying on school lotteries implies that the treatment estimates capture the joint impact of outsourcing *and* the provider. We provide treatment effects across a list of providers, carefully vetted by the government, and show that the provider matters.

Recent theoretical and experimental results have highlighted the role of state capacity in service delivery (Ladner and Persson 2009; Besley and Persson 2010; Muralidharan, Niehaus, and Sukhtankar 2016). We complement these results by showing the strength and weaknesses of outsourcing as an alternative to improve service delivery in the absence of state capacity. Our results highlight that the success of public-private partnerships hinge on the details of the partnership. At least under certain conditions,

[8] A related paper to ours increased the supply of schools through a public-private partnership in Pakistan (Barrera-Osorio et al. 2013). However, it is difficult to disentangle the effect of increasing the supply of schools from the effect of privately provided, but publicly funded schools.

leveraging the private sector can improve service delivery in fragile states. This is promising. But our results also highlight the importance of procurement rules and contracting details in aligning public and private interests. Contracts are by nature incomplete and subject to regulatory capture; competition requires active encouragement. More theoretical and empirical research is needed to understand how different arrangements of procurement, contracts, and entry and exit dynamics affect the long-term outcomes of public-private partnerships such as this one.

1.2 Experimental design

1.2.1 The program

Context

The PSL program breaks new ground in Liberia by delegating management of government schools and employees to private providers. Nonetheless, a strong role for private actors — such as NGOs and USAID contractors — in providing school meals, teacher support services, and other assorted programs in government schools is the norm, not an innovation. Over the past decade, Liberia's basic education budget has been roughly $40 million per year (about 2-3% of GDP), while external donors contribute about $30 million. This distinguishes Liberia from most other low-income countries in Africa, which finance the vast bulk of education spending through domestic tax revenue (UNESCO 2016). The Ministry spends roughly 80% of its budget on teacher salaries (Ministry of Education - Republic of Liberia 2017), while almost all of the aid money bypasses the Ministry, flowing instead through an array of donor contractors and NGO programs covering non-salary expenditures. For instance, in 2017 USAID tendered a $28 million education program to be implemented by a U.S. contractor in public schools over a five year period (USAID 2017). The net result of this financing system is that many "public" education services in Liberia beyond teacher salaries are provided by non-state actors. On top of that, more than half of children in preschool and primary attend private schools (Ministry of Education - Republic of Liberia 2016a).

A second broad feature of Liberia's education system, relevant for the PSL program, is its performance: Not only are learning levels low, but access to basic education and progression through school remains inadequate. The Minister of Education has cited the perception that "Liberia's education system is in crisis" as the core justification for the PSL program (Werner 2017). While the world has made great progress towards universal primary education in the past three decades (worldwide net enrollment was almost 90% in 2015), Liberia has been left behind. Net primary enrollment stood at only 38% in 2014 (WB 2014). Low *net* enrollment is partially explained by an extraordinary backlog of over-age children (see Figure 1.1): The median student in early childhood education is eight years old and over 60% of 15 years olds are still enrolled in early childhood or primary education (LISGIS 2016). Learning levels are low: Only 25% of adult women who finish elementary school can read a complete sentence (LISGIS 2014) (there is no information for men).

Intervention

The Partnership Schools for Liberia (PSL) program is a public-private partnership (PPP) for school *management*. The Government of Liberia contracted multiple non-state providers to run ninety-three existing public primary and pre-primary schools.[9] Providers receive funding on a per-pupil basis. In exchange they are responsible for the daily management of the schools.

Eight providers were allocated rights to manage public schools by the govern-

[9] There are nine grades per school: three early childhood education grades (Nursery, K1, and K2) and six primary grades (grade 1 - grade 6).

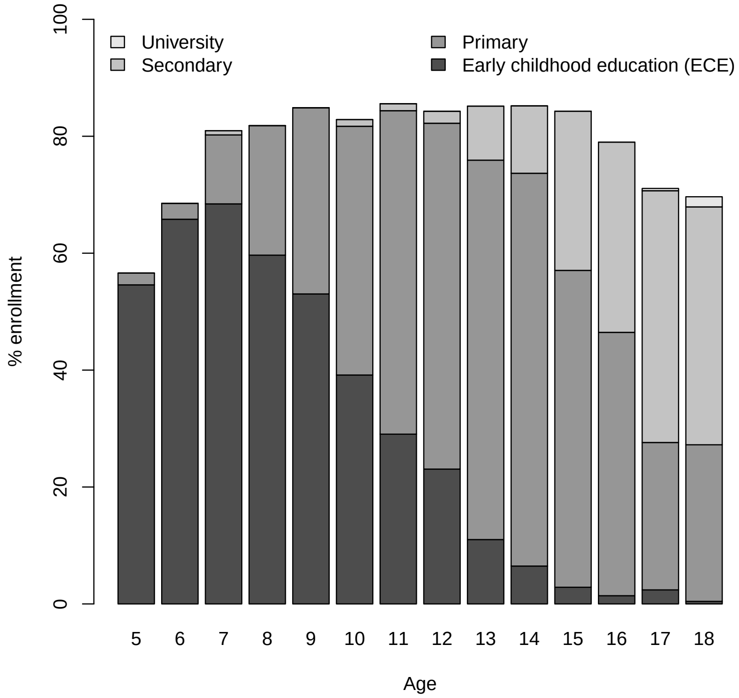


Figure 1.1: Enrollment by age
Note: Authors' calculations based on 2014 Household Income and Expenditures Survey.

ment under the PSL program. The organizations are as follows, ordered by the number of schools they manage that are part of the RCT: Bridge International Academies (23 schools), BRAC (20 schools), Omega Schools (19 schools), Street Child (12 schools), More than Me (6 schools), Rising Academies (5 schools), Youth Movement for Collective Action[10] (4 schools), and Stella Maris (4 schools).[11]

Rather than attempting to write a complete contract specifying private providers'

[10] Youth Movement for Collective Action began the evaluation as "Liberian Youth Network," or LIYONET. The group has since changed its name.

[11] Bridge International Academies is managing two additional demonstration schools that were not randomized and are thus not part of our sample. Omega Schools opted not to operate two of their assigned schools, which we treat as non-compliance. Rising Academies opted not to operate one of their assigned schools (which we treat as non-compliance), and was given one non-randomly assigned school in exchange (which is outside our sample). Therefore, the set of schools in our analysis is not identical to the set of schools actually managed by PSL providers.

full responsibilities, the government opted instead to select organizations it deemed aligned with its mission of raising learning levels.[12] After an open and competitive bidding process led by the Ministry of Education with the support of the Ark Education Partnerships Group (henceforth Ark, a UK charity), the Liberian government selected seven organizations, of which six passed financial due diligence. Stella Maris did not complete this step and, although included in our sample, was never paid. The government made a separate agreement with Bridge International Academies (not based on a competitive tender), but considers Bridge part of the PSL program.

PSL schools remain public schools that should be free of charge and non-selective (i.e., providers are not allowed to charge fees or to discriminate in admissions, for example on learning levels). While PSL schools should be free at all levels, traditional public schools are not fully free. Public primary education is nominally free starting in Grade 1,[13], but tuition for early childhood education in traditional public schools is stipulated at LBD 3,500 per year (about $38).

PSL school buildings remain under the ownership of the government. Teachers in PSL schools are civil servants, drawn from the existing pool of government teachers. The Ministry of Education's financial obligation to PSL schools is the same as all government-run schools: It provides teachers and maintenance, valued at about USD 50 per student. A noteworthy feature of PSL is that providers receive *additional* funding of USD 50 per student (with a maximum of USD 3,250 or 65 students per grade). Neither Bridge International Academies nor Stella Maris received the extra $50 per pupil. As mentioned above, Stella Maris did not complete financial due diligence. Bridge

[12] Some agency problems related to contracting out the provision of a public good are alleviated by "mission-matching" (Besley and Ghatak 2005; Akerlof and Kranton 2005). At the time of writing, an expansion of the program was underway. Preliminary details from this expansion suggest that there will be some type of results-based accountability, in which part of the providers' payments will be conditional on achieving predetermined milestones.

[13] Officially, public schools are free, but in reality most charge informal fees. See Section 1.3.4 for statistics on these fees.

International Academies had a separate agreement with the Ministry of Education and relied entirely on direct grants from donors. Providers have complete autonomy over the use of these funds (e.g., they can be used for teacher training, school inputs, or management personnel).[14] On top of that, providers may raise more funds on their own.

Providers must teach the Liberian national curriculum, but may supplement it with remedial programs, prioritization of subjects, longer school days, and non-academic activities. They are also welcome to provide more inputs such as extra teachers, books or uniforms, as long as they pay for them.

The intended differences between treated (PSL) and control (traditional public) schools are summarized in Table 1.1. First, PSL schools are managed by private organizations. Second, PSL schools were theoretically guaranteed one teacher per grade in each school, plus extra funding. Third, private providers are authorized to cap class sizes. Finally, while both PSL and traditional public schools are free for primary students starting in first grade, public schools charge early-childhood education (ECE) fees.

What do providers do?

Providers enjoy considerable flexibility in defining the intervention. They are free to choose their preferred mix of, say, new teaching materials, teacher training, and managerial oversight of the schools' day-to-day operations.

[14] Providers may spend some of their funds hiring more teachers (or other school staff); thus is possible that some of the teachers in PSL schools are not civil servants. However, this rarely occurred in practice. Only 8% of teachers in PSL schools were paid by providers at the end of the school year. Information interviews with providers indicate that in most cases, the providers are paying these salaries while awaiting placement of the teachers on the government payroll, and they expect to be reimbursed by the government once that occurs.

Table 1.1: Policy differences between treatment and control schools

	Control schools	PSL treatment schools
Management		
Who owns school building?	Government	Government
Who employs and pays teachers?	Government	Government
Who manages the school and teachers?	Government	Provider
Who sets curriculum?	Government	Government + provider supplement
Funding		
Primary user fees (annual USD)	Zero	Zero
ECE user fees (annual USD)	$38	Zero
Extra funding per pupil (annual USD)	NA	$50[a] + independent fundraising
Staffing		
Pupil-teacher ratios	NA	Promised one teacher per grade, allowed to cap class sizes at 45-65 pupils[b]
New teacher hiring	NA	First pick of new teacher-training graduates[c]

[a] Neither Bridge International Academies nor Stella Maris received the extra $50 per pupil.

[b] Bridge International Academies was authorized to cap class sizes at 55 (but in practice capped them at 45 in most cases as this was allowed by the MOU), while other providers were authorized to cap class sizes at 65.

[c] Bridge International Academies has first pick, before other providers, of the new teacher-training graduates.

Rather than relying on providers' own description of their model — where the incentives to exaggerate may be strong, and activities may be defined in non-comparable ways across providers — we administered a survey module to teachers in all treatment schools, asking if they had heard of the provider, and if so, what activities the provider had engaged in. We summarize teachers' responses in Figure 1.2, which shows considerable variation in the specific activities and the total activity level of providers.

For instance, teachers reported that two providers (Omega and Bridge) frequently provided computers to schools, which fits with the stated approach of these two international, for-profit firms. Other providers, such as BRAC and Street Child, put slightly more focus on teacher training and observing teachers in the classroom, though these differences were not dramatic. In general, providers such as More than Me and Rising Academies showed high activity levels across dimensions, while teacher surveys confirmed administrative reports that Stella Maris conducted almost no activities in its assigned schools.

Cost data and assumptions

The government designed the PSL program based on the estimate that it spends roughly \$50 per child on teacher salaries in all public schools, and it planned to continue to do so in PSL schools (Werner 2017).[15] On top of this, providers would be offered a \$50 per-pupil payment to cover their costs.[16] This cost figure was chosen because \$100 was deemed a realistic medium-term goal for public expenditure on primary education nationwide (Werner 2017). To locate this in a global context, \$50 is about what was spent per primary pupil by governments in Sierra Leone in 2012, Burundi in 2005, the Central African Republic in 2006, or Guinea in 2008. \$100 is comparable to Lao PDR in 2010, Chad in 2010, Zambia in 2000, or Tanzania in 2007 (WB 2015b, 2015a).[17]

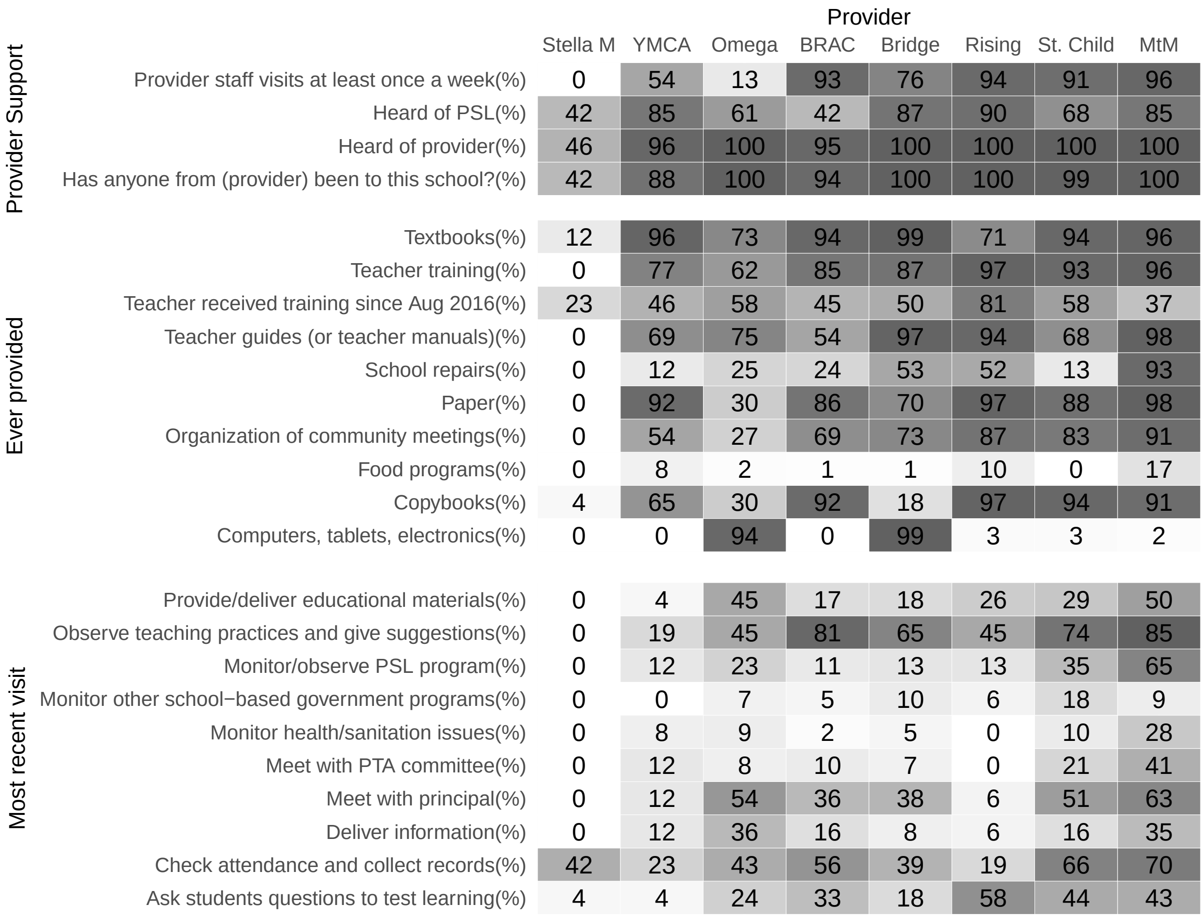
In the first year, providers spent far more than this amount.[18] *Ex ante* per-pupil

[15] As shown in Section 1.3, PSL led to reallocation of additional teaching staff to treatment schools and reduced pupil-teacher ratios in treatment schools, raising the Ministry's per-pupil cost to close to \$70.

[16] As noted above, neither Bridge International Academies nor Stella Maris received the extra \$50 per pupil.

[17] To make expenditures comparable across time, we transform all figures to 2010 US dollars.

[18] Several caveats apply to the cost figures here, which are our own estimates based on providers' self-reported budget data, and combine start-up costs, fixed costs, and variable costs. At the time of



		Provider							
		Stella M	YMCA	Omega	BRAC	Bridge	Rising	St. Child	MtM
Provider Support	Provider staff visits at least once a week(%)	0	54	13	93	76	94	91	96
	Heard of PSL(%)	42	85	61	42	87	90	68	85
	Heard of provider(%)	46	96	100	95	100	100	100	100
	Has anyone from (provider) been to this school?(%)	42	88	100	94	100	100	99	100
Ever provided	Textbooks(%)	12	96	73	94	99	71	94	96
	Teacher training(%)	0	77	62	85	87	97	93	96
	Teacher received training since Aug 2016(%)	23	46	58	45	50	81	58	37
	Teacher guides (or teacher manuals)(%)	0	69	75	54	97	94	68	98
	School repairs(%)	0	12	25	24	53	52	13	93
	Paper(%)	0	92	30	86	70	97	88	98
	Organization of community meetings(%)	0	54	27	69	73	87	83	91
	Food programs(%)	0	8	2	1	1	10	0	17
	Copybooks(%)	4	65	30	92	18	97	94	91
	Computers, tablets, electronics(%)	0	0	94	0	99	3	3	2
Most recent visit	Provide/deliver educational materials(%)	0	4	45	17	18	26	29	50
	Observe teaching practices and give suggestions(%)	0	19	45	81	65	45	74	85
	Monitor/observe PSL program(%)	0	12	23	11	13	13	35	65
	Monitor other school−based government programs(%)	0	0	7	5	10	6	18	9
	Monitor health/sanitation issues(%)	0	8	9	2	5	0	10	28
	Meet with PTA committee(%)	0	12	8	10	7	0	21	41
	Meet with principal(%)	0	12	54	36	38	6	51	63
	Deliver information(%)	0	12	36	16	8	6	16	35
	Check attendance and collect records(%)	42	23	43	56	39	19	66	70
	Ask students questions to test learning(%)	4	4	24	33	18	58	44	43

Figure 1.2: What did providers do?

The figure reports simple proportions (not treatment effects) of teachers surveyed in PSL schools who reported whether or not the provider responsible for their school had engaged in each of the activities listed. The sample size, n, of teachers interviewed with respect to each provider is: Stella Maris, 26; Omega, 141; YMCA, 26; BRAC, 170; Bridge, 157; Street Child, 80; Rising Academy, 31; More than Me, 46. This sample only includes compliant treatment schools.

budgets submitted to the program secretariat before the school year started (on top of the Ministry's costs) ranged from a low of approximately $57 for Youth Movement for Collective Action to a high of $1,050 for Bridge International Academies (see Figure 1.3a). *Ex post* per-pupil expenditure submited to the evaluation team at the end of

writing, the most comparable cost data we have access to are providers' *ex ante* budgets, rather than actual expenditures. Five providers submitted (self-reported) data to the evaluation team on actual expenditures at the end of the school year.

the school year (on top of the Ministry's costs) ranged from a low of approximately $48 for Street Child to a high of $663 for Bridge International Academies (see Figure 1.3b). These differences in costs are large relative to differences in treatment effects on learning, implying that cost-effectiveness may be driven largely by cost assumptions.

In principle, the costs incurred by private providers would be irrelevant for policy evaluation in a public-private partnership with this structure. If the providers are willing to make an agreement in which the government pays $50 per pupil, providers' losses are inconsequential to the government (philanthropic donors have stepped in to fund some providers' high costs under PSL).[19] Thus we present analyses in this report using both the Ministry's $50 long-term cost target and providers' actual budgets.[20]

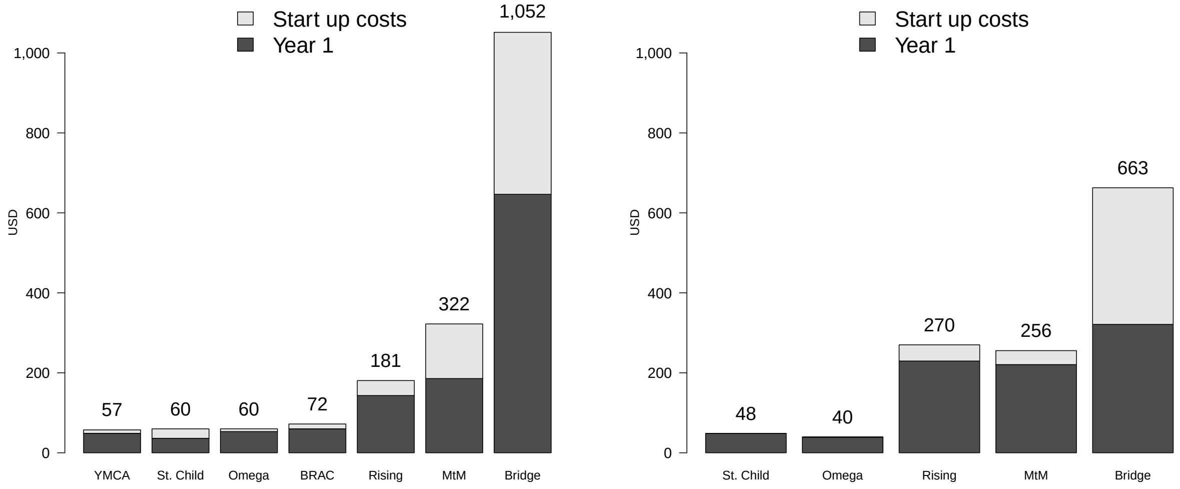
Providers' budgets for the first year of the program are likely a naïve measure of program cost, as these budgets combine start-up costs, fixed costs, and variable costs.[21] It is possible to distinguish start-up costs from the other costs as shown in Figure 1.3, and these make up a small share of the first-year totals for most providers. But it is not possible to distinguish fixed from variable costs in the current budget data. In informal interviews, some providers (e.g., Street Child) profess operating mostly a variable-cost model, implying that each additional school costs roughly the same amount to operate. Others (e.g., Bridge) report that their costs are almost entirely fixed, and unit costs

[19] These costs matter to the government under at least two scenarios. First, if providers are spending more during the first years of the program to prove effectiveness, they may lower expenditure (and quality) once they have locked in long-term contracts. Second, if private provider's aren't financially sustainable, they may suddenly close schools and disrupt student learning.

[20] While some providers relied almost exclusively on the $50 per child subsidy from the PSL pool fund, others have raised additional money from donors. Notably, Bridge International Academies relied entirely on direct grants from donors and opted not to participate in the competitive bidding process for the $50 per pupil subsidy which closed in June 2016. However, Bridge did subsequently submit an application for this funding in January 2017, which was not approved, but allows us access to their budget data. Bridge instead followed a bilateral memorandum of understanding (MOU) signed with the government of Liberia (Ministry of Education - Republic of Liberia 2016b). In practice, they operated as part of the larger PSL program. A noteworthy difference is that Bridge was authorized to cap class sizes somewhere between 45 and 55 students per class, while other providers were authorized to cap them at 65.

[21] Another possibility is that providers are spending more during the first years of the program to prove effectiveness, but will lower expenditure once they are locked in a long-term contract.

would fall precipitously if scaled; however, we have no direct evidence of this. Our best estimate is that Bridge's international operating cost, at scale, is between $191 and $220 per pupil annually.[22]



(a) Ex ante budget per pupil

(b) Ex post cost per pupil

Figure 1.3: Budget and costs as reported by providers

Note: Numbers in 1.3a are based on providers' ex-ante budgets, as submitted to the program secretariat in a uniform template (inclusive of both fixed and variable costs). Stella Maris did not provide budget data. Numbers in 1.3b are based on self-reported data on ex post expenditures (inclusive of both fixed and variable costs) submitted to the evaluation team by five providers in various formats. Numbers do not include the cost of teaching staff borne by the Ministry of Education.

[22] In written testimony to the UK House of Commons, Bridge stated that its fees were between $78 and $110 per annum in private schools, and that it had approximately 100,000 students in both private and PPP schools (Bridge International Academies 2017; Kwauk and Robinson 2016). Of these, roughly 9,000 are in PPP schools and pay no fees. In sworn oral testimony, Bridge co-founder Shannon May stated that the company had supplemented its fee revenue with more than $12 million in the previous year (May 2017). This is equivalent to an additional $120 per pupil, and implies Bridge spends between $191 and $220 per pupil at its current global scale.

1.2.2 Experimental design

Sampling and random assignment

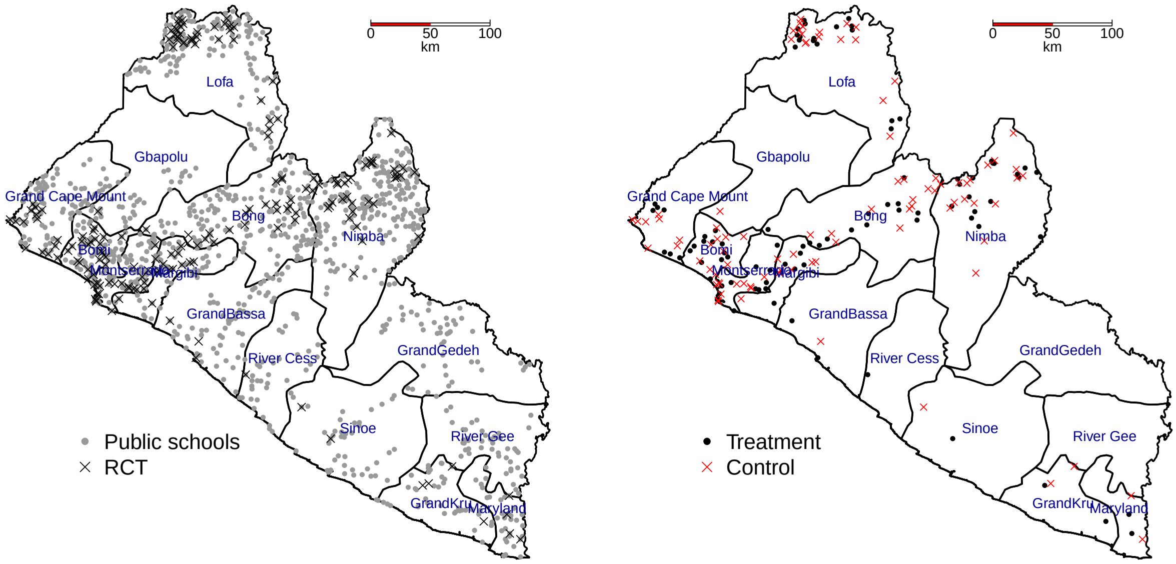
Liberia has 2,619 public primary schools. Private providers and the government agreed that potential PSL schools should have at least six classrooms and six teachers, good road access, a single shift, and should not contain a secondary school on their premises.[23] Only 299 schools satisfied all the criteria, although some of these are "soft" constraints that can be addressed if the program expands. For example, the government can build more classrooms and add more teachers to the school staff. On average, schools in the experiment are closer to the capital (Monrovia), have more students, greater resources, and better infrastructure.[24] Figure 1.4a shows all public schools in Liberia and those within our sample. Table A.1 in Appendix A.1 has details on the differences between schools in the experiment and other public schools.

Two providers, Omega Schools and Bridge International Academies, required schools with 2G connectivity. In addition, each provider submitted to the government a list of the regions they were willing to work in (Bridge International Academies had first pick of schools). Based on preferences and requirements the list of eligible schools was partitioned across providers. Then, we paired schools in the experiment sample within each district according to a principal component analysis (PCA) index of school resources.[25] This pairing stratified treatment by school resources within each private

[23] Additionally, a few schools were added to the list at the request of Bridge International Academies. Some of these schools had double shifts.

[24] While schools in the RCT generally have better facilities and infrastructure than most schools in the country, they still have deficiencies. For example, the average school in Liberia has 1.8 permanent classrooms — the median school has zero permanent classrooms — while the average school in the RCT has 3.16 classrooms.

[25] We calculated the index using the first eigenvector of a principal component analysis that included the following variables: students per teacher; students per classroom; students per chair; students per desk; students per bench; students per chalkboard; students per book; whether the school has a permanent building; whether the school has piped water, a pump or a well; whether the school has a toilet; whether the school has a staff room; whether the school has a generator; and the number of enrolled students.



(a) Geographical distribution of all public schools in Liberia and those within the RCT.

(b) Geographical distribution of treatment and control schools, original treatment assignment.

Figure 1.4: Public primary schools in Liberia

provider, but not across providers. We gave a list of "counterparts" to each provider based on their location preferences and requirements, so that each list had twice the number of schools they were to operate. Once each provider approved this list, we randomized the treatment assignment within each pair.[26] Appendix A.10 has details on the geographical distribution of the difference in school characteristics across providers. In short, schools are assigned to a provider, then paired, and then randomly assigned to treatment or control.

Private providers did not manage all the schools originally assigned to treatment and we treat them as non-compliant, presenting results in an intention-to-treat framework. After providers visited their assigned schools to start preparing for the upcoming school year, two treatment schools turned out to be private schools that were incorrectly labeled in the EMIS data as public schools. Two other schools had only two

[26] There is one triplet due to logistical constraints in the assignment of schools across counties, which resulted in one extra treatment school.

classrooms each. Of these four schools, two had originally been assigned to More Than Me and two had been assigned to Street Child. Omega Academies opted not to operate two of their assigned schools and Rising Academies opted not to operate one of their assigned schools. In short, there are 7 non-compliant treatment schools.[27] Figure 1.4b shows the treatment assignment.

Treatment assignment may change the student composition across schools. Thus, to prevent differences in the composition of students from driving differences in test scores, we sampled 20 students per school (from K1 to grade 5) from enrollment logs from 2015/2016, the year before the treatment was introduced. We associate each student with his or her "original" school, regardless of what school (if any) he or she attended in subsequent years. The combination of random treatment at the school level with sampling from a fixed and comparable pool of students allows us to provide clean estimates of the program's intention-to-treat (ITT) effect on test scores within the student population originally attending study schools, uncontaminated by selection.

Timeline of research and intervention activities

We collected data in schools twice: At the beginning of the school year in September/October 2016 and at the end of the school year in May/June 2017. A third round of data collection will take place in March/April 2019 conditional on continuation of the project and preservation of the control group (see Figure A.1 in Appendix A.1

[27] More than Me and Street Child were provided with replacement schools, presenting them with a new list of counterparts and informing them, as before, that they would operate one of each pair of schools (but not which one). Providers approved the list before we randomly assigned replacement schools from it. However, we do not use this list as our main sample since it is not fully experimental. We analyzed results for this "final" treatment and control school list, and they are almost identical to the results for the "original" list — perhaps unsurprisingly, given that they only differ by four pairs of schools. Results for this final list of treatment and control schools are available upon request. Bridge International Academies is managing two extra demonstration schools that were not randomized and are not part of our sample. Rising Academies was given one non-randomly assigned school, which is not part of our sample either. Therefore, the set of schools in our analysis is not identical to the set of schools actually managed by PSL providers. Table A.2 summarizes the overlap between schools in our main sample and the set of schools actually managed by PSL providers.

for a detailed timeline of intervention and research activities). We collected the first round of data 2 to 8 weeks after the beginning of treatment. While we intended the first survey wave to serve as a baseline, logistical delays led it to take place shortly after the beginning of the school year. We see evidence of treatment effects within this 1-2 month time frame and treat this early wave as a very short-term outcome survey. We do not use techniques like ANCOVA or difference-in-differences that consider these outcomes to be balanced.[28] We focus on fixed covariates and administrative data collected before the program began when checking balance between treatment and control schools to verify whether treatment was truly randomly assigned (see Section 1.2.2).

Test design

In our sample, literacy cannot be assumed at any grade level, precluding the possibility of written tests. We opted to conduct one-on-one tests in which an enumerator sits with the student, asks questions, and records the answers.[29] For the math portion of the test, we provided students with scratch paper and a pencil. We designed the tests to capture a wide range of student abilities. To make the test scores comparable across grades we constructed a single adaptive test for all students. The test has stop rules that skip higher-order skills if the student is not able to answer questions related to more basic skills. Appendix A.4 has details on the construction of the test.

[28] Our pre-analysis plan was written on the assumption we would be able to collect baseline data. Hence, the pre-analysis plan includes an ANCOVA specification along with the main specifications we use in this paper. We report these results in Table A.4 in Appendix A.1. We view the differences in short-term outcomes as treatment effects rather than "chance bias" in randomization for the following reasons. First, time-invariant student characteristics are balanced across treatment and control (see Table 1.2). Second, the effects on English and math test scores appear to materialize in the later weeks of the fieldwork, as shown in Figure A.2, consistent with a treatment effect rather than imbalance. Third, there is no significant effect on abstract reasoning, which is arguably less amenable to short-term improvements through teaching (although the difference between a significant English/math effect and an insignificant abstract reasoning effect here is not itself significant). We report the ANCOVA style specification results in Table A.4 in Appendix A.1.

[29] In addition, school-based tests would be contaminated by any effects arising from shifts in enrollment and attendance due to treatment.

We estimate an item response theory (IRT) model for each round of data collection.[30] IRT models are the standard in the assessments literature for generating comparative test scores.[31] There are two important and relevant characteristics of IRT models in this setting: First, they simultaneously estimate the test taker's ability and the difficulty of the questions, which allows the contribution of "correct answers" to the ability measure to vary from question to question. Second, they provide a comparable measure of student ability across different grades and survey rounds, even if the question overlap is imperfect. A common scale across grades allows us to estimate treatment effects as additional years of schooling. Following standard practice, we normalize the IRT scores with respect to the control group.

Additional data

We surveyed all the teachers in each school and conducted in-depth surveys with those teaching math and English. We asked teachers about their time use and teaching strategies. We also obtained teacher opinions on the PSL program. For a randomly selected class within each school, we conducted a classroom observation using the Stallings Classroom Observation Tool (World Bank 2015). Furthermore, we conducted school-level surveys to collect information about school facilities, the teacher roster, input availability (e.g., textbooks) and expenditures.

Enumerators collected information on some school practices. Specifically, enumerators recorded whether the school has an enrollment log and what information it

[30] The overlap between rounds of data collection is small, and therefore we do not estimate the same IRT model across rounds.

[31] For example, IRT models are used to estimate students' ability in the Graduate Record Examinations (GRE), the Scholastic Assessment Test (SAT), the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS) assessments. The use of IRT models in the development and education literature in economics is less prevalent, but becoming common: For example, see Das and Zajonc (2010), Andrabi et al. (2011), Andrabi, Das, and Khwaja (2017), Singh (2015b, 2016), Muralidharan, Singh, and Ganimian (2016), and Mbiti et al. (2017). Das and Zajonc (2010) provide a nice introduction to IRT models, while Linden (2017) provides a full treatment of IRT models.

stores; whether the school has an official time table and whether it is posted; whether the school has a parent-teacher association (PTA) and if the principal knows the PTA head's contact information (or where to find it); and whether the school has a written budget and keeps a record (and receipts) of past expenditures.[32] Additionally, we asked principals to complete two commonly used human resource instruments to measure individuals' "intuitive score" (Agor 1989) and "time management profile" (Schermerhorn et al. 2011).

For the second wave of data collection, we surveyed a random subset of households from our student sample, recording household characteristics and attitudes of household members. We also gathered data on school enrollment and learning levels for all children 4-8 years old living in these households.

Balance and attrition

As mentioned above, the first wave of data was collected 2 to 8 weeks after the beginning of treatment; hence, we focus on time-invariant characteristics (fixed covariates) when checking balance across treatment and control. Observable (time-invariant) characteristics of students and schools are balanced across treatment and control (see Table 1.2). Eighty percent of schools in our sample are in rural areas, over an hour away from the nearest bank (which is usually located in the nearest urban center); over 10% need to hold some classes outside due to insufficient classrooms. Boys make up 55% of our students and the students' average age is 12. According to pre-treatment administrative data (EMIS), the number of students, infrastructure, and resources available to students were not statistically different across treatment and control schools (for details, see Table A.3 in Appendix A.1).

[32] While management practices are difficult to measure, previous work has constructed detailed instruments to measure them in schools (e.g., see Bloom et al. (2015), Crawfurd (2017), and Lemos and Scur (2016)). Due to budget constraints, we checked easily observable differences in school management.

We took great care to avoid differential attrition: enumerators conducting student assessments participated in extra training on tracking and its importance, and dedicated generous time to tracking. Students were tracked to their homes and tested there when not available at school. Attrition in the second wave of data collection from our original sample is balanced between treatment and control and is below 4% overall (see Panel C). Appendix A.3 has more details on the tracking and attrition that took place in each round of data collection.

1.3 Experimental results

In this section, we first explore how the PSL program affected access to and quality of education. We then turn to mechanisms, looking at changes in material inputs, staffing, and school management.[33]

1.3.1 Test scores

Following our pre-analysis plan, we report treatment-effect estimates based on three specifications. The first specification amounts to a simple comparison of post-treatment outcomes for treatment and control individuals, in which Y_{isg} is the outcome of interest for student i in school s and group g (denoting the matched pairs used for randomization); α_g is a matched-pair fixed effect (i.e., stratification-level dummies); $treat_s$ is an indicator for whether school s was randomly chosen for treatment; and ε_{isg}

[33] A randomized controlled trial registry entry and the pre-analysis plan, are available at: https://www.socialscienceregistry.org/trials/1501.

Table 1.2: Balance: Observable, time-invariant school and student characteristics

	(1) Treatment	(2) Control	(3) Difference	(4) Difference (F.E)
Panel A: School characteristics (N = 185)				
Facilities (PCA)	-0.080	-0.003	-0.077	-0.070
	(1.504)	(1.621)	(0.230)	(0.232)
% holds some classes outside	13.978	14.130	-0.152	0.000
	(34.864)	(35.024)	(5.138)	(5.094)
% rural	79.570	80.435	-0.865	-0.361
	(40.538)	(39.888)	(5.913)	(4.705)
Travel time to nearest bank (mins)	75.129	68.043	7.086	7.079
	(69.099)	(60.509)	(9.547)	(8.774)
Panel B: Student characteristics (N = 3,496)				
Age in years	12.390	12.292	0.098	0.052
	(2.846)	(2.934)	(0.169)	(0.112)
% male	54.825	56.253	-1.427	-1.720
	(49.781)	(49.622)	(2.048)	(1.269)
Wealth index	-0.006	0.025	-0.031	0.010
	(1.529)	(1.536)	(0.140)	(0.060)
% in top wealth quartile	0.199	0.219	-0.020	-0.017
	(0.399)	(0.414)	(0.026)	(0.014)
% in bottom wealth quartile	0.266	0.284	-0.018	-0.012
	(0.442)	(0.451)	(0.039)	(0.019)
ECE before grade 1	0.834	0.820	0.014	0.013
	(0.372)	(0.384)	(0.025)	(0.017)
Panel C: Attrition in the second wave of data collection (N = 3,499)				
% interviewed	95.98	96.01	-0.03	-0.23
	(19.64)	(19.57)	(0.63)	(0.44)

Mean and standard error of the mean (in parentheses) for the control (Column 1) and treatment (Column 2). Difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including “pair” fixed effects) in Column 4. The school infrastructure index is made up of the first component in a principal component analysis of indicator variables for classrooms, staff room, student and adult latrines, library, playground, and an improved water source. The wealth index is the first component of a principal component analysis indicators for whether the student’s household has a television, radio, electricity, a refrigerator, a mattress, a motorbike, a fan, and a phone. Attrition rate is the proportion of students interviewed at the first round of data collection who we were unable to interview in the second wave. The standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

is an error term.

$$Y_{isg} = \alpha_g + \beta_{1.1} treat_s + \varepsilon_{isg} \tag{1.1}$$

$$Y_{isg} = \alpha_g + \beta_{1.2} treat_s + \gamma_{1.2} X_i + \delta_{1.2} Z_s + \varepsilon_{isg} \tag{1.2}$$

$$Y_{isg} = \alpha_g + \beta_{1.3} treat_s + \gamma_{1.3} X_i + \delta_{1.3} Z_s + \zeta_{1.3} Y_{isg,-1} + \varepsilon_{isg} \tag{1.3}$$

The second specification adds controls for time-invariant characteristics measured at the individual level (X_i) and school level (Z_s).[34] Finally, in equation (1.3) we use an ANCOVA specification (i.e., controlling for pre-treatment individual outcomes). However, as mentioned before, the first wave of data was collected after the beginning of treatment, so we lack a true baseline of student test scores. [35].

Table 1.3 shows results from student tests. The first three columns show differences between control and treatment schools' test scores after 1-2 months of treatment (September/October 2016), while the last three columns show the difference after 9-10 months of treatment (May/June 2017). After 1-2 months of treatment student test scores increase by 0.06σ in math (p-value=0.07) and 0.07σ in English (p-value=0.03). Part of these short-term improvements can be explained by the fact that most providers started the school year on time, while most traditional public schools began classes 1-4 weeks later. Hence, most students were already attending classes on a regular basis in treatment schools during our field visit, while their counterparts in control schools were not. In addition, we estimate the treatment effect separately for students tested during the first and the second half of the first round of data collection (see Figure A.2 in Appendix A.1), and show that the treatment effects fade in during the course of field work.

[34] These controls were specified in the pre-analysis plan and are listed in Table A.14.

[35] We report an ANCOVA-style specification in Table A.4 in Appendix A.1, and the results are still statistically significant, but mechanically downward biased.

In our preferred specification (Column 6) the treatment effect of PSL after one academic year is .18σ for English (p-value < 0.001) and .18σ for math (p-value < 0.001). We focus on the ITT effect, but the treatment-on-the-treated (ToT) effect (i.e., the treatment effect only for students that actually attended a PSL school in 2016/2017) can be computed using the fraction of students originally assigned to treatment schools who are actually in treatment schools at the end of the 2016/2017 schools year (77%) and the fraction of students assigned to control schools who are in treatment schools at the end of the 2016/2017 schools year (0%). For details, see Table A.6 in Appendix A.1 which shows both the ITT and the ToT. Our results are robust to different measures of student ability (see Table A.7 in Appendix A.1 for details).

An important concern when interpreting these results is whether they represent real gains in learning or better test-taking skills resulting from "teaching to the test". We show suggestive evidence that these results represent real gains. First, the treatment effect over new modules that were not in the first wave test (and unknown to the providers or the teachers) is significant (.19σ, p-value < 0.001), and statistically indistinguishable from the treatment effect over all the items (.19σ, p-value < 0.001). Second, the treatment effect over the conceptual questions (which do not resemble the format of standard textbook exercises) is positive and significant (.12σ, p-value .0013). However, we cannot rule out that providers narrowed the curriculum by focusing on English and mathematics or, conversely, that they generated learning gains in other subjects that we did not test. We find no evidence of heterogeneity by students' socio-economic status, gender, or grade (see Table A.5 in Appendix A.1).

Although reporting the impact of interventions in standard deviations is the norm in the education and experimental literature, we also report results as "equivalent years of schooling" (EYOS) following Evans and Yuan (2017). Results in this format

Table 1.3: ITT treatment effects on learning

	First wave (1-2 months after treatment)			Second wave (9-10 months after treatment)		
	Difference	F.E.	Controls	Difference	F.E.	Controls
	(1)	(2)	(3)	(4)	(5)	(6)
English	0.05	0.09*	0.07**	0.17**	0.17***	0.18***
	(0.08)	(0.05)	(0.03)	(0.08)	(0.04)	(0.03)
Math	0.08	0.08*	0.06*	0.17***	0.19***	0.18***
	(0.07)	(0.04)	(0.03)	(0.07)	(0.04)	(0.03)
Abstract	0.04	0.05	0.04	0.05	0.05	0.05
	(0.06)	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)
Composite	0.07	0.08*	0.06*	0.17**	0.19***	0.19***
	(0.07)	(0.05)	(0.03)	(0.07)	(0.04)	(0.03)
New				0.17**	0.20***	0.19***
				(0.07)	(0.04)	(0.04)
Conceptual				0.12**	0.13***	0.12***
				(0.05)	(0.04)	(0.04)
Observations	3,496	3,496	3,496	3,492	3,492	3,492

Columns 1-3 are based on the first wave of data and show the difference between treatment and control (Column 1), and the difference taking into account the randomization design — i.e., including "pair" fixed effects — (Column 2), and the difference taking into account other student and school controls (Column 3). Columns 4-6 are based on the second wave of data and show the difference between treatment and control (Column 4) in test scores, the difference taking into account the randomization design — i.e., including "pair" fixed effects — (Column 5), and the difference taking into account other student and school controls (Column 6).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

are easier to communicate to policymakers and the general public, by juxtaposing treatment effects with the learning from business-as-usual schooling. In our data the average increase in test scores for each extra year of schooling in the control group is .31σ in English and .28σ in math. Thus, the treatment effect is roughly 0.56 EYOS for English and 0.65 EYOS for math. See Appendix A.8 for a detailed explanation of the methodology to estimate EYOS, and a comparison of EYOS and standard deviation across countries. Additionally, Appendix A.9 shows absolute learning levels in treatment and control schools for a subset of the questions that are comparable to other settings, to allow direct comparisons with learning levels in other countries. Despite the positive treatment effect of the program, students in treatment schools are still behind their international peers.

1.3.2 Enrollment, attendance, and student selection

The previous section showed that education quality, measured in an ITT framework using test scores, increases in PSL schools. We now ask whether the PSL program increases access to education. To explore this question we focus on three outcomes which were committed to in the pre-analysis plan: Enrollment, student attendance, and student selection. PSL increased enrollment overall, but in schools where enrollment was already high and classes were large, the program led to a significant decline in enrollment. This does not appear to be driven by selection of "better" students, but by providers capping class sizes and eliminating double shifts.[36] As shown in Section 1.5.4, almost the entirety of this phenomenon is explained by Bridge International Academies.

Enrollment changes across treatment and control schools are shown in Panel A of Table 1.4. There are a few noteworthy items. First, treatment schools are slightly larger before treatment: They have 34 (p-value .094) students more on average before treatment.[37] Second, PSL schools have on average 57 (p-value < 0.001) more students than control schools in the 2016/2017 academic year, which results in a net increase (after controlling for pre-treatment differences) of 25 (p-value .088) students per school.[38]

[36] Three Bridge International Academies treatment schools (representing 28% of total enrollment in Bridge treatment schools) had double shifts in 2015/2016, but not in 2016/2017. One Omega Schools treatment school (representing 7.2% of total enrollment in Omega treatment schools) had double shifts in 2015/2016, but not in 2016/2017. The MOU between Bridge and the Ministry of Education explicitly authorized eliminating double shifts (Ministry of Education - Republic of Liberia 2016b).

[37] Table A.3 uses EMIS data, while Table 1.4 uses data independently collected by IPA. While the difference in enrollment in the 2015/2016 academic year is only significant in the latter, the point estimates are remarkably similar across both tables.

[38] Once the EMIS data for the 2016/2017 school year are released, we will reexamine this issue to study whether increases in enrollment come from children previously out-of-school or from children previously enrolled in other schools.

Since provider compensation is based on the number of students enrolled rather than the number of students actively attending school, increases in enrollment may not translate into increases in student attendance. An independent measure of student attendance conducted by our enumerators during a spot check shows that students are 16 (p-value < 0.001) percentage points *more* likely to be in school during class time in treatment schools (see Panel A, Table 1.4).

Turning to the question of student selection, we find no evidence that any group of students is systematically excluded from PSL schools. The proportion of students with disabilities is not statistically different in PSL schools and control schools (Panel A, Table 1.4).[39] Among our sample of students (i.e., students sampled from the 2015/2016 enrollment log), students are equally likely across treatment and control to be enrolled in the same school in the 2016/2017 academic year as they were in 2015/2016, and equally likely to be enrolled in any school (see Panel B, Table 1.4). Finally, selection analysis using student-level data on wealth, gender, and age finds no evidence of systematic exclusions (see Table A.8 in Appendix A.1).

Providers are authorized to cap class sizes, which could lead to students being excluded from their previous school (and either transferred to another school or to no school at all). We estimate whether the caps are binding for each student by comparing the average enrollment prior to treatment in her grade cohort and the two adjacent grade cohorts (i.e., one grade above and below) to the theoretical class-size cap under PSL. We average over three cohorts because some providers used placement tests to reassign students across grade levels. Thus the "constrained" indicator is defined by the number of students enrolled in the student's 2016/2017 "expected grade" (as predicted

[39] The fraction of students identified as disabled in our sample is an order of magnitude lower than estimates for the percentage of disabled students in the U.S and worldwide using roughly the same criteria (both about 5%) (Brault 2011; UNICEF 2013).

Table 1.4: ITT treatment effects on enrollment, attendance, and selection

	(1) Treatment	(2) Control	(3) Difference	(4) Difference (F.E)
Panel A: School level data (N = 175)				
Enrollment 2015/2016	298.45	264.11	34.34	34.18*
	(169.74)	(109.91)	(21.00)	(20.28)
Enrollment 2016/2017	309.71	252.75	56.96***	56.89***
	(118.96)	(123.41)	(18.07)	(16.29)
15/16 to 16/17 enrollment Δ	11.55	-6.06	17.61	24.60*
	(141.30)	(82.25)	(17.19)	(14.35)
Attendance % (spot check)	48.02	32.84	15.18***	15.56***
	(24.52)	(26.54)	(3.81)	(3.13)
% of students with disabilities	0.59	0.39	0.20	0.21
	(1.16)	(0.67)	(0.14)	(0.15)
Panel B: Student level data (N = 3,627)				
% enrolled in the same school	80.74	83.34	-2.61	0.79
	(39.45)	(37.27)	(3.67)	(2.07)
% enrolled in school	94.14	94.00	0.14	1.22
	(23.49)	(23.76)	(1.33)	(0.87)
Days missed, previous week	0.85	0.85	-0.00	-0.06
	(1.42)	(1.40)	(0.10)	(0.07)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) and treatment (Column 2) groups, as well as the difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 4. Our enumerators conducted the attendance spot check in the middle of a school day. If the school was not in session during a regular school day we mark all students as absent. Standard errors are clustered at the school level. The sample is the original treatment and control allocation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

based on normal progression from their 2015/2016 grade) and adjacent grades, divided by the "maximum capacity" in those three grades in 2016/2017 (as specified in our pre-analysis plan):

$$c_{igso} = \frac{Enrollment_{is,g-1} + Enrollment_{is,g} + Enrollment_{is,g+1}}{3 * Maximum_o},$$

where c_{igso} is our "constrained" measure for student i, expected to be in grade g in 2016/2017, at school s, in a "pair" assigned to provider o. $Enrollment_{is,g-1}$ is enrollment in the grade below the student's expected grade, $Enrollment_{is,g}$ is enrollment in the student's expected grade, and $Enrollment_{is,g+1}$ is enrollment in the grade above the student's expected grade. $Maximum_o$ is the class cap approved for provider o. We label a grade-school combination as "constrained" if $c_{igso} > 1$.

Enrollment in constrained school-grades decreases, while enrollment in unconstrained school-grades increases (see Column 1 in Table 1.5). Thus, schools far below the cap have positive treatment effects on enrollment and schools near or above the cap offset it with declining enrollment. Our student data reveal this pattern as well: Columns 2 and 3 in Table 1.5 show the ITT effect on enrollment depending on whether students were enrolled in a constrained class in 2015/2016. In unconstrained classes students are more likely to be enrolled in the same school (and in any school). But in constrained classes students are less likely to be enrolled in the same school. While there is no effect on overall school enrollment, switching schools may be disruptive for children (Hanushek, Kain, and Rivkin 2004). Finally, test-scores improve for students in constrained classes. This result is difficult to interpret as it includes the positive treatment effect over students who did not change schools (possibly compounded by smaller class sizes) with the effect over students removed from their schools. These results are robust to excluding adjacent grades from the "constrained" measure (see Table A.9 in Appendix A.1).

Table 1.5: ITT treatment effects, by whether class size caps are binding

	(1) Δ enrollment	(2) % same school	(3) % in school	(4) Test scores
Constrained=0 × Treatment	5.30***	4.04***	1.64**	0.15***
	(1.11)	(1.39)	(0.73)	(0.034)
Constrained=1 × Treatment	-11.7*	-12.8	0.070	0.35***
	(6.47)	(7.74)	(4.11)	(0.11)
No. of obs.	1,635	3,625	3,485	3,490
Mean control (Unconstrained)	-0.75	82.09	93.38	0.13
Mean control (Constrained)	-7.73	84.38	94.81	-0.08
α_0 :Constrained-Unconstrained	-17.05	-16.79	-1.57	0.20
p-value ($H_0 : \alpha_0 = 0$)	0.01	0.03	0.71	0.07

Column 1 uses school-grade level data. Columns 2 - 4 use student level data. The independent variable in Column 4 is the composite test score. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. There were 194 constrained classes before treatment (holding 30% of students), and 1,468 unconstrained classes before treatment (holding 70% of students).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.3.3 Intermediate inputs

In this section we explore the effect of the PSL program on school inputs (including teachers), school management (with a special focus on teacher behavior and pedagogy), and parental behavior.

Inputs and resources

Teachers, one of the most important inputs of education, change in several ways (see Panels A/B in Table 1.6). PSL schools have 2.6 more teachers on average (p-value < 0.001), but this is not merely the result of operators hiring more teachers. Rather, the Ministry of Education agreed to release some underperforming teachers from PSL schools,[40] replace those teachers, and provide additional ones. Ultimately, the extra teachers result in lower pupil-teacher ratios (despite increased student enrollment). This re-shuffling of teachers means that PSL schools have younger and less-experienced teachers, who are more likely to have worked in private schools in the past and have higher test scores (we conducted a simple memory, math, word association, and abstract thinking test).[41] While the program's contracts made no provisions to pay teachers differently in treatment and control schools, teachers in PSL schools report higher wages. However large unconditional increases in teacher salaries have been shown elsewhere to have no effect on student performance in the short run (Ree et al. 2015).

Our enumerators conducted a "materials" check during classroom observations (See Panels C - Table 1.6). Since we could not conduct classroom observations in schools that were out of session during our visit, Table A.10 in Appendix A.1 presents Lee (2009) bounds on these treatment effects (control schools are more likely to be out of session). Conditional on the school being in session during our visit, students in PSL schools are 23 percentage points (p-value < 0.001) more likely to have a textbook and 8.2 percentage points (p-value .049) more likely to have writing materials (both a pen and a copybook). However, we cannot rule out that there is no overall effect as zero is between the Lee (2009) bounds.

School management

Two important management changes are shown in Table 1.7: PSL schools are 8.7 percentage points more likely to be in session (i.e., the school is open, students

[40] Once the EMIS data for the 2016/2017 school year are released, we will reexamine this issue to study whether teachers who were fired were allocated to other public schools. While the majority of released teachers are on the government's payroll, some of the dismissed teachers are thus they have not necessarily been assigned to other public schools.

[41] Replacement and extra teachers are recent graduates from the Rural Teacher Training Institutes. See King et al. (2015) for details on this program.

Table 1.6: ITT treatment effects on inputs and resources

	(1)	(2)	(3)	(4)
	Treatment	Control	Difference	Difference
				(F.E)
Panel A: School-level outcomes (N = 185)				
Number of teachers	9.62	7.02	2.60***	2.61***
	(2.82)	(3.12)	(0.44)	(0.37)
Pupil-teacher ratio (PTR)	32.20	39.95	-7.74***	-7.82***
	(12.29)	(18.27)	(2.31)	(2.12)
New teachers	4.81	1.77	3.03***	3.01***
	(2.56)	(2.03)	(0.34)	(0.35)
Teachers dismissed	3.35	2.17	1.18**	1.16**
	(3.82)	(2.64)	(0.48)	(0.47)
Panel B: Teacher-level outcomes (N = 1,167)				
Age in years	39.09	46.37	-7.28***	-7.10***
	(11.77)	(11.67)	(1.02)	(0.68)
Experience in years	10.59	15.79	-5.20***	-5.26***
	(9.20)	(10.77)	(0.76)	(0.51)
% has worked at a private school	47.12	37.50	9.62**	10.20***
	(49.95)	(48.46)	(3.76)	(2.42)
Test score in standard deviations	0.13	-0.01	0.14*	0.14**
	(1.02)	(0.99)	(0.07)	(0.06)
% certified (or tertiary education)	60.11	58.05	2.06	4.20
	(48.99)	(49.39)	(4.87)	(2.99)
Salary (USD/month) $\|$salary> 0	121.36	104.54	16.82**	13.90***
	(44.42)	(60.15)	(6.56)	(4.53)
Panel C: Classroom observation (N = 143)				
	(15.43)	(20.26)	(2.94)	(2.61)
% with chalk	96.39	78.87	17.51***	16.58***
	(18.78)	(41.11)	(5.29)	(5.50)
% of students with textbooks	37.08	17.60	19.48***	22.60***
	(43.22)	(35.25)	(6.33)	(6.32)
% of students with pens/pencils	88.55	79.67	8.88**	8.16**
	(19.84)	(30.13)	(4.19)	(4.10)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) and treatment (Column 2) groups, as well as the difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 4. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

and teachers are on campus, and classes are taking place) during a regular school day (p-value .057), and have a longer school day that translates into 3.9 more hours per week of instructional time (p-value < 0.001). In addition, although principals in PSL schools have scores in the "intuitive" and "time management profile" scale that are almost identical to their counterparts in traditional public schools, they spend more of their time on management-related activities (e.g., supporting other teachers, monitoring student progress, meeting with parents) than actually teaching, suggesting a change in the role of the principal in these schools — perhaps as a result of additional teachers, principals in PSL schools did not have to double as teachers. Additionally, management practices (as measured by a PCA index[42] normalized to a mean of zero and standard deviation of one in the control group) are $.4\sigma$ (p-value < 0.001) higher in PSL schools. This effect size can be viewed as a boost for the average treated school from the 50th to the 66th percentile in management practices.

Teacher behavior

An important component of school management is teacher accountability and its effects on teacher behavior. As mentioned above, teachers in PSL schools are drawn from the pool of unionized civil servants with lifetime appointments and are paid directly by the Liberian government. In theory, private providers have limited authority to request teacher reassignments and no authority to promote or dismiss civil service teachers. Thus, a central hypothesis underlying the PSL program is that providers can hold teachers accountable through monitoring and support, rather than rewards and

[42] The index includes whether the school has an enrollment log and what information is in it, whether the school has an official time table and whether it is posted, whether the school has a parent-teacher association (PTA) and whether the principal has the PTA head's number at hand, and whether the school keeps a record of expenditures and a written budget. Table A.11 has details on every component of the good practices index.

Table 1.7: ITT treatment effects on school management

	(1)	(2)	(3)	(4)
	Treatment	Control	Difference	Difference
				(F.E)
% school in session	92.47	83.70	8.78^{*}	8.66^{*}
	(26.53)	(37.14)	(4.75)	(4.52)
Instruction time (hrs/week)	20.40	16.50	3.90^{***}	3.93^{***}
	(5.76)	(4.67)	(0.77)	(0.73)
Intuitive score (out of 12)	4.08	4.03	0.04	0.02
	(1.35)	(1.38)	(0.20)	(0.19)
Time management score (out of 12)	5.60	5.69	-0.09	-0.10
	(1.21)	(1.35)	(0.19)	(0.19)
Working time (hrs/week)	21.43	20.60	0.83	0.84
	(11.83)	(14.45)	(1.94)	(1.88)
% of time spent on management	74.06	53.64	20.42^{***}	20.09^{***}
	(27.18)	(27.74)	(4.12)	(3.75)
Index of good practices (PCA)	0.41	-0.00	0.41^{***}	0.40^{***}
	(0.64)	(1.00)	(0.12)	(0.12)
Observations	93	92	185	185

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) and treatment (Column 2) groups, as well as the difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 4. Intuitve score is measured using Agor (1989)'s instrument and time management profile using Schermerhorn et al. (2011)'s instrument. The index of good practices is the first component of a principal component analysis of the variables in Table A.11. The index is normalized to have mean zero and standard deviation of one in the control group. Standard errors are clustered at the school level. The sample is the original treatment and control allocation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

threats.[43]

To study teacher behavior, we conducted unannounced spot checks of teacher attendance and collected student reports of teacher behavior (see Panels A/B in Table 1.8). Also, during these spot checks we used the Stallings classroom observation instrument to study teacher time use and classroom management (see Panel C in Table 1.8).

[43] While providers could have provided teachers with performance incentives, we have no evidence that any of them did.

Teachers in PSL schools are 20 percentage points (p-value < 0.001) more likely to be in school during a spot check (from a base of 40%) and the unconditional probability of a teacher being in a classroom increases by 15 percentage points (p-value < 0.001). Our spot checks align with student reports on teacher behavior. According to students, teachers in PSL schools are 7.6 percentage points (p-value < 0.001) less likely to have missed school the previous week. In addition, students in PSL schools also report that teachers are 6.6 percentage points (p-value .0099) less likely to hit them.

Classroom observations also show changes in teacher behavior and pedagogical practices. First, teachers in PSL schools are 15 percentage points (p-value .0023) more likely to engage in either active instruction (e.g., teacher engaging students through lecture or discussion) or passive instruction (e.g., students working in their seat while the teacher monitors progress) and 25 percentage points (p-value < 0.001) less likely to be off-task.[44] Although these are considerable improvements, the treatment group is still far off the Stallings, Knight, and Markham (2014) good practice benchmark of 85 percent of total class time used for instruction, and below the average time spent on instruction across five countries in Latin America (Bruns and Luque 2014).

These estimates combine the effects on individual teacher behavior with changes to teacher composition. To estimate the treatment effect on teacher attendance over a fixed pool of teachers, we perform additional analyses in Appendix A.1 using administrative data (EMIS) to restrict our sample to teachers who worked at the school the year before the intervention began (2015/2016). We treat teachers who no longer worked at the school in the 2016/2017 school year as (non-random) attriters and estimate Lee (2009) bounds on the treatment effect. Table A.10 in Appendix A.1 shows

[44] See Stallings, Knight, and Markham (2014) for more details on how active and passive instruction, as well as time off-task and student engagement, are coded.

Table 1.8: ITT treatment effects on teacher behavior

	(1) Treatment	(2) Control	(3) Difference	(4) Difference (F.E)
Panel A: Spot checks (N = 185)				
% on schools campus	60.32	40.38	19.94***	19.79***
	(23.10)	(25.20)	(3.56)	(3.48)
% in classroom	47.02	31.42	15.60***	15.37***
	(26.65)	(25.04)	(3.80)	(3.62)
Panel B: Student reports about teachers (N = 185)				
Missed school previous week (%)	17.69	25.12	-7.43***	-7.55***
	(10.75)	(14.92)	(1.91)	(1.94)
Never hits students (%)	54.71	48.21	6.50**	6.56***
	(18.74)	(17.06)	(2.63)	(2.52)
Helps outside the classroom (%)	50.00	46.59	3.41	3.55
	(18.22)	(18.05)	(2.67)	(2.29)
Panel C: Classroom observations (N = 185)				
Instruction (% of class time)	49.68	35.00	14.68***	14.51***
	(32.22)	(37.08)	(5.11)	(4.70)
Class management (% class time)	19.03	8.70	10.34***	10.25***
	(20.96)	(14.00)	(2.62)	(2.73)
Teacher off-task (% class time)	31.29	56.30	-25.01***	-24.77***
	(37.71)	(42.55)	(5.91)	(5.48)
Student off-task (% class time)	50.41	47.14	3.27	2.94
	(33.51)	(38.43)	(5.30)	(4.59)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) and treatment (Column 2) groups, as well as the difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 4. Our enumerators conducted the attendance spot check in the middle of a school day. If the school was not in session during a regular school day we mark all teachers not on campus as absent and teachers and students as off-task in the classroom observation. Table A.10 has the results without imputing values. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

an ITT treatment effect of 14 percentage points (p-value < 0.001) on teacher attendance. Importantly, zero is not part of the Lee (2009) bounds for this effect. This aligns with previous findings showing that management practices have significant effects on worker performance (Bloom et al. 2014; Bloom et al. 2013; Bennedsen et al. 2007).

1.3.4 Other outcomes

Student data (Table 1.9, Panel C) and household data (Table 1.9, Panel A) show that the program increases both student and parental satisfaction. Students in PSL schools are happier (measured by whether they think going to school is fun or not), and parents with children in PSL schools (enrolled in 2015/2016) are 7.4 percentage points (p-value .022) more likely to be satisfied with the education their children are receiving. Table A.21 in Appendix A.6 has detailed data on student, parental, and teacher support and satisfaction with PSL.

Providers are not allowed to charge fees and PSL should be free at all levels, including early-childhood education (ECE) for which fees are normally permitted in government schools. We interviewed both parents and principals regarding fees. In both treatment and control schools parents are more likely to report paying fees than schools are to report charging them. Similarly, the amount parents claim to pay in school fees is much higher than the amount schools claim to charge (see Panel A and Panel B in Table 1.9). Since principals may be reluctant to disclose the full amount they charge parents, especially in primary school (which is nominally free), this discrepancy is normal. While the likelihood of charging fees decreases in PSL schools by 26 percentage points according to parents and by 19 percentage points according to principals, 48% of parents still report paying some fees in PSL schools.

On top of reduced fees, providers often provide textbooks and uniforms free of charge to students (see Section 1.2.1). Indeed, household expenditures on fees, textbooks, and uniforms drop (see Table A.12 for details). In total, household expenditures on children's education decrease by 6.7 USD (p-value .1) in PSL schools.

A reduction in household expenditure in education reflects a crowding out response (i.e., parents decrease private investment in education as school investments increase). To explore whether crowding out goes beyond expenditure, we ask parents about engagement in their child's education, but see no change in this margin (we summarize parental engagement using the first component from a principal component analysis across several measures of parental engagement; see Table A.13 for the effect on each component).

To complement the effect of the program on cognitive skills, we study student attitudes and opinions (see Table 1.10). Some of the control group rates are noteworthy: 50% of children use what they learn in class outside school, 69% think that boys are smarter than girls, and 79% think that some tribes in Liberia are bad. Turning to treatment effects, children in PSL schools are more likely to think school is useful, more likely to think elections are the best way to choose a president, and less likely to think some tribes in Liberia are bad. The effect on tribe perceptions is particularly important in light of the recent conflict in Liberia and the ethnic tensions that sparked it. Our results also align with previous findings from Andrabi et al. (2010), who show that children in private schools in Pakistan are more "pro-democratic" and exhibit lower gender biases (we do not find any evidence of lower gender biases in this setting). Note, however, that our treatment effects are small in magnitude. It is also impossible to tease out the effect of who is providing education from the effect of better education, and the effect of younger and better teachers. Hence, our results show the net change in students' opinions, and cannot be attributed to providers per se but rather to the program as a whole.

Table 1.9: ITT treatment effects on household behavior and fees

	(1) Control	(2) Difference	(3) Difference (F.E)
Panel A: Household behavior (N = 1,115)			
% satisfied with school	67.46	7.42**	7.44**
	(23.95)	(3.20)	(3.23)
% paying any fees	73.56	-25.45***	-25.69***
	(44.14)	(4.73)	(3.26)
Fees (USD/year)	8.04	-2.32**	-2.89***
	(9.73)	(0.96)	(0.61)
Expenditure (USD/year)	73.61	-8.09	-6.74
	(79.53)	(6.96)	(4.13)
Engagement index (PCA)	-0.09	-0.02	-0.03
	(0.91)	(0.07)	(0.06)
Panel B: Fees (N = 184)			
% with $>$ 0 ECE fees	30.77	-18.94***	-18.98***
	(46.41)	(5.92)	(5.42)
% with $>$ 0 primary fees	29.67	-16.77***	-16.79***
	(45.93)	(5.95)	(5.71)
ECE Fee (USD/year)	1.42	-0.85**	-0.87***
	(2.78)	(0.35)	(0.33)
Primary Fee (USD/year)	1.22	-0.68**	-0.70**
	(2.40)	(0.31)	(0.31)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) group, as well as the difference between treatment and control (Column 2), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 3. Standard errors are clustered at the school level. The index for parent engagement is the first component from a principal component analysis across several measures of parental engagement (see Table A.13). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.4 Unbundling the treatment effect

The question of mechanisms can be divided into two parts: What changed? And which changes mattered for learning outcomes? We answered the first question in the previous section. In this section we use non-experimental variation to answer the latter question. The key assumption underlying these results is that we can identify the casual effect of intermediate inputs on learning in the absence of experimental variation

Table 1.10: ITT treatment effects on student attitudes

	(1) Control	(2) Difference	(3) Difference (F.E)
School is fun	0.53	0.05**	0.05**
	(0.50)	(0.02)	(0.02)
I use what I'm learning outside of school	0.49	0.04	0.04***
	(0.50)	(0.02)	(0.02)
If I work hard, I will succeed.	0.55	0.05*	0.04***
	(0.50)	(0.03)	(0.02)
Best way to choose president:Elections	0.88	0.03*	0.03***
	(0.33)	(0.01)	(0.01)
Boys are smarter than girls	0.69	-0.00	0.01
	(0.46)	(0.02)	(0.01)
Some tribes in Liberia are bad	0.79	-0.03	-0.03**
	(0.41)	(0.02)	(0.01)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) group, as well as the difference between treatment and control (Column 2), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 3. Standard errors are clustered at the school level. The index for parent engagement is the first component from a principal component analysis across several measures of parental engagement (see Table A.13). N = 3,492. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

in these inputs across schools.

There are three related goals in the analysis below: (i) to highlight which mechanisms correlate with learning gains; (ii) to uncover how much of the treatment effect is the result of an increase in resources (e.g., teachers and per-child expenditure); and (iii) to estimate whether PSL schools are more productive (i.e., whether they use resources more effectively to generate learning). To attain these goals we use mediation analysis, and follow the general framework laid out in Imai, Keele, and Yamamoto (2010) and Imai, Keele, and Tingley (2010).[45]

The mediation effect of a learning input (e.g., teacher attendance) is the change in learning gains that can be attributed to changes in this input caused by treatment.

[45] This framework is closely related to the framework used by Heckman, Pinto, and Savelyev (2013) and Heckman and Pinto (2015). There is a direct mapping between the two.

Formally, we can estimate the mediation effect via the following two equations:

$$\begin{aligned} M_{isg} &= \alpha_g + \beta_{1.4} treat_s + \gamma_{1.4} X_i + \delta_{1.4} Z_s + u_{isg} \qquad (1.4) \\ Y_{isg} &= \alpha_g + \beta_{1.5} treat_s + \gamma_{1.5} X_i + \delta_{1.5} Z_s + \theta_{1.5} M_{isg} + \varepsilon_{isg}, \qquad (1.5) \end{aligned}$$

in which Y_{isg} is the test score for student i in school s and group g (denoting the matched pairs used for randomization); α_g is a matched-pair fixed effect (i.e., stratification-level dummies); $treat_s$ is an indicator for whether school s was randomly chosen for treatment; and ε_{isg} and u_{isg} are error terms. X_i and Z_s are individual and school-level time-invariant controls, while M_{isg} are the potential mediators for treatment (i.e., learning inputs measured during the second wave of data collection). Equation 1.4 is used to estimate the effect of treatment on the mediator ($\beta_{1.4}$), while equation 1.5 is used to estimate the effect of the mediator on learning ($\theta_{1.5}$).

The mediation effect is $\beta_{1.4} \times \theta_{1.5}$, i.e., the effect of the mediator on learning gains ($\theta_{1.5}$) combined with changes in the mediator caused by treatment ($\beta_{1.4}$). $\beta_{1.5}$ captures the treatment effect that is not mediated by M_{isg}. $\beta_{1.5}$ is often refereed to as the "direct effect", but it can be a treatment effect mediated by unmeasured mediators. The mediation effect ($\beta_{1.4} \times \theta_{1.5}$) and the direct effect ($\beta_{1.5}$) are in the same units (the units of Y_{isg}), and are therefore comparable.

The crux of a mediation analysis is to get consistent estimators of $\theta_{1.5}$ (and therefore of $\beta_{1.5}$). Imai, Keele, and Yamamoto (2010) show that the OLS estimators for $\beta_{1.5}$ and $\theta_{1.5}$ are consistent under the following assumption:

Assumption 1 (Sequential ignorability)**.**

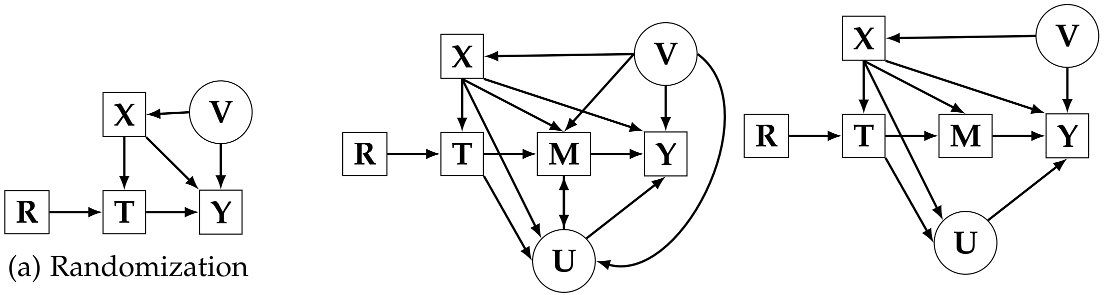
$$Y_i(t,m), M_i(t) \perp\!\!\!\perp T_i | X_i = x \tag{1.6}$$

$$Y_i(t,m) \perp\!\!\!\perp M_i(t) | X_i = x, T_i = t \tag{1.7}$$

where $Y_i = Y_i(t,m)$ denotes the potential outcome for individual i under treatment t and mediators m, $M_i(t)$ denotes the potential mediator for individual i under treatment t; $Pr(T_i = t|X_i = x) > 0$; and $Pr(m_i(t) = m|T_i = t, X_i = x) > 0$ for all values of t, x and m.

Figure 1.5 shows the difference between a randomization model without mediation (1.5a), a mediation model with all the possible causal relationships (1.5b), and a mediation model under assumption 1 (1.5c). Randomization guarantees that there is no causal relationship between unobserved variables and treatment status (there is no arrow between V and T). Once mediators are included, these may be correlated to unobserved variables (including unobserved or unmeasured mediators). Assumption 1 implies that unobserved variables do not cause changes in inputs (once observable variables are taken into account), and that there is no relationship between unmeasured and measured mediators (i.e., there are no arrows from V to neither M or U, and there are no arrows between M and U).

While randomization implies that equation 1.6 in Assumption 1 is met, we do not have experimental variation in any of the possible mediators and thus unobserved variables may confound the relationship between mediators and learning gains, violating equation 1.7 in Assumption 1 (Green, Ha, and Bullock 2010; Bullock and Ha 2011). To mitigate omitted variable bias we use the rich data we have on soft inputs (e.g., hours of instruction and teacher behavior) and hard inputs (e.g., textbooks and number of teachers) and include a wide set of variables in M_{is}. But two problems arise: 1) As Bullock and Ha (2011) state, "it is normally impossible to measure all possible



(a) Randomization

(b) Mediation

(c) Mediation under assumption 1

Figure 1.5: Causal relationships under different models

Note: This figured is based on Figure 1 in Heckman and Pinto (2015) and shows the mechanisms of causality for treatment effects. Arrows represent causal relationships. Circles represent unobserved variables. Squares represent observed variables. Y are test scores. V are unobserved variables. T is the treatment variable. X are time-invariant covariates. R is the random device used to assign treatment status. M are measured mediators. U are unmeasured mediators.

mediators. Indeed, it may be impossible to merely *think* of all possible mediators". Thus, despite being extensive, the list may be incomplete. 2) It is unclear what the relevant mediators are, and adding an exhaustive list of them will reduce the degrees of freedom in the estimation and lead to multiple-inference problems. As a middle ground between these two issues, we use "Double Lasso" (Belloni, Chernozhukov, and Hansen 2014b, 2014a; Urminsky, Hansen, and Chernozhukov 2016) to select controls that are relevant from a statistical point of view, as opposed to having the researcher choose them *ad hoc*. "Double Lasso" is akin to Lasso, but provides standard errors that are valid after model selection.[46]

We use two sets of mediators. The first only includes raw inputs: teachers per student, textbooks per student, and teachers' characteristics (age, experience, and ability). Results from estimating equation 1.5 with these mediators are shown in Columns 2 and 3 of Table 1.11. The second includes raw inputs as well as changes in the use of these inputs (e.g., teacher behavior measurements, student attendance, and hours of instructional time per week). Results from estimating equation 1.5 with these

[46] Lasso is similar to OLS but penalizes according to the number of controls used. See James et al. (2014) for a recent discussion.

mediators are shown in Columns 4 and 5 of Table 1.11. For reference, we include a regression with no mediators (Column 1) which replicates the results from Table 1.3. The dependent variable is the composite test score (IRT score using both math and English questions).

The "direct" treatment effect of PSL is positive after controlling for more and better inputs (Columns 2 and 3). However, the drop in the point estimate, compared to Column 1, suggests that changes in inputs explain about half of the total treatment effect. The persistence of a "direct" treatment effect in these columns suggests that changes in the use of inputs are an important mechanism as well. The results from Columns 3 and 4 provide ancillary evidence that changes in the use of inputs (i.e., management) are important pathways to impact. After controlling for how inputs are used (e.g., teacher attendance) the "direct" treatment effect is close to zero.

In Section 1.3 we estimated equation (1.4) for several mediators. Combining those results with the results from Table 1.11, we show in Figure 1.6 the mediation effect ($\beta_{1.4} \times \theta_{1.5}$) for the intermediate outcomes selected by "Double Lasso", as well as the direct effect ($\beta_{1.5}$). The left panel uses only raw inputs as mediators, while the right panel also includes changes in the use of inputs. Figure A.4 in Appendix A.1 includes all the possible intermediate outcomes.

Over half of the overall increase (60.8%–62.4%) in learning appears to have been due to changes in the composition of teachers (measured by teacher's age, a salient characteristic of new teaching graduates). Once we allow changes in the use of inputs to act as mediators, teacher attendance accounts for 15.4% of the total treatment effect. Although changes to teacher composition make it impossible to claim that teacher attendance increases purely due to management changes, our estimates from Section 1.3.3 suggest that providers are able to increase teacher attendance even if the pool of

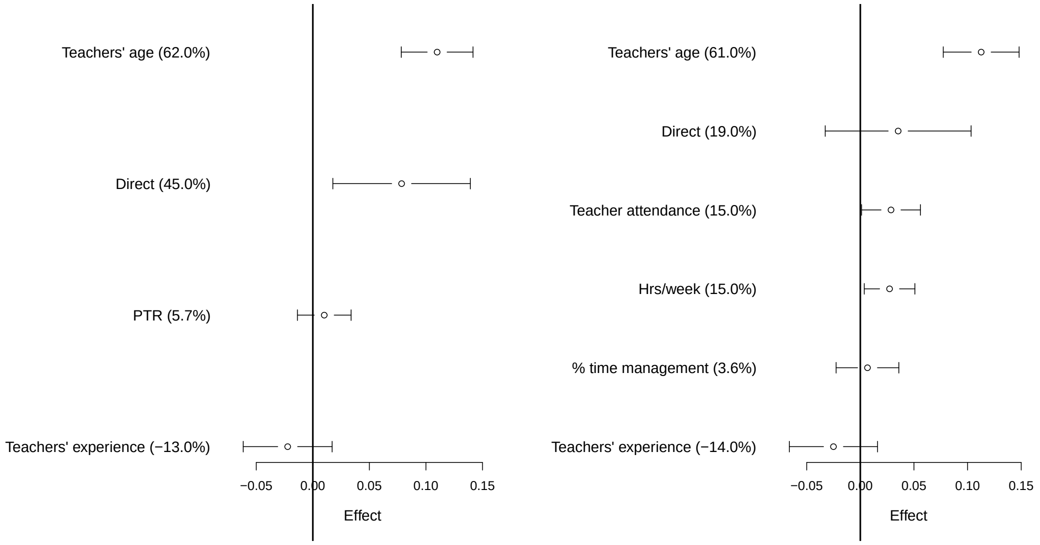
Table 1.11: Effect of mediator on learning

		Inputs		Inputs+Management	
	(1)	(2)	(3)	(4)	(5)
Treatment	0.188***	0.091**	0.115**	0.034	0.032
	(0.032)	(0.044)	(0.048)	(0.051)	(0.055)
PTR		-0.001	-0.000		-0.002
		(0.002)	(0.002)		(0.001)
Teachers' age		-0.014***	-0.014***	-0.013***	-0.010***
		(0.003)	(0.003)	(0.002)	(0.002)
Teachers' experience		0.006	0.008*	0.006	0.005
		(0.005)	(0.005)	(0.005)	(0.005)
Textbooks			-0.001		-0.000
			(0.001)		(0.001)
Writing materials			-0.000		-0.000
			(0.001)		(0.001)
% w/private school exp.			-0.000		-0.000
			(0.000)		(0.000)
Teachers' test score			0.056		0.073
			(0.049)		(0.048)
Certified teachers			0.001		0.000
			(0.001)		(0.001)
% time on management				0.027	0.009
				(0.091)	(0.082)
Teacher attendance				0.002**	0.002*
				(0.001)	(0.001)
Hrs/week				0.008**	0.008*
				(0.004)	(0.004)
Good practices (PCA)					0.079***
					(0.024)
Student attendance					-0.048
					(0.081)
Instruction (% class time)					-0.000
					(0.001)
No. of obs.	3,492	3,458	3,458	3,492	3,458
R2	0.53	0.54	0.55	0.54	0.55
Mediators	None	Lasso	All	Lasso	All

Independent variable is the composite IRT score. Dependent variables are standardized (mean zero standard deviation of 1). Column 1 replicates the results in Table 1.3. Columns 2 and 3 include raw inputs. Columns 4 and 5 include raw inputs and the use of inputs. Column 2 and column 4 include mediators selected by "Double Lasso". Columns 3 and 5 include all mediators. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

teachers is held constant. Finally, 44.5% of the total treatment effect is a residual (the direct effect) when we only control for changes in inputs, but this drops to 19% when we control for changes in the *use of* inputs.

In short, roughly half of the overall increase in learning appears to have been due to changes in the composition of teachers. Teacher attendance (which may reflect underlying managerial practice) explains much of the residual not explained by the younger, better-trained teachers. Extra resources (new and younger teachers) are an important pathway to impact in the PSL program, but changes in management practices play an equally important role. As a complementary exercise, we estimate $\theta_{1.5}$ using only variation from the control schools, and estimate the "direct effect" as the residual treatment effect not explained by the mediators (see Table A.15 in Appendix A.1). These results suggest that, holding the productivity of inputs fixed in treatment school, over 70% of the treatment effect cannot be explained by a change in inputs.



(a) Inputs

(b) Inputs & Management

Figure 1.6: Direct and mediation effects

Note: Direct ($\beta_{1.5}$) and mediation effects ($\beta_{1.4} \times \theta_{1.5}$) for the mediators selected via "Double Lasso". Note that the direct effect captures the treatment effect that is not mediated via the mediators. The percentage of the total treatment effect explained by each variable is in parenthesis. The point estimates in each panel are directly comparable to each other. Point estimates and 90% confidence intervals are plotted. Panel 1.6a shows treatment effects allowing only change in inputs as mediators. Panel 1.6b shows treatment effects allowing change in inputs and in the use of inputs as mediators.

1.5 Provider comparisons

The main results in Section 1.3 address the impact of the PSL program from a policy-maker's perspective, answering the question, "What can the Liberian government achieve by contracting out management of public schools to a variety of private organizations?" However, these results mask a great deal of heterogeneity across providers.

1.5.1 Methodology: Bayesian hierarchical model

There are two hurdles to estimating provider-specific treatment effects. First, the assignment of providers to schools was not random, which resulted in (non-random) differences in schools and locations across providers (see Appendix A.10 for more details). While the estimated treatment effects for each provider are internally valid, they are not comparable to each other without further assumptions. Second, the sample sizes for most providers are too small to yield reliable estimates.

To mitigate the bias due to differences in locations and schools we control for a comprehensive set of school characteristics (to account for the fact that some providers' schools will score better than others for reasons unrelated to PSL), as well as interactions of those characteristics with a treatment dummy (to account for the fact that raising scores through PSL relative to the control group will be easier in some contexts than others). We control for both student (age, gender, wealth, and grade) and school characteristics (pre-treatment enrollment, facilities, and rurality).

Because randomization occurred at the school level and some providers are managing only four or five treatment schools, the experiment is under-powered to estimate their effects.[47] Additionally, since the "same program" was implemented

[47] There are not enough schools per provider to get reliable standard errors by clustering at the school level. Therefore, when comparing providers we collapse the data to the school level.

by different providers, it would be naïve to treat providers' estimators as completely independent from each other.[48] We take a Bayesian approach to this problem, estimating a hierarchical model (Rubin 1981) (see Gelman et al. (2014) and Meager (2016) for a recent discussion). Intuitively, by allowing dependency across providers' treatment effects, the model "pools power" across providers, and in the process pulls estimates for smaller providers toward the overall average (a process known as "shrinkage"). The results of the Bayesian estimation are a weighted average of providers' own performance and average performance across all providers, and the proportions depend on the provider's sample size. We apply the Bayesian estimator after adjusting for baseline school differences and estimating the treatment effect of each provider on the average school in our sample.[49]

Formally, let

$$Y_{isgc} = \alpha_g + \beta_c treat_s + \varepsilon_{isgc} \tag{1.8}$$

where Y_{isgc} is the test score for student i in school s in group g (denoting the matched pairs used for randomization), assigned to provider c; α_g is a matched-pair fixed effect (i.e., stratification-level dummies); $treat_s$ is an indicator for whether school s was randomly chosen for treatment; and ε_{isgc} are the error terms. The difference between equation 1.8 and equation 1.1 is that the treatment effect (β_c) is provider specific.

Asymptotically, the estimator of the treatment effect for each provider is normally

[48] In a frequentist framework treatment estimates for providers are considered independent when compared to each other.

[49] Coincidentally, the textbook illustration of a Bayesian hierarchical model is the estimate of treatment effects for an education intervention run in eight different schools with varied results (Rubin 1981; Gelman et al. 2014).

distributed (assuming the standard error is known):[50]

$$\hat{\beta}_c \sim N(\beta_c,\sigma_c^2) \tag{1.9}$$

The bayesian hirerichal model further assumes that

$$\beta_c \sim N(\mu,\tau^2) \tag{1.10}$$

Finally, we place a prior distribution over μ and τ^2, and estimate the posterior distribution of β_c. In the main results shown below we use flat priors ("improper uniform priors"). By imposing some structure over the treatment effects for each provider (β_c), the posterior standard errors for each treatment effect become smaller, and the posterior treatment effects are pulled towards the overall average ("shrinkage"). In Appendix A.5 we show that the results are robust to the prior; how the posterior treatment effects (and standard errors) vary with τ; and the posterior distribution of τ for the case in the case of a flat prior.

1.5.2 Baseline differences

As discussed in Section 1.2.2 and shown in Table A.1, PSL schools are not a representative sample of public schools. Furthermore, there is heterogeneity in school characteristics across providers. This is unsurprising since providers stated different preferences for locations and some volunteered to manage schools in more remote and marginalized areas. We show how the average school for each provider differs from the average public school in Liberia in Table 1.12 (Table A.25 in Appendix A.10 shows simple summary statistics for the schools of each provider). We reject

[50] In reality, the standard error is unknown and therefore $\frac{\hat{\beta}_c-\beta_c}{\hat{\sigma}_c^2}$ follows a t-student distribution. However, we assume the standard error is known for exposition purposes.

the null that providers' schools have similar characteristics on at least three margins: number of students, pupil/teacher ratio, and the number of permanent classrooms. Bridge International Academies is managing schools that were considerably bigger (in 2015/2016) than the average public school in Liberia (by over 150 students), and these schools are larger than those of other providers by over 100 students. Most providers have schools with better infrastructure than the average public school in the country, except for Omega and Stella Maris. Finally, while all providers have schools that are closer to a paved road than other public schools, Bridge's and BRAC's schools are about 2 km closer than other providers' schools.

Table 1.12: Baseline differences between treatment schools and average public schools, by provider

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	BRAC	Bridge	YMCA	MtM	Omega	Rising	St. Child	Stella M	p-value equality
Students	31.94	156.19***	-23.03	35.49	-0.83	31.09	-19.16	-22.53	.00092
	(27.00)	(25.48)	(49.01)	(27.69)	(53.66)	(34.74)	(59.97)	(59.97)	
Teachers	1.23*	2.72***	1.42	1.70**	1.16	0.59	1.13	0.76	.66
	(0.70)	(0.66)	(1.28)	(0.72)	(1.40)	(0.90)	(1.56)	(1.56)	
PTR	-4.57	5.77*	-8.47	-5.45	-6.02	2.34	-10.62	-7.29	.079
	(3.27)	(3.09)	(5.94)	(3.36)	(6.50)	(4.21)	(7.27)	(7.27)	
Latrine/Toilet	0.18**	0.28***	0.26*	0.25***	0.23	0.22**	0.06	0.18	.96
	(0.08)	(0.07)	(0.14)	(0.08)	(0.16)	(0.10)	(0.17)	(0.17)	
Solid classrooms	0.63	2.81***	2.64*	-0.11	1.85	1.59*	-1.95	1.30	.055
	(0.75)	(0.71)	(1.36)	(0.77)	(1.49)	(0.97)	(1.67)	(1.67)	
Solid building	0.28***	0.22***	0.19	0.09	0.26*	0.19*	0.23	0.23	.84
	(0.08)	(0.07)	(0.14)	(0.08)	(0.15)	(0.10)	(0.17)	(0.17)	
Nearest paved road (KM)	-9.25***	-10.86***	-7.13*	-8.22***	-4.47	-7.13***	-4.56	-7.79*	.78
	(2.03)	(1.91)	(3.67)	(2.08)	(4.01)	(2.60)	(4.48)	(4.48)	

This table presents the difference between public schools and the schools operated by each provider. The information for all schools is taken from the 2015/2016 EMIS data, and therefore is pre-treatment information. Column 9 shows the p-value for testing $H_0 : \beta_{BRAC} = \beta_{Bridge} = \beta_{YMCA} = \beta_{MtM} = \beta_{Omega} = \beta_{Rising} = \beta_{St.Child} = \beta_{StellaM}$. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. Since some providers had no schools with classes above the class caps, there is no data to estimate treatment effects over constrained classes. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

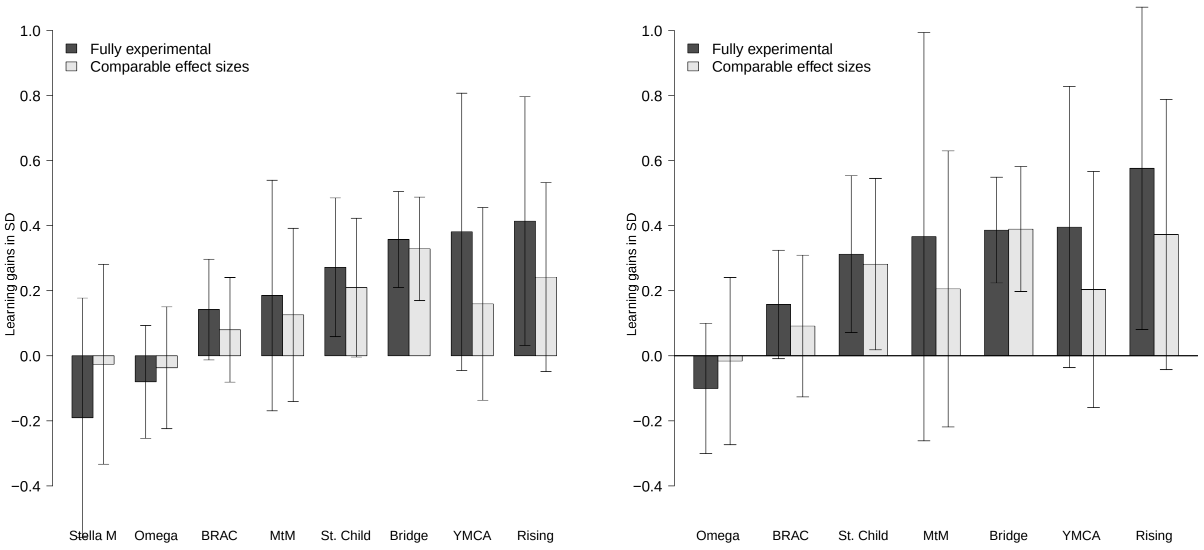
1.5.3 Learning outcomes

The raw treatment effects on test scores for each individual provider shown in Figure 1.7 are internally valid, but not comparable. They are positive and significantly different from zero for three providers: Rising Academies, Bridge International Academies, and Street Child. They are positive but statistically insignificant for Youth Movement for Collective Action, More Than Me, and BRAC. The estimates which we label as "comparable treatment effects" differ in two respects: They adjust for baseline differences and "shrink" the estimates for smaller providers using the Bayesian hierarchical model. While the comparable effects are useful for comparisons, the raw experimental estimates remain cleaner for non-comparative statements (e.g., whether a provider had an effect or not).[51]

Intention-to-treat (ITT) treatment effects are shown in Figure 1.7a (i.e., over all students enrolled in a treatment school in 2015/2016, regardless of whether they attended an actual PSL school in 2016/2017). Treatment-on-the-treated (ToT) treatment effects are shown in Figure 1.7b (i.e., the effect for students who actually attended a PSL school in 2016/2017). Non-compliance can happen either at the school level (if a provider opted not to operate a school or the school did not meet the eligibility criteria), or at the student level (if the student no longer attends a treatment school). Comparable ITT treatment effects across providers from the Bayesian hierarchical model are also shown in Panel A of Table 1.13.

There is considerable heterogeneity in the results. The data suggest providers' learning impacts fall into three categories, based on a k-means clustering algorithm. In the first group, YMCA, Rising Academies, Street Child, and Bridge International Academies generated an increase in learning of 0.26σ across all subjects. In the second

[51] Figure A.5 in Appendix A.1 shows the the effects after adjusting for differences in school characteristics (before the Bayesian hierarchical model) and the effects after applying a Bayesian hierarchical model (but without adjusting for school differences).



(a) Intention-to-treat (ITT) effect

(b) Treatment-on-the-treated effect (ToT)

Figure 1.7: Treatment effects by provider

Note: These figures show the raw, fully experimental treatment effects and the comparable treatment effects after adjusting for differences in school characteristics and applying a Bayesian hierarchical model. Figure 1.7a shows the intention-to-treat (ITT) effect, while Figure 1.7b shows the treatment-on-the-treated (ToT) effect. The ToT effects are larger than the ITT effects due to providers replacing schools that did not meet the eligibility criteria, providers refusing schools, or students leaving PSL schools. Stella Maris had full non-compliance at the school level and therefore there is no ToT effect for this provider.

group, BRAC and More than Me generated an increase in learning of 0.12σ. In the third group, consisting of Omega and Stella Maris,[52] estimated learning gains are on the order of -0.03σ, and indistinguishable from zero in both cases.

Below we explore whether these gains impose negative externalities on the broader education system (i.e., whether better performance came at a cost to the education system as a whole).[53]

[52] Non-compliance likely explains the lack of effect for these two providers. Stella Maris never took control of its assigned schools, and Omega had not taken control of all its schools by the end of the school year. Our teacher interviews reflect these providers' absence: in 3 out of four Stella Maris schools, all of the teachers reported that no one from Stella had been at the school in the previous week, and in 6 out of 19 Omega schools all of the teachers reported that no one from Omega had been at the school in the previous week.

[53] We had committed in the pre-analysis plan to compare for-profit to non-profit providers. This comparison yields no clear patterns.

1.5.4 Are public and private interests aligned under PSL?

Economists typically approach outsourcing in a principal-agent framework: A government (the principal) seeks to write a complete contract defining the responsibilities of the private provider (the agent). This evaluation is part of that effort. In real-world settings, contracts are inevitably incomplete. It is impossible to pre-specify every single action and outcome that a private provider must concern themselves with when managing a school. Economists have offered a number of responses to contractual incompleteness. One approach focuses on fostering competition among providers via the procurement process and parental choice (Hart, Shleifer, and Vishny 1997). Another, more recent approach puts greater focus on the identity of the providers, on the premise that some agents are more "mission motivated" than others (Besley and Ghatak 2005; Akerlof and Kranton 2005). If providers have intrinsic motivation and goals that align with the principal's objectives then they are unlikely to engage in pernicious behavior. This may be the case for non-profit providers whose core mission is education. In the particular case of Liberia, this may also be true for for-profit providers who are eager to show their effectiveness and attract investors and philanthropic donors. But, if providers define their objectives more narrowly than the government, they may neglect to pursue certain government goals.

We examine three indicators illustrating how public and private goals may diverge under PSL: providers' willingness to manage any school (as opposed to the best schools); providers' willingness to work with existing teachers and improve their pedagogical practices and behavior (as opposed to having the worst performing teachers transferred to other public schools, imposing a negative externality on the broader school system); and providers' commitment to improving access to quality education (rather than learning gains for a subset of pupils). In short, we're concerned with providers rejecting "bad" schools, "bad" teachers, and excess pupils.

We already studied school selection in Section 1.5.2. To measure teacher selection, we study the number of teachers dismissed and the number of new teachers recruited (Table 1.13 - Panel B). As noted above, PSL led to the assignment of 2.6 additional teachers per school and 1.2 additional teachers exiting per school. However, large-scale dismissal of teachers was unique to one provider (Bridge International Academies), while successful lobbying for additional teachers was common across several providers. Although weeding out bad teachers is important, a reshuffling of teachers is unlikely to raise average performance in the system as a whole.

While enrollment increased across all providers, the smallest treatment effect on this margin is for Bridge, which is consistent with that provider being the only one enforcing class size caps (see Panel C in Table 1.13 and Figure A.6 in Appendix A.1 for more details). As shown above, in classes where class-size caps were binding (10% of all classes holding 30% of students at baseline), enrollment fell by 12 students per grade.

Table 1.13: Comparable ITT treatment effects by provider

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	BRAC	Bridge	YMCA	MtM	Omega	Rising	St. Child	Stella M	p-value
Panel A: Student test scores									
English (standard deviations)	0.14	0.26***	0.17	0.02	0.23	0.21*	0.03	0.24	0.10
	(0.09)	(0.09)	(0.14)	(0.11)	(0.16)	(0.12)	(0.17)	(0.17)	
Math (standard deviations)	0.04	0.35***	0.10	-0.05	0.22	0.19	-0.05	0.10	0.0090
	(0.10)	(0.10)	(0.17)	(0.11)	(0.18)	(0.13)	(0.19)	(0.18)	
Composite (standard deviations)	0.08	0.33***	0.13	-0.04	0.24	0.21	-0.03	0.16	0.019
	(0.10)	(0.10)	(0.16)	(0.11)	(0.18)	(0.13)	(0.19)	(0.18)	
Panel B: Changes to the pool of teachers									
% teachers dismissed	-8.59	49.54***	13.93	-6.22	0.52	-0.79	-1.66	12.00	<0.001
	(6.48)	(7.17)	(11.09)	(6.76)	(11.94)	(9.01)	(12.92)	(12.96)	
% new teachers	38.15***	70.80***	47.19**	22.61*	20.56	36.01**	-9.64	35.69*	0.0060
	(11.14)	(13.13)	(18.75)	(11.91)	(20.12)	(15.23)	(26.28)	(21.10)	
Age in years (teachers)	-5.50***	-9.13***	-7.80***	-5.74***	-8.08***	-6.54***	-6.00**	-3.50	0.16
	(1.71)	(2.18)	(2.56)	(1.73)	(2.74)	(2.10)	(2.71)	(3.51)	
Panel C: Enrollment and access									
Δ enrollment	31.89	7.61	12.60	28.84	16.39	25.39	15.79	27.57	0.48
	(25.45)	(26.73)	(32.73)	(25.02)	(32.89)	(28.71)	(34.03)	(34.18)	
Δ enrollment (constrained grades)	41.89	-29.68**	41.42	-3.48	41.63	22.52	–	–	0.48
	(43.93)	(14.60)	(44.08)	(36.68)	(43.75)	(47.11)	(–)	(–)	
Student attendance (%)	18.44***	12.81*	20.75**	17.54***	19.03**	19.39**	16.68*	17.45*	0.48
	(6.59)	(7.53)	(9.16)	(6.69)	(8.96)	(7.96)	(9.47)	(9.03)	
% students still attending any school	-1.99	1.30	-4.83	-2.03	-3.84	-1.98	-3.20	-3.18	0.35
	(3.36)	(3.69)	(5.93)	(3.62)	(5.61)	(4.24)	(5.28)	(5.57)	
% students still attending same school	0.53	2.36	0.34	0.66	0.72	0.25	0.28	0.16	0.44
	(1.76)	(1.91)	(2.58)	(1.87)	(2.58)	(2.23)	(2.64)	(2.78)	
Observations	40	45	8	12	38	10	24	8	

ITT treatment effect for each provider, after adjusting for differences in baseline school characteristics with a Bayesian hierarchical model. Column 9 shows p-value for testing $H_0 : \beta_{BRAC} = \beta_{Bridge} = \beta_{YMCA} = \beta_{MtM} = \beta_{Omega} = \beta_{Rising} = \beta_{St.Child} = \beta_{StellaM}$. Some operators had no schools with class sizes above the caps. Table A.16 in Appendix A.1 has the raw experimental treatment effects by provider. Standard errors are shown in parentheses. Estimation is conducted on collapsed, school-level data. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.6 Cost-effectiveness analysis

From a policy perspective, the relevant question is not only whether the PSL program had a positive impact (especially given its bundled nature), but whether it is the best use of scarce funds. Cost-effectiveness analysis compares programs designed to achieve a common outcome with a common metric — in this case learning gains — by their cost per unit of impact. Inevitably, this type of analysis requires a host of assumptions, which must be tailored to a given user and policy question (see Dhaliwal et al. (2013) for a review). Section 1.2.1 outlined various assumptions behind the cost estimates for each provider.[54]

Given the contested nature of these assumptions and the difficulty of modeling the long-term unit cost of PSL in a credible way, we opt to present only basic facts here. We encouraged operators to publish their *ex post* expenditure data in the same repository as our survey data, and some have agreed to do this.

We make a conservative assumption and perform a single cost-effectiveness calculation assuming a cost of $50 per pupil (the lowest possible cost associated with the program). Given that the ITT treatment effect is $.19\sigma$, test scores increased 0.38σ per $100 spent.[55] Taking these estimates at face value suggests that in its first year PSL is not a cost-effective program for raising learning outcomes. While many education interventions have either zero effect or provide no cost data for cost-effectiveness calculations (Evans and Popova 2016), a review by Kremer, Brannen, and Glennerster (2013) of other interventions subject to experimental evaluation in developing countries highlights various interventions that yield higher per-dollar gains than PSL (see Figure 1.8).

[54] We do not present a cost-effective comparison of the effect of the program on access to schooling since the overall treatment effect on enrollment is not statistically different from zero.

[55] Note that given our design, we are unable to take into account any test score gains associated with drawing new students into school.

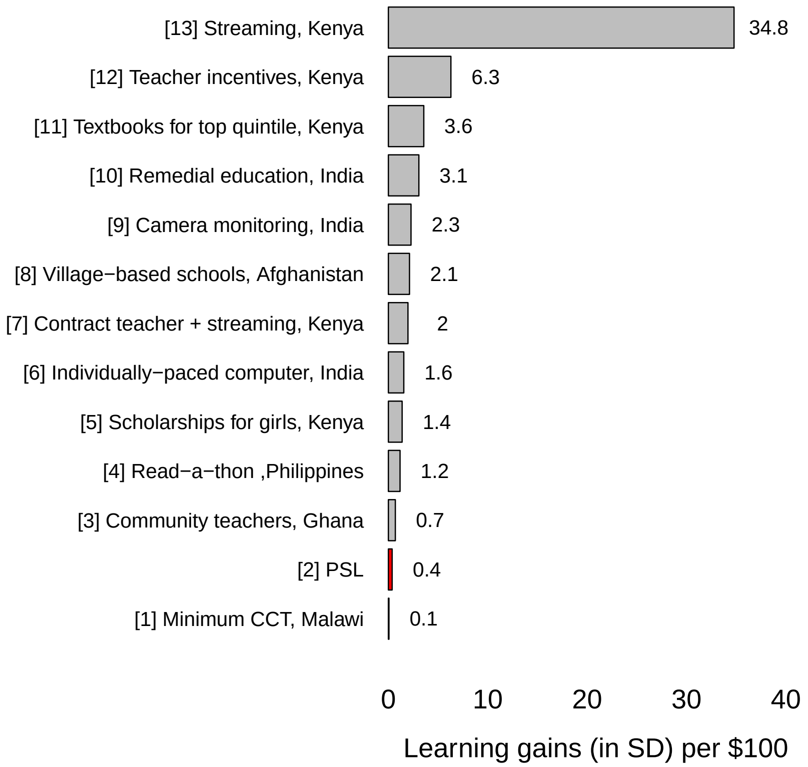


Figure 1.8: Cost per child and treatment effects for several education interventions
Note: Figures show the learning gains per 100 (2011) USD. For more details on the calculations for [1], [4]-[13] see https://www.povertyactionlab.org/policy-lessons/education/increasing-test-score-performance. Data for [3] is taken from Kiessel and Duflo (2014). The original studies of each intervention are as follows: [7] and [13] Duflo, Dupas, and Kremer (2011, 2015); [1] Baird, McIntosh, and Özler (2011); [4] Abeberese, Kumler, and Linden (2014); [5] Kremer, Miguel, and Thornton (2009); [6] and [10] Banerjee et al. (2007); [8] Burde and Linden (2013); [9] Duflo, Hanna, and Ryan (2012); [11] Glewwe, Kremer, and Moulin (2009); [12] Glewwe, Ilias, and Kremer (2010).

However, it is unclear whether cost-effectiveness calculations from other contexts and interventions are relevant to the Liberian context and comparable to our results. First, test design is crucial to estimates of students' latent ability (and thus to treatment effects on this measure).[56] Since different interventions use different exams to measure students' ability, it is unclear that the numerator in these benefit-cost ratios is comparable.[57] The second problem is external validity. Even if treatment estimates were

[56] For example, Table A.7 shows how PSL treatment estimates vary depending on the measure of students' ability we use.

[57] For more details, see Singh (2015a)'s discussion on using standard deviations to compare interven-

comparable across settings, treatment effects probably vary across contexts. This does not mean we cannot learn from different programs around the world, but implementing the same program in different settings is unlikely to yield identical results everywhere. Finally, the cost of implementing a program *effectively* (the denominator) is also likely to be variable across settings.

An important feature of our experiment is its real-world setting, which may increase the likelihood that gains observed in this pilot could be replicated at a larger scale. Interventions successfully implemented by motivated non-government organizations (NGO) often fail when implemented at scale by governments (e.g., see Banerjee, Duflo, and Glennerster (2008), Bold, Kimenyi, and Sandefur (2013), Dhaliwal and Hanna (2014), Kerwin and Thornton (2015), and Cameron and Shah (2017)). The public-private partnership is designed to bypass the risk of implementation failure when taken up by the government, simply because the government is never the implementing agency. However, the program may still fail if the government withdraws support or removes all oversight.

1.7 Conclusions

Public-private partnerships in education are controversial and receive a great deal of attention from policy makers. Yet, the evidence for or against them is almost nonexistent, especially in developing countries (Aslam, Rawal, and Saeed 2017). Advocates argue that privately provided but publicly funded education is a means to inject cost-efficiency, through private providers, into education without compromising equity. Critics argue that outsourcing will lead to student selection and low-quality, expensive schools.

tions.

We present empirical evidence that both advocates and critics are partially right. The Partnership Schools for Liberia program, a public-private partnership that delegated *management* of 93 public schools ($\sim$ 3.4% of all public schools) to eight different private organizations, was an effective way to circumvent low state capacity and improve the quality of education. The ITT treatment effect on test scores of PSL program students after one academic year of treatment are .18σ for English (p-value < 0.001) and .18σ for math (p-value < 0.001).

We find no evidence that providers engage in student selection — the probability of remaining in a treatment school is unrelated to age, gender, household wealth, or disability. However, costs were high, performance varied across providers, and the largest provider pushed excess pupils and under-performing teachers into other government schools.

One interpretation of our results is that contracting rules matter. Changing the details of the contract might improve the overall results of the program. For instance, contracts could forbid class-size caps or require that students previously enrolled in a school be guaranteed re-admission once a school joins the PSL program. Similarly, contracts could require prior permission from the Ministry of Education before releasing a public teacher from their place of work.

However, fixing the contracts and procurement process is not just a question of technical tweaks; it reflects a key governance challenge for the program. Contract differences are endogenous: The largest provider opted not to participate in the competitive bidding process and made a separate bilateral agreement with the government. Ultimately, a different contract allowed pushing excess pupils and under-performing teachers into other government schools. This underlines the importance of uniform contracting rules and competitive bidding in a public-private partnership.

On the other hand, contracts are by nature incomplete and subject to regulatory capture. While Hart, Shleifer, and Vishny (1997) focus on incomplete contracts when deciding whether outsourcing is wise, the mission matching literature a la Besley and Ghatak (2005) focuses on heterogeneity in contractors' intrinsic motivation. We examine a setup where eight providers were offered to participate in the same program. We observe significant heterogeneity in learning outcomes and in actions that might generate negative spillovers for the broader education system. Heterogeneity in both efficiency and mission appears to be a first order concern here.

To our knowledge, we provide the first experimental estimates of the intention-to-treat (ITT) effect of outsourcing the management of existing schools to private providers in a developing country. In contrast to the U.S. charter school literature, which focuses on experimental effects for the subset of schools and private provider where excess demand necessitates an admissions lottery, we provide treatment effects from across the distribution of outsourced schools in this setting.

But an assortment of questions remain open for future research. First, given the bundled nature of this program, more evidence is needed to isolate the effect of outsourcing management. Variations of outsourcing also need to be studied (e.g., not allowing any teacher re-assignments, or allowing providers to hire teachers directly).

Second, while we identify sources of possible externalities from the program – e.g., pushing pupils or teachers into nearby schools – we are unable to study the effect of these externalities (positive or negative). Another key potential negative externality for other public schools is the opportunity cost of the program: PSL may deprive other schools of scarce resources by garnering preferential allocations of teachers or funding. On the other hand, traditional public schools may learn good management and pedagogical practices from nearby PSL schools. In addition, the program may lead to changes within the Ministry of Education that improve performance of the system as a whole.[58]

More broadly, future research is needed to understand how procurement rules affect the long term outcomes of PPP programs such as this one. For example, a key difference between the private and the public sector is the dynamics of entry and exit. Underperforming public schools are never closed, and underperforming education officers and teachers are rarely dismissed. In contrast, in the private sector consumer choice (and exit), together with hard budget constraints, force underperforming schools out of the market (Pritchett 2013). Competition requires active encouragement. A challenge for PPP programs is whether the government procurement rules can create entry and exit dynamics that mimic the private sector, filtering out bad providers (in a relevant public cost effectiveness sense). If not, then in steady state the program may replicate the (undesirable) exit dynamics of the public sector, and lead to under performing PPP schools.

Chapter 1, in full, is currently being prepared for submission for publication of the material. Romero, Mauricio; Sandefur, Justin; Sandholtz, Wayne Aaron. "Outsourcing Service Delivery in a Fragile State: Experimental Evidence from Liberia". The dissertation author was the primary investigator and author of this material.

[58] For example, the Ministry is reforming some of measurement systems, to monitor provider performance.

Chapter 2

Cross-Age Tutoring: Experimental Evidence from Kenya

(Co-authors: Lisa Chen and Noriko Magari)

2.1 Introduction

Over the past three decades, access to primary school has dramatically increased in low- and middle-income countries (United Nations 2015). However, the quality of education remains poor despite increases in enrollment.[1] This is especially worrisome as evidence suggests that the quality of education, not the quantity, is what matters for growth (Hanushek and Kimko 2000; Hanushek and Wößmann 2007). The combination of low learning outcomes and a fiscally constrained environment in developing countries incites a search for cost-effective ways to improve education quality.

Interventions that tailor teaching to student learning levels are consistently signaled by the literature as having the largest effects on learning levels across different settings (for three recent reviews of the literature see Glewwe and Muralidharan (2016), Evans and Popova (2016), and Snilstveit et al. (2016)). However, teachers often lack the time (or incentives) to give each child personalized instruction tailored to their needs and providing schools with additional teachers to do so is expensive. Cross-age tutoring, where older students tutor younger students, is an inexpensive alternative to teacher-provided tutoring. It substitutes a trained instructor (the teacher) with an untrained one (the older student) at the cost of the older student's time. To the extent that tutoring can also provide benefits to tutors (e.g., mastering knowledge and increasing social skills), cross-age tutoring could result in an overall welfare improvement. We present results from a large randomized control trial over 180 schools, 15,000 tutees, and 15,000 tutors in Kenya, in which schools are randomly selected to implement a cross-age tutoring program in either English or math.

Cross-age tutoring has taken place since at least 95 CE (Quintilianus and Halm

[1] For example, despite enrollment rates of over 90%, less than 50% of children in Argentina, Colombia, Morocco, Uganda, Namibia, and Malawi attain "minimum literacy standards" (World Bank 2007). Estimates from Mexico and Brazil show that more than 50% of children lack minimal competency in mathematics (Filmer, Hasan, and Pritchett 2006). Jones et al. (2014) find that the majority of children in grade 3 across East Africa are unable to recognize a single word in their medium of instruction.

1869) and is now widely used across the world. In the typical setup, the tutor is a few years older than the tutee and works with him or her on specific problems in a particular subject. Cross-age tutoring is often thought to have positive effects for both tutors and tutees in terms of academic achievement and social-emotional outcomes (Cohen, Kulik, and Kulik 1982). Yet the evidence on the subject is mixed, based on small-scale experiments or observational data, and mostly from developed countries. An early review of the literature showed that cross-age tutoring had a positive effect (both in terms of academic performance and attitudes) on both tutees and tutors (Cohen, Kulik, and Kulik 1982). A more recent review that included only randomized control trials came to the conclusion that cross-age tutoring in math has non-significant effects on math test scores and cross-age tutoring in reading has a small (statistically significant) positive effect (Shenderovich, Thurston, and Miller 2016).

In our setting, tutoring took place each school day of the 2016 academic year. At the end of every day, older students tutored younger students in either English or math for 40 minutes. Tutors were five grades above tutees. In some schools the tutoring focused on math, while in others it focused on English. Whether math or English tutoring took place was randomized across schools. Therefore, within a school all grades participated in either math or English tutoring.[2]

Defining an appropriate counterfactual is a common challenge with interpreting curriculum interventions. Most curriculum interventions involve additional instructional time, so measuring whether any changes in learning are due to additional instructional time or to the nature of the instruction is difficult. A noteworthy feature of this experiment is that all schools in our sample implement a tutoring program (i.e., there was no "pure" control group that did not receive any tutoring). The random assignment determines whether a school implemented English or a math tutoring.

[2] Section 2.2.2 has details on the math and English tutoring interventions.

Therefore, all of our results should be interpreted as the impact of tutoring in math relative to tutoring in English (or vice versa). Although the lack of a "pure" control group may give the impression that our estimates are difficult to interpret, in this experiment we know exactly what the counterfactual for time use is across groups. An alternative would have been to provide tutoring in some schools but not in others. Since time is finite, we would either need to control how tutors and tutees use the time allocated for tutoring in control schools or let schools/students choose how to use this time. Is not clear that either case would lead to a better counterfactual.

Cross-age tutoring in math, relative to tutoring in English, has a small positive effect (0.06 SD, p-value of 0.073) on math test scores. These results do not hold true for English tutoring: relative to math tutoring, it has no positive effect on English test scores (we can rule out an effect of 0.077 SD with 95% confidence). There is considerable heterogeneity in the results when broken down by the student's baseline learning level. Specifically, the effect of math tutoring, relative to English tutoring, on math test scores is largest for students in the middle of the ability distribution (0.144 SD, p-value of 0.005). The point estimate is almost zero for students with either very low or very high baseline learning levels. This is consistent with: a) tutors not being able to help students who are advanced learners and need an instructor with a high level of expertise to guide them through more advanced material; and b) tutors not being able to help tutees lagging behind grade level competencies who may need more specialized instruction to catch up.

However, there is no heterogeneity by tutees' gender or age. Similarly, there is no heterogeneity by school characteristics (pupil-teacher ratio, class size, or tutor-tutee ratio). Since we do not have data on tutor/tutee matches and teachers were responsible for matching tutees and tutors, we can only analyze the average characteristics of possible tutors for a specific tutee, and find no heterogeneity by tutors' average age, gender or proficiency level (baseline test scores).

In short, in this setting math tutoring is more effective than English tutoring in raising test scores (in the subject of tutoring), and although tutors are limited in their ability to help certain students, they can effectively increase test scores for students in the middle of the distribution of baseline learning levels.

Two central issues to the research design are multi-tasking and cross-domain spillover effects. For example, treatment could induce pupils to concentrate their attention on the subject they are being tutored in, negatively affecting their performance in other subjects. It is also possible that tutoring increases the performance of students in other subjects by releasing study time that would otherwise be devoted to the tutored subject. Although our research design does not explicitly allows us to rule out multi-tasking or spillover effects, we do not believe these are issues in practice. First, had we found effects of English tutoring on English and math tutoring on math, a possible explanation, akin to multi-tasking, would have been that tutoring in one subject erodes performance in the other subject. Second, tutoring has no effect (positive or negative) on Swahili. The lack of effect on other subjects does not rule out the possibility of cross-domain spillover effects, but the effect on other subjects would need to be the same across English and math tutoring to yield no difference when comparing the two.

Our results speak directly to three strands in the literature. First, they relate to the literature that seeks to understand the underlying production function for cognitive achievement (Todd and Wolpin 2003) and studies the impact of different education policies and programs (Glewwe and Muralidharan 2016; Evans and Popova 2016; Snilstveit et al. 2016). We present evidence on a novel approach to improving the amount of personalized instruction at low cost. Although the size of the program's effects are modest, it is essentially free, and therefore may be cost-effective relative to other alternatives for providing personalized instruction. For example, while contract teachers have been found to increase test scores by 1.97 SD per 100 USD invested (Kremer, Brannen, and Glennerster 2013), cross-age tutoring has resulted in an increase of 18 SD per 100 USD invested.

Second, our results speak to the literature on peer effects and their effect on learning (Sacerdote 2001; Zimmerman 2003; Munley, Garvey, and McConnell 2010). We explore how the quality of different tutors (at the school level) affects the outcomes of the program. Finally, we communicate with the literature on peer-learning programs. Only two of the studies reviewed by Shenderovich, Thurston, and Miller (2016) involved other elementary school students providing tutoring (as opposed to adults, community volunteers, or university students), and both of those studies focus only on reading. None of the interventions in those studies were implemented in a low- or middle-income country. To the best of our knowledge, ours is the first RCT implemented on cross-age tutoring in which tutors are students in the same school as tutees. Furthermore, it is the first study of this kind in a low-income country.

2.2 Experimental Design

2.2.1 Context

Despite high net enrollment rates in primary schools (∼95% in 2013), the quality of education in Kenya is low: Children often fail to attain proficiency in reading and numeracy in the early grades. The annual nationwide learning assessments carried out by Twaweza (the Uwezo test) consistently show that only half of grade 3 students can read a simple story at a grade 2 level in English (the national language and the language of instruction) or successfully demonstrate grade 2 numerical skills (Jones et al. 2014).

Bold, Kimenyi, and Sandefur (2013) argue that the abolition of fees for primary schools in 2003 led to a decline in the quality ("or at least perceived quality") of public schools and in response the demand for (and supply of) private primary education increased dramatically. According to World Bank statistics, the proportion of students enrolled in private primary schools more than doubled from 4.5% in 2004 to over 10.5% in 2009.

Kenya is not the only country where there has been a surge in private school enrollment. Recently several chains of for-profit, low-cost private schools have emerged around the world. These chains leverage technology to deliver lessons and manage teachers more effectively (Mbiti 2016). In this article, we work with a large low-cost private school provider, Bridge International Academies (Bridge), in which schools within their network are randomly selected to implement either a math or an English tutoring program. Bridge opened its first school in Nairobi in January 2009. By November 2014, it was operating nearly 400 schools across Kenya and had enrolled over 100,000 students.[3]

Bridge tries to takes advantage of economies of scale in school management, teacher training, and teaching guides to lower the marginal cost of delivering education.[4] English is the language of instruction in all Bridge schools, which are located across East Africa, West Africa, and India, but mainly in Kenya. The company relies heavily on technology to maintain a constant feedback loop.[5]

[3] See http://www.bridgeinternationalacademies.com/company/history/

[4] For example, in each Bridge academy, unlike in low-cost, "mom-and-pop" private schools, management consists of just one employee. This is because the vast majority of non-instructional activities that the Bridge "Academy Manager" would normally have to deal with (billing, payments, expense management, payroll processing, and more) are automated and centralized. Similarly, Bridge hires experts to develop comprehensive teacher guidelines and training programs, which are then used in all of their schools. Schools charge on average a monthly fee of USD 6 and cater to families living on USD 2 a day per person or less.

[5] Bridge follows the 8-4-4 curriculum framework mandated by the national government, but provides detailed teacher guides for each lesson that are used by teachers across the network. The guides are developed by Bridge staff at its offices in Boston and Nairobi and are then streamed to teachers' personal tablets. Teachers use tablets to upload students' information (e.g., test scores) to Bridge country headquarter offices as well.

From a research standpoint, an advantage of working with Bridge data is that all students take the same tests across all schools, and Bridge collects data on students' performance to detect levels of content mastery. This data is also used to measure and improve on teacher quality. Students are tested six times per academic year. Each academic year has three terms, and each term has a midterm and an endterm exam. Additionally, at the beginning of the academic year students in primary grades (Standards 1 - 6) take a diagnostic exam. Randomized control trials to study the effectiveness of different approaches to improving learning can be implemented relatively easily with no additional cost for data collection (often the most expensive part of a field experiment). This is the first of such trials implemented across schools in the Bridge network.

2.2.2 Intervention

The intervention took place every school day during the 2016 academic year. At the end of each school day, older students tutored younger students in either English or math for 40 minutes (3:35-4:15 pm). Tutors were five grades above tutees (see Table 2.1 for details). In some schools the tutoring focused on math, while in others it focused on English. Whether math or English tutoring took place was randomized across schools. Therefore, within a school students in all grades participated in either math or English tutoring. Table 2.2 has details on the math tutoring intervention, while Table 2.3 provides details on the English tutoring intervention.

The main objective of the math (English) tutoring program was to raise math (English) achievement among tutees (BC-Grade 2 pupils). A secondary objective was to develop communication and leadership skills among tutors (Grade 3-Grade 7 pupils) and build a school community through development of sibling-like relationships

Table 2.1: Tutors and Tutees

Tutors	Tutees
Grade 3 →	Baby Class (BC)
Grade 4 →	Nursery (NU)
Grade 5 →	Preunit (PU)
Grade 6 →	Grade 1
Grade 7 →	Grade 2

between tutees and tutors.

Tutors were given manuals with problems and activities to complete with tutees each day. Teachers led the mentoring sessions, deliver the "tutor manual", and chose how to pair tutees with tutors. Teachers were also allowed to vary the tutor-tutee pairs each day. During the first two weeks of the 2016 academic year, the mentoring sessions consisted of "mentor training". During this mentor training, teachers instructed mentors to keep pupils focused and use the "ask-tell-show-repeat" procedure to correct pupil work. "Ask-tell-show-repeat" is a four-step process in which tutees are asked to do a problem again if they answer incorrectly; they then receive verbal instructions on the correct solution if the mistake is repeated; they are then shown the correct solution if they make a mistake again; and finally, the pupil is asked to repeat the problem one last time. The idea was to provide a simple structure for mentor-pupil interaction.

For math, minor changes were introduced during the last quarter of the school year. Specifically, brief instructions for mentors replaced the teacher demonstration. This was done to shift the focus of the mentoring session from the teacher to the mentoring pairs. "Tutor manuals" then instructed tutors to provide immediate feedback to tutees, telling them whether the answer was correct or incorrect as soon as they answered a question. A simplified version of the "ask-tell-show-repeat" correction method was implemented: the "ask-show-repeat" method. Instead of first verbally instructing the pupil how to obtain the correct solution in case of a repeated mistake, the mentor was instructed to immediately show them the correct solution. Finally, teachers were instructed to "check-respond-leave" with mentors exclusively, thus empowering mentors to take responsibility for their pupils' performance (see Table 2.2 for details).

For English, minor changes were introduced during the last two quarters of the school year; most of the changes affected how much time was allocated to different activities (see Table 2.3 for details).

Table 2.2: Math tutoring intervention

	Term 1 and Term 2	Term 3
Timing	3:35 - 4:15 pm.	3:35 - 4:15 pm.
C1/C2	Intro: 3 min Teacher demo: 5 min Mentoring: 30 min Guide with 18 problems 1 topic	Intro: 3 min Mentoring 1: 22 min Mentoring 2: 15 min Guide with 60 problems 2 topics
PU	Intro: 3 min Warm-up exercise: 10 min Teacher demo: 5 min Mentoring: 15 min Guide with 10 problems 1 topic	Intro: 3 min Mentoring 1: 22 min Mentoring 2: 15 min Guide with 56 problems 2 topics
BC/NU	Introduction: 3 min Counting: 7 min Rhyming: 3 min ID numbers: 7 min ID frames: 7 min Rhyming: 3 min ID shapes: 8 min Closing: 2 min	Introduction: 3 min Counting with mentors: 7 min Rhyming: 3 min Writing numbers with mentors: 7 min Drawing frames with mentors: 7 min Rhyming: 3 min Drawing shapes with mentors: 8 min Closing: 2 min
Tutor duties	Keep pupil focused Use ask-tell-show-repeat	Correct tutee after every two problems Use ask-show-repeat
Teacher duties	Demonstrate mentoring process Circulate	Check-respond-leave

Table 2.3: English tutoring intervention

	T1	T2	T3
Timing	3:35 - 4:15 pm.	3:35 - 4:15 pm.	3:35 - 4:15 pm.
C1/C2	Introduction: 2 min Dialogue practice: 5 min Mentoring instructions: 3 min Words: 5 min Reading: 15 min Writing: 9 min	Introduction: 2 min Dialogue practice: 5 min Words: 8 min Writing: 15 min Reading: 9 min	Introduction: 3 min Words: 10 min Writing: 10 min Reading: 15 min Closing: 2 min
PU	Introduction: 3 min Dialogue practice: 5 min Practice book: 7 min Sight words: 5 min Reading: 15 min	Introduction: 3 min Dialogue practice: 5 min Sight words: 12 min Reading: 15 min	Introduction: 2 min Words: 8 min Reading: 15 min Sight words: 15 min
BC/NU	Introduction & song: 5 min Practice set: 7 min Finding words: 4 min Rhyming: 3 min Finding letters: 5 min Letter sound chant: 2 min Dialogue practice: 5 min Closing: 2 min	Introduction & song: 5 min Words: 11 min Rhyming: 3 min Finding letters: 5 min Letter sound chant: 2 min Dialogue practice: 5 min Closing: 2 min	Introduction & song: 5 min Words: 11 min Rhyming: 3 min Finding rhyme words: 7 min Letter sound chant: 2 min Finding letters: 5 min Dialogue practice: 5 min Closing: 2 min
Mentor duties	Keep pupil focused Use ask-tell-show-repeat Correction method	Keep pupil focused Use ask-tell-show-repeat Correction method	Keep pupil focused Use ask-tell-show-repeat Correction method
Teacher duties	Circulate	Circulate	Circulate

2.2.3 Sampling

Bridge has a network of over 400 schools across Kenya, but only 187 schools were eligible to participate in the trial.[6]. Randomization was stratified at the "former province" level (Kenya's provinces were replaced by a system of counties in 2013) and by average baseline test scores at each academy. Estimations take into account the randomization design by including the appropriate fixed effects (Bruhn and McKenzie 2009). Figure 2.1 shows the distribution of schools across the country. Math tutoring took place in 137 academies, while English tutoring took place in 50 academies.

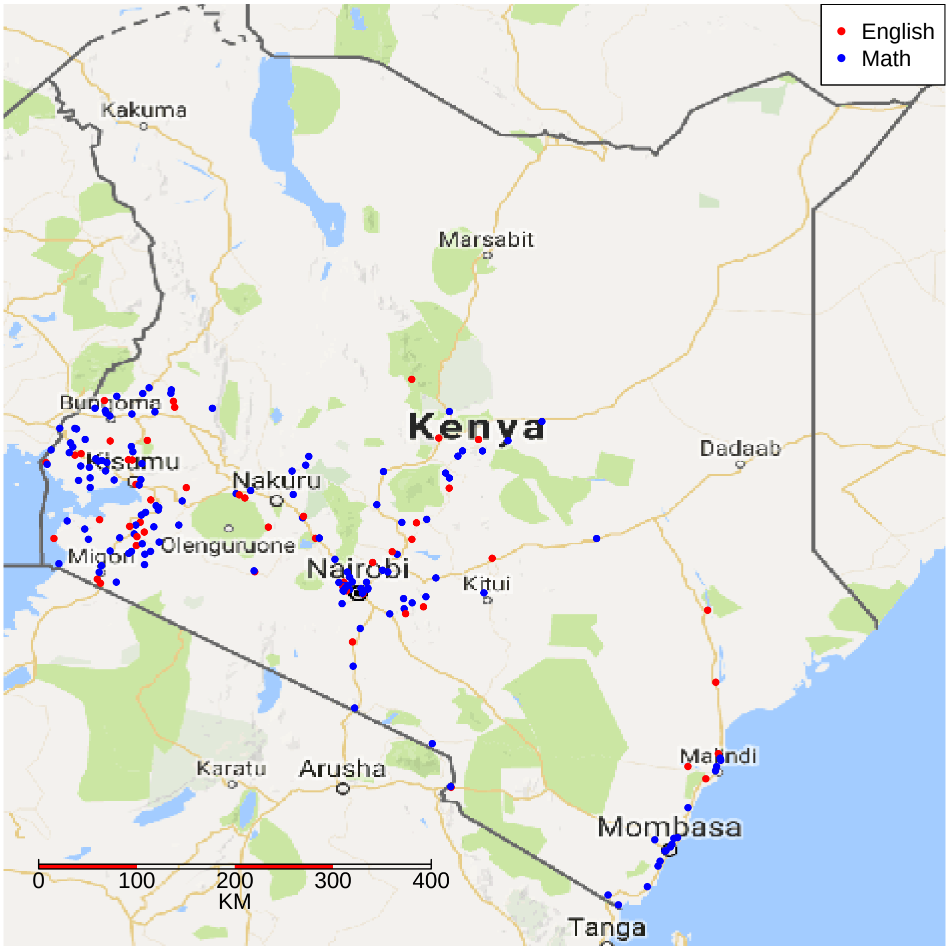


Figure 2.1: Schools with math and English tutoring

[6] Schools where a pilot of the program was tested during the 2015 academic year were excluded.

2.2.4 Data and summary statistics

As mentioned above, students were tested six times per academic year. Each academic year has three terms, and each term has a midterm and an endterm exam. Additionally, at the beginning of the academic year students in primary grades (Standards 1 - 6) took a diagnostic exam. Table 2.4 shows the dates of each exam. Two exams (T3ET15 and T1DG16) were taken by students before tutoring began, and six exams were taken after. Since students in Preunit, Nursery and Baby Class were not tested at the beginning of 2016 (T1DG16), we use both T1DG16 and T3ET15 as our baseline test scores. For students in Baby Class (BC) we have no baseline test scores.

All students at each grade level across schools in Bridge's network take the same exam, making test scores for students in different schools comparable. However, the exams are not vertically linked (i.e., there are no overlapping questions across exams for each grade level), and therefore we standardized test scores in each term (such that in English tutoring schools the mean score is zero and the standard deviation is 1).

Table 2.4: Learning assessments

Year	Term	Exam	Dates	Code
2015	3	Endterm	11/10/2015 - 11/12/2015	T3ET15
2016	1	Diagnostic	1/13/2016 - 1/14/2016	T1DG16
2016	1	Midterm	2/16/2016 - 2/18/2016	T1MT16
2016	1	Endterm	4/5/2016 - 4/7/2016	T1ET16
2016	2	Midterm	6/14/2016 - 6/16/2016	T2MT16
2016	2	Endterm	8/9/2016 - 8/11/2016	T2ET16
2016	3	Midterm	9/26/2016 - 9/27/2016	T3MT16
2016	3	Endterm	10/25/2016 - 10/27/2016	T3ET16

Schools randomly assigned to math tutoring are similar to those assigned to English tutoring: They were inaugurated around the same time (in operation for two years by January 1, 2016), and have similar teacher salaries and pupil-teacher ratios

(PTR) of 22 students per teacher (see Table 2.5). Pupils (Table 2.6) in English and math tutoring schools are similar across all characteristics.[7] Tutors (Table 2.7) are also similar across English and math tutoring schools. On average, pupils are 6.5 years old and tutors are 4.5 years older than their tutees.

[7] Except for Science and Social Sciences in T1DG16 (see Table B.1) where students in English tutoring schools seem to be doing better. However, neither of these subjects is the focus of tutoring. Moreover, when correcting for multiple-hypothesis testing, these differences are no longer significant.

Table 2.5: School characteristics in English and math tutoring schools

	(1) English Tutoring	(2) Math Tutoring	(3) Difference	(4) Difference (F.E)
Days since launch (as of Jan. 1, 2016)	672.960	693.310	20.347	-16.257
	(406.417)	(405.017)	(66.887)	(46.838)
Monthly teacher wage of 11,250 KSH	0.060	0.110	0.049	0.014
	(0.240)	(0.313)	(0.043)	(0.026)
Monthly teacher wage of 10,400 KSH	0.180	0.100	-0.078	-0.073
	(0.388)	(0.304)	(0.061)	(0.061)
Monthly teacher wage of 7,970 KSH	0.760	0.790	0.028	0.058
	(0.431)	(0.410)	(0.070)	(0.065)
Teachers	7.440	7.530	0.093	0.077
	(0.541)	(0.619)	(0.093)	(0.092)
Enrollment (beginning of the school year)	167.760	167.180	-0.585	-2.554
	(75.793)	(84.627)	(12.894)	(11.451)
PTR	22.240	21.980	-0.257	-0.478
	(9.363)	(10.367)	(1.589)	(1.401)

Days since launch: number of days that have passed since the schools opened, as of January 1, 2016. Bridge has three teacher wage levels. Monthly teacher wage indicates the proportion of schools within each wage schedule. Teachers is the number of teachers at the school. The enrollment is measured across all grades at the school at the beginning of the school year. PTR is the pupil-teacher ratio. Each row presents the mean for schools which receive English tutoring (Column 1), schools which receive math tutoring (Column 2), the difference between the two (Column 3), and the difference taking into account the randomization design (Column 4). In the first two columns the standard deviation is shown in parentheses, while in the third and fourth columns the standard error of the difference is in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.6: Pupil characteristics: ECE, Grade 1 and Grade 2

	(1) English	(2) Math	(3) Difference	(4) Difference (F.E)
Panel A: Tutees' time-invariant characteristics				
Age	6.600	6.500	-0.097*	-0.024
	(1.617)	(1.595)	(0.054)	(0.037)
Male	0.520	0.520	0.002	0.000
	(0.500)	(0.500)	(0.011)	(0.010)
Age entered Bridge	5.440	5.390	-0.057	0.013
	(1.669)	(1.643)	(0.076)	(0.073)
Panel B: Tutees' test-scores in T3ET15				
English (Reading)	0.000	-0.010	-0.013	-0.058
	(1.000)	(1.021)	(0.074)	(0.071)
English (Writing)	0.000	-0.040	-0.038	-0.064
	(0.999)	(1.014)	(0.064)	(0.058)
Swahili (Reading)	0.000	-0.020	-0.025	-0.064
	(1.000)	(1.020)	(0.083)	(0.082)
Swahili (Writing)	0.000	-0.070	-0.072	-0.113
	(1.000)	(1.102)	(0.111)	(0.090)
Math	0.000	0.040	0.041	0.011
	(0.999)	(0.974)	(0.056)	(0.052)
Science	0.000	-0.070	-0.069	-0.098
	(1.000)	(1.009)	(0.089)	(0.081)
S.S.	0.000	-0.020	-0.018	-0.055
	(1.000)	(1.030)	(0.093)	(0.084)

Math, Language (English), Swahili, Science, and S.S. (Social Sciences) represent the standardized test scores (mean zero and standard deviation 1 in English tutoring schools). Each row presents the mean for schools that receive English tutoring (Column 1), schools that receive math tutoring (Column 2), the difference between the two (Column 3), and the difference taking into account the randomization design (Column 4). In the first two columns the standard deviation is shown in parentheses, while in the third and fourth columns the standard error of the difference is in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We have an unbalanced panel, where very few students have test score data for all periods. This is due to a combination of lack of compliance, software issues and network failures in which teachers enter the data but it is not uploaded to servers at

Table 2.7: Tutors's characteristics: Grade3 - Grade 7

	(1) English	(2) Math	(3) Difference	(4) Difference (F.E)
Panel A: Tutors' time-invariant characteristics				
Age	11.040	11.070	0.030	0.023
	(1.980)	(2.017)	(0.097)	(0.062)
Male	0.500	0.520	0.020^{**}	0.023^{***}
	(0.500)	(0.500)	(0.009)	(0.008)
Age entered Bridge	9.660	9.710	0.053	0.045
	(2.269)	(2.316)	(0.140)	(0.098)
Panel B: Tutors' test scores in T3ET15				
English (Reading)	0.000	0.070	0.070	0.047
	(0.999)	(1.038)	(0.051)	(0.046)
English (Writing)	0.000	0.070	0.069	0.034
	(0.999)	(0.967)	(0.054)	(0.045)
Swahili (Reading)	0.000	0.050	0.055	0.053
	(0.999)	(1.042)	(0.056)	(0.046)
Swahili (Writing)	0.000	0.140	0.138^{*}	0.114^{*}
	(0.999)	(0.941)	(0.081)	(0.059)
Math	0.000	0.050	0.047	0.027
	(0.999)	(1.009)	(0.063)	(0.048)
Science	0.000	0.060	0.060	0.026
	(0.999)	(1.010)	(0.052)	(0.044)
S.S.	0.000	0.070	0.070	0.034
	(0.999)	(0.967)	(0.056)	(0.049)

Math, Language (English), Swahili, Science, and S.S. (Social Sciences) represent the standardized test scores (mean zero and standard deviation 1 in English tutoring schools). Each row presents the mean for schools that receive English tutoring (Column 1), schools that receive math tutoring (Column 2), the difference between the two (Column 3), and the difference taking into account the randomization design (Column 4). In the first two columns the standard deviation is shown in parentheses, while in the third and fourth columns the standard error of the difference is in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Bridge HQ.[8] Table 2.8 shows the fraction of students tested each time. More than 25% of the data is missing (and often more than 30%)[9]. In particular, the endterm exam

[8] In addition, students may have been absent from school on the day of the test. However, in most cases if test score data is missing for a student, it is also missing for their entire grade.

[9] For the purposes of this paper, the missing data numbers include pupils who enrolled but have not

in the second period (T2ET16) is missing over 60% of the test scores for math due to a glitch in a programming update. Additionally, a software problem prevented more than 25% of the academies from entering test-score data for T2ET16. Since this is noisy data, we remove it from our sample in the main text, but we provide robustness checks that include the data in Appendix B.1.

paid fees — and hence are not allowed to sit through classes — in a given period.

Table 2.8: Available data

	T1MT16	T1ET16	T2MT16	T2ET16	T3MT16	T3ET16	Total
Math	0.751	0.591	0.711	0.399	0.570	0.532	0.590
	(0.432)	(0.492)	(0.453)	(0.490)	(0.495)	(0.499)	(0.492)
English (Writing)	0.739	0.575	0.710	0.472	0.564	0.517	0.594
	(0.439)	(0.494)	(0.454)	(0.499)	(0.496)	(0.500)	(0.491)
English (Reading)	0.738	0.566	0.709	0.449	0.553	0.512	0.586
	(0.439)	(0.496)	(0.454)	(0.497)	(0.497)	(0.500)	(0.493)
Observations	192346						

Fraction of students for whom test-score data is available in math, English (reading), and English (writing) in each test.

Whether the data for a particular student is missing is orthogonal to whether that student is receiving math or English tutoring (see Table 2.9). Since attrition is high in any given period (over 30%) we do not perform Lee (2009) bounds as these are too wide to be informative.[10] Additionally, since a large number of students do not have baseline test scores we impute scores for those students and add a dummy variable to all our regressions for whether the baseline test score was inputted.

Table 2.9: Differential attrition: Students in math and English tutoring schools

	(1) Math	(2) English	(3) Swahili
Math tutoring	-0.0027	-0.0053	-0.0098
	(0.022)	(0.022)	(0.028)
Mean English	0.63	0.61	0.61
N. of obs.	81195	81209	55019
Number of schools	187	187	187

Differential attrition between students in math tutoring schools and students in English tutoring schools. The estimation data set does not include T2ET16 data; see Table B.3 for an estimation including that data. Clustered standard errors, by school, in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.3 Results

2.3.1 Treatment effects

In order to estimate the effect of tutoring on test scores we use the following specification:

[10] The T2ET16 testing rates are different across math and English tutoring (see Figure B.2). However, we believe this difference is merely coincidental.

$$Y_{isgd,t} = \alpha_0 + \beta T_s + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t} \tag{2.1}$$

where $Y_{isgd,t}$ is the test score of student i in grade g at school s located in province d at time t (and $Y_{isgd,t=0}$ is his test score before treatment). γ_d is a set of province and strata fixed effects, γ_t are time fixed effects, and γ_g are grade fixed effects. X_i is a set of student time-invariant characteristics (month of birth and gender), and X_s are school characteristics at baseline (pupil-teacher ratio, monthly school fees and teachers' wages). T_s indicates whether the student is in a school with a math tutoring program (if not, he is in a school with English tutoring). Standard errors are clustered at the school level. β is the coefficient of interest here and estimates the effect of math tutoring on test scores compared to English tutoring. This specification assumes that the treatment effect (β) is time-invariant and grade-invariant (in Section 2.3.2 we relax these assumptions).

Tutees

Math tutoring has a small positive effective of 0.6 SD on math scores (see Table 2.10, Column 1). However, English tutoring (or the lack of it) has no effect on English test scores–we can rule out an effect bigger than $0.077SD$ with a confidence of 95% (Table 2.10, Column 2). Neither math nor English tutoring seem to have an effect on Swahili (see Table 2.10, Column 3).

Tutors

We can rule out an effect greater than 0.09 SD with a confidence level of 95% in math (for the math tutoring program). Similarly, we can rule out an effect greater than 0.06 SD with a confidence level of 95% in English (for the English tutoring program).

Table 2.10: Effect on tutees' test scores

	Tutees			Tutors		
	(1) Math	(2) English	(3) Swahili	(4) Math	(5) English	(6) Swahili
Math tutoring	0.063^{*}	-0.0061	0.035	0.029	-0.019	-0.020
	(0.034)	(0.035)	(0.047)	(0.031)	(0.035)	(0.036)
N. of obs.	50424	48204	32736	48741	46938	46512
Number of schools	187	187	186	187	187	187

The independent variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does not include T2ET16 data. See Tables B.4 and B.5 for versions of these estimates that include T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

See columns 4 and 5 in Table 2.10 for details.

2.3.2 Heterogeneity

In this section we test for heterogeneous treatment effects in tutees.[11] Overall, the math tutoring program is most effective after the first quarter (except for T2ET16, the exam with a high attrition rate and therefore unreliable results). We also find that math tutoring is most effective for students in the middle of the ability distribution at baseline. We do not find any heterogeneity by grade, age, gender, average tutor characteristics (age, gender, baseline test scores), or average school characteristics (pupil-teacher ratio, school size, or tutor-tutee ratio).

[11] Results for tutors are available upon request.

Periods

In order to estimate the effect of tutoring on test scores across time we use the following specification:

$$Y_{isgd,t} = \alpha_0 + \sum_{\tau=1}^{6} \beta_\tau T_s \times 1_{t=\tau} + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \tag{2.2}$$

where β_1 measures the treatment effect in period T1MT15, β_2 measures the treatment effect in T1ET15, and so on, until β_6 which measures the treatment effect in period T3ET15. The treatment effect for math (of math tutoring relative to English tutoring) increases after the first marking period (except for T2ET16, the period with a high attrition rate). On the other hand, math tutoring, relative to English tutoring, does not seem to have a negative effect on English test scores, with point estimates close to zero after the first marking period. Figure 2.2 provides additional details.

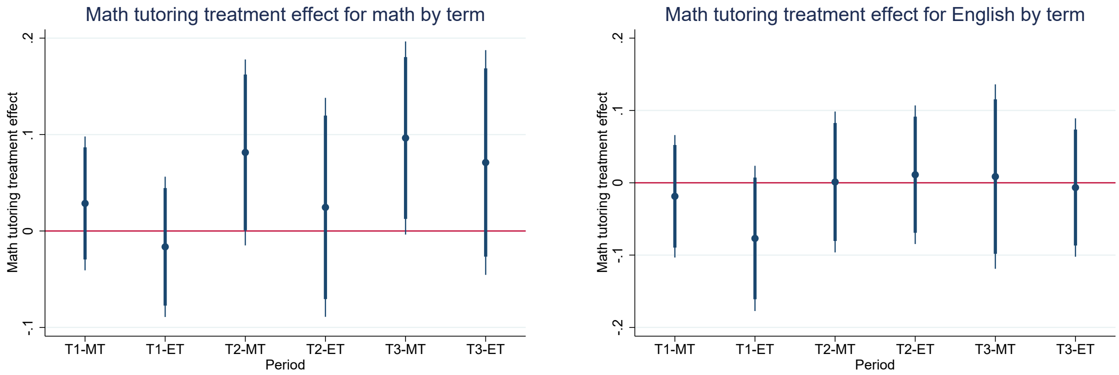


Figure 2.2: Evolution of the treatment effect of math tutoring, relative to English tutoring, on math (left panel) and English (right panel) test scores. Bars represent 90% and 95% confidence intervals (thick lines and thin lines, respectively).

Grade

In order to estimate the effect of tutoring on test scores across time we use the following specification:

$$Y_{isgd,t} = \alpha_0 + \sum_{\tau=1}^{5} \beta_\tau T_s \times 1_{t=\tau} + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \quad (2.3)$$

where β_1 measures the treatment effect for BC, β_2 for NU, β_3 for PU, β_4 for STD 1 and β_5 for STD 2. Although the point estimate for STD 2 is the largest in math, there does not seem to be a systematic pattern in which the oldest students benefit more than younger ones from math tutoring, and we cannot reject the hypothesis that the effect is the same across grades. Similarly, there seems to be no systematic pattern in the effect on English test scores. Figure 2.3 provides more details.

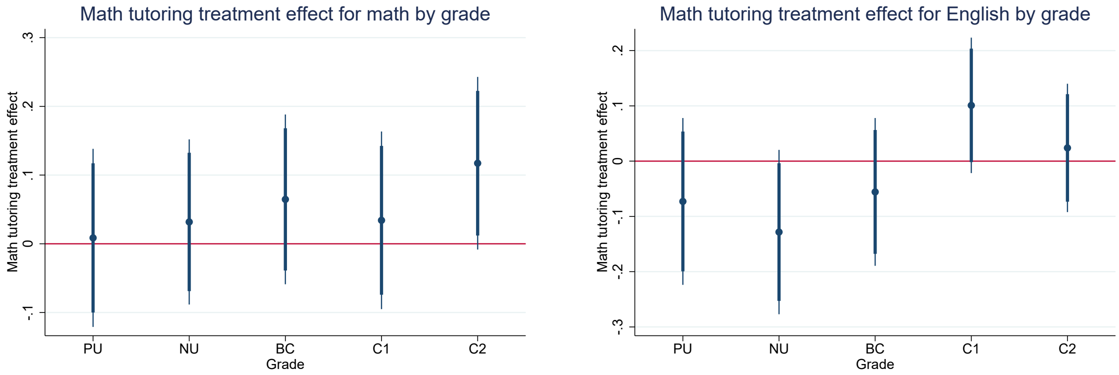


Figure 2.3: Treatment effect of math tutoring, relative to English tutoring, in math (left panel) and English (right panel) test scores by grade. Bars represent 90% and 95% confidence intervals (thick lines and thin lines, respectively).

Baseline test scores

In order to estimate the effect of tutoring across baseline test scores we use the following specification:

$$Y_{isgd,t} = \alpha_0 + \sum_{i=0}^{10} \beta_i T_s \times c_i + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \qquad (2.4)$$

where c_i is the decile of the student's test score in math in T3ET15. We have 11 categories for c_i: 10 deciles and a category for those students with missing test scores. Figure 2.4 shows the estimates for all the βs which correspond to the treatment effect for students in a given category. The effect for students with missing test scores is similar to the average treatment effect (0.06 SD). Although the treatment effect is positive for all students (except for students in the top 10% at baseline for which there is a very small, insignificant, negative effect), students in the middle of the distribution benefit more from the math tutoring (0.15 SD compared to the average effect of 0.06 SD). This is consistent with: a) tutors not being able to help students who are advanced learners and need an instructor with a high level of expertise to guide them through more advanced concepts; and b) tutors not being able to help tutees lagging behind grade level competencies who may need more specialized instruction to catch up. Along those lines one might expect that low achieving tutors might benefit from reviewing material they do not master completely. Figure B.3 in Appendix B shows that this is not the case. The effect is indistinguishable from zero for all tutors, regardless of baseline test scores, without any discernible pattern.

Tutee, tutor and school characteristics

In order to estimate the effect of tutoring on test scores across tutee, tutor and school characteristics we use the following specification:

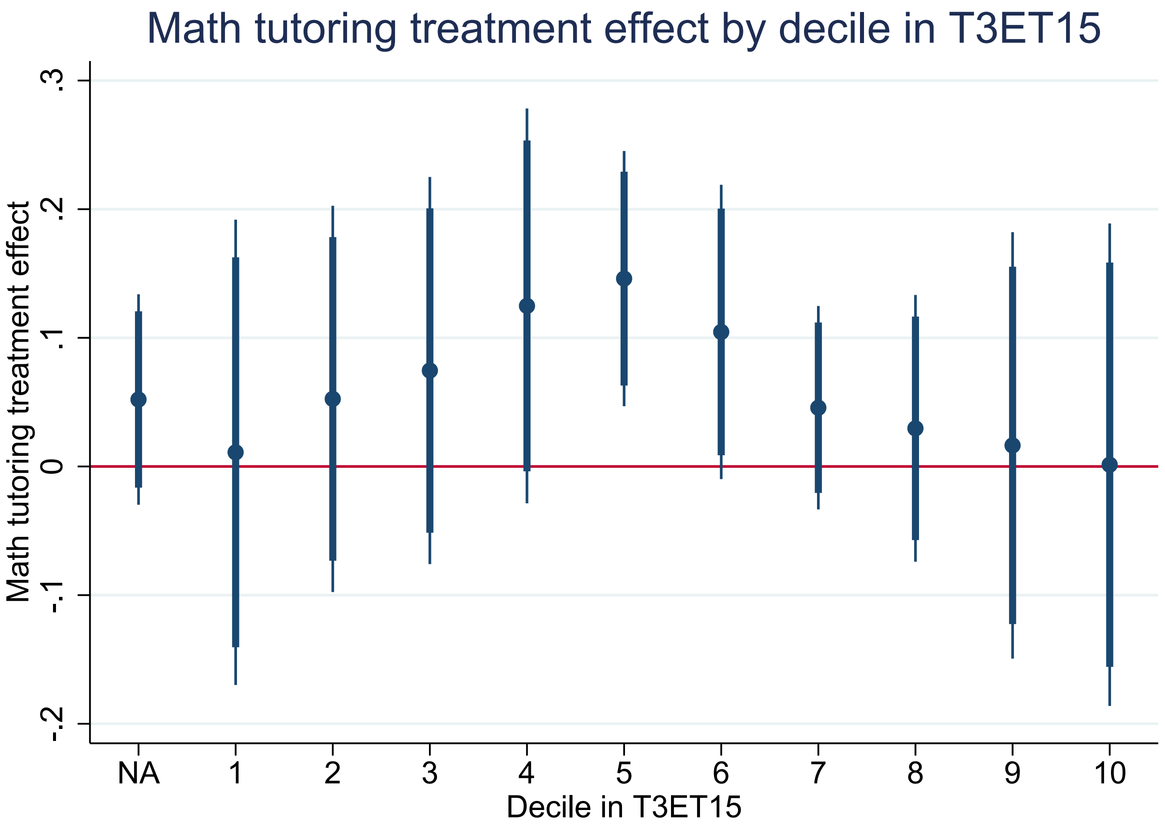


Figure 2.4: Treatment effect of math tutoring, relative to English tutoring, on math test scores by ability decile in T3ET15. Bars represent 90% and 95% confidence intervals (thick lines and thin lines, respectively).

$$Y_{isgd,t} = \alpha_0 + \beta_1 T_s + \beta_2 T_s \times c_i + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \quad (2.5)$$

where c_i denotes the characteristics along which we wish to measure heterogeneity and β_2 allows us to test whether there is any differential treatment effect. Since we do not know how teachers matched students we can only measure heterogeneity across the average characteristics of all the possible tutors a tutee might have (e.g., all the Standard 5 students for Pre-Unit tutees). Table 2.11 show the results from estimating

β_2 across different characteristics.[12] The first three columns show heterogeneity by student characteristics, the middle three columns by the average characteristics of all the possible tutors, and the last three columns by school characteristics.

There is evidence of heterogeneity by tutee's age (see Column 1). Specifically, older students benefit more from math tutoring. In this context the age distribution in each grade has wide tails and they often overlap (see Figures B.1 in Appendix B). There is no heterogeneity by the tutee's gender (see Column 2). This stands in contrast to several education interventions in which girls benefit more.[13] Students that join Bridge later benefit the most from math tutoring, consistent with the idea that tutoring allows students lagging behind to catch up (see Column 3).

Columns 4-6 show that there is no differential effect by tutors' average age, gender, or baseline test score (a PCA index across all subjects), while Columns 7-9 show that there is no differential effect by the tutors' pupil-teacher ratio (PTR), tutee-tutor ratio (TTR) or school size (number of enrolled students).

[12] Table 2.11 has results for math test scores. Table B.2 has the results for English test scores.

[13] For example, Anderson (2008) reviews several early childhood interventions with larger effects for girls and Chetty, Friedman, and Rockoff (2014) report larger impacts of teacher quality on girls than boys.

Table 2.11: Heterogeneity: Math test scores

	Tutee characteristics			Tutor characteristics			School characteristics		
	(1) Age	(2) Male	(3) Age joined Bridge	(4) Age	(5) Male	(6) Score in T3ET15	(7) PTR	(8) TTR	(9) Enrollment
Panel A: Math									
Math tutoring × Covariate	0.023*	-0.026	0.023*	0.018	-0.18	-0.024	0.0026	0.034	0.00034
	(0.013)	(0.029)	(0.012)	(0.019)	(0.22)	(0.032)	(0.0038)	(0.029)	(0.00050)
Observations	50820	50934	50820	50538	50538	40891	50934	50913	50934
Adjusted R^2	0.222	0.220	0.221	0.220	0.220	0.230	0.220	0.220	0.220

The independent variable is the standardized math test score (mean 0 and standard deviation of 1 in English tutoring schools). Each column shows heterogeneity by a different covariate. The covariates in columns 1-3 are the tutee's age (in 2016), gender, and the age at which they joined Bridge (in 2016). The covariates used in columns 4-6 are tutors' average characteristics (age, gender and test scores at baseline). Columns 7-9 include school-level characteristics (pupil-teacher ratio (PTR), tutee-tutor ratio (TTR), and number of enrolled students). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does not include T2ET16 data. See Table B.6 for a version of this table that includes T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.4 Conclusions

There is an increasing wealth of evidence showing that teaching appropriate to a student's learning level can improve learning outcomes in low-income countries. However, teachers often lack the time (or incentives) to give each child personalized instruction tailored to their needs and providing schools with additional teachers to do so is expensive. Cross-age tutoring, where older students tutor younger students, is an inexpensive alternative to providing personalized instruction to younger students in that it substitutes a trained instructor (the teacher) with an untrained one (the older student) at the cost of the older student's time.

We present results from a large randomized control trial (over 180 schools, 15,000 tutees, and 15,000 tutors) in Kenya, in which schools are randomly selected to implement a cross-age tutoring program in either English or math. Cross-age tutoring in math, relative to tutoring in English, has a small positive effect (0.06 SD, p-value of 0.073) on math test scores. These results do not hold true for English tutoring, however: relative to math tutoring, it has no positive effect on English test scores (we can rule out an effect of 0.077 SD with 95% confidence). There is considerable heterogeneity by the students' baseline learning levels. Specifically, the effect is largest for students in the middle of the ability distribution (0.144 SD, p-value of 0.005), while the point estimates are almost zero for students with either very low or very high baseline learning levels. This is consistent with: a) tutors not being able to help students who are advanced learners and need an instructor with a high level of expertise to guide them through more advanced concepts; and b) tutors not being able to help tutees lagging behind grade level competencies who may need more specialized instruction to catch up. Surprisingly, we find no heterogeneity by school characteristics or tutor characteristics.

These results translate into some policy conclusions. First, cross-age tutoring is more effective for math than languages. Second, these types of interventions can help average students, but not stellar students or students who are really struggling and need more skilled assistance. Third, although the program has modest effect sizes, it is essentially free, and therefore cost-effective relative to several other alternatives for providing personalized instruction. For example, contract teachers have been shown to increase student learning by 0.26 σ in Kenya (Duflo, Dupas, and Kremer 2015) and 0.16 σ in India (Muralidharan and Sundararaman 2013). Cross-age tutoring is akin to the contract teacher approach, in which teachers that are not professionally trained are hired, by delegating older kids to teach. Contract teachers have been found to increase test scores by 1.97 SD per 100 USD invested (Kremer, Brannen, and Glennerster 2013).[14] The total cost of this intervention was 97,000 USD for both the math and the English tutoring program.[15]. While only 187 schools (over 15,000 tutees) participated in the field experiment, over 400 schools implemented the program (i.e., over 32,000 students). Thus, the total cost of the program is around 3 USD per student, which translates into test score increases of 18 SD per 100 USD invested.[16]

Chapter 2, in full, is currently being prepared for submission for publication of the material. Romero, Mauricio; Chen,Lisa; Magari, Noriko. “Cross-Age Tutoring: Experimental Evidence from Kenya”. The dissertation author was the primary investigator and author of this material.

[14] See https://www.povertyactionlab.org/policy-lessons/education/increasing-test-score-performance for cost-effectiveness comparisons across interventions.

[15] This includes the cost of the original pilot, the development and testing of teaching guides for tutors, and the monitoring of the program.

[16] The cost of implementing the program in future years is projected to decrease as the bulk of the cost was a fix investment: development of teaching guides for tutors.

Chapter 3

Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania

(Co-authors: Isaac Mbiti, Karthik Muralidharan, Youdi Schipper, Constantine Manda, and Rakesh Rajani)

3.1 Introduction

The idea that complementarities across policies and programs can lead to increasing returns from joint provision has been posited in several economic settings, and has been a central theme in development economics (Johnston and Mellor 1961; Ray 1998). This belief is implicit in the design of prominent anti-poverty programs, such as the Millennium Villages Project (Sachs 2006; Munk 2013) and the graduation programs for ultra-poor households (Banerjee et al. 2015a; Bandiera et al. 2017).[1] Yet most empirical research in economics has focused on estimating the impact of single or bundled interventions and there is limited experimental evidence of complementarities to date.

This approach is exemplified by the economics of education literature, where the growing use of randomized experiments has allowed researchers to credibly study the effects of a wide range of education interventions (Muralidharan 2017; Fryer 2017). Yet, while this literature has successfully studied the impacts of several individual education interventions, it has typically not tested for complementarities across them.[2] Since policies are rarely implemented in isolation, ignoring complementarities (if they exist) may lead to misleading inference about the marginal effectiveness of policies in practice.

We test for complementarities among education policies using a large-scale randomized evaluation. Our study is set in Tanzania, a developing country where two key constraints to education quality are a lack of school resources and low teacher mo-

[1] Both sets of programs provide selected villages or individuals/households with a combination of physical capital, human capital, and ongoing engagement and support. While there is more evidence on the efficacy of graduation programs in raising the incomes of the poor, the design of both sets of programs are based on the likelihood of complementarities across program components in alleviating poverty.

[2] Evaluations of "bundled" interventions are inclusive of complementarities between specific components of the intervention. But such evaluations cannot test for the existence of complementarities in the absence of treatment arms with each of the individual components.

tivation and effort (World Bank 2012). We study the impact of two programs, designed to alleviate each of these constraints; as well as complementarities between them. The first one aimed to alleviate resource constraints by providing schools with grants that nearly *tripled* the per-student resources available to them (not including infrastructure and teacher salaries). The second one aimed to improve teacher motivation and effort by providing them with performance-based bonuses — based on the number of their students who passed basic tests of Swahili, math, and English. A teacher with average enrollment could earn up to 55% of base pay as a bonus. Both programs were implemented by Twaweza, a leading Tanzanian non-profit organization with a strong track record of working to improve education in East Africa, including conducting independent assessments of student learning in Kenya, Uganda, and Tanzania (Uwezo 2017).

We conducted the experiment in a nationally representative sample of 350 public schools across 10 districts in mainland Tanzania.[3] We randomly allocated schools to four groups (stratified within districts): 70 schools received unconditional school grants, 70 schools received the teacher performance pay program, 70 schools received *both* programs, and 140 schools were assigned to a control group. The study was powered adequately to test for complementarities, and we gave the same importance to testing for complementarities as testing for the main effects of the two programs.[4]

We report four main sets of results. First, the school grant significantly increased per-student expenditure in treated schools. Consistent with prior findings (as in Das et al. (2013)) we find evidence of crowding out of school and household spending in response to the extra school grant. Despite this reduction, there was a near doubling of

[3] See Heckman and Smith (1995) for a discussion of the threats to external validity of experiments resulting from non-random site selection in experimental studies. Allcott (2015) provides evidence of such site-selection bias. Muralidharan and Niehaus (2017) discuss the use and value of random assignment in representative samples for improving the external validity of experimental studies.

[4] Trial registry and pre-analysis plan available at https://www.socialscienceregistry.org/trials/291

net spending per student in treated schools (excluding teacher salaries). However, this increase in spending had *no impact* on student learning outcomes on low-stakes tests in Swahili (language), math, or English after both one and two years of the program. The estimates are precise and we can rule out effects larger than 0.1σ after two years.

Second, teacher performance pay improved student test scores. At the end of two years, students in treated schools were 37%, 17%, and 70% more likely to pass independently-administered high-stakes tests in math, Swahili, and English — the outcome that teacher bonuses were based on. These correspond to a 7.7, 7.3, and 2.1 percentage-point increase in the passing rate relative to control group means of 21%, 44%, and 3% in these subjects. These also correspond to a test-score increase of 0.21σ across subjects in treated schools. However, in a parallel set of low-stakes tests administered by the research team, we find no impact on test scores in incentive schools. Since Twaweza employed strict protocols to minimize cheating on the high-stakes tests (see Section 3.4.2), the differences between high-stakes and low-stakes testing are most likely explained by greater student effort on the high-stakes tests (with the magnitude of the difference being comparable to that reported by Levitt et al. (2016) and Hirshleifer (2017)).

Third, students in schools that received both inputs and incentives had significantly higher test scores (relative to the control group) in all subjects on both the high-stakes and low-stakes tests. At the end of two years, student passing rates on the high-stakes tests (which determined the teacher bonus payments) were 49%, 31%, and 116% higher in math, Swahili, and English (a 10.3, 13.6, and 3.5 percentage-point increase relative to the control means of 21%, 44%, and 3%). Student test scores on these high-stakes tests were 0.36σ higher than those in control schools. After two years, test scores were also 0.23σ higher on the low-stakes tests administered by the research team.

Fourth, we find strong evidence of complementarities between inputs and incentives. At the end of two years, the test score gains in the "combination" group were significantly higher than the sum of the gains in the "school grant" and the "teacher incentives" groups in *each* of the three subjects (math, Swahili, and English). Using a composite measure of test-scores across subjects, the "interaction" effect was equal to 0.18σ ($p < 0.01$).[5] In short, school inputs are more effective when teachers have incentives to use them effectively. Conversely, motivated teachers (either intrinsically or through incentives) can be more effective with additional educational inputs.

To help interpret our results, we present a simple theoretical framework that specifies an education production function and a teacher's optimization problem regarding how much effort to exert. The key insights from the model are the following: First, the observed effects of policy changes (like providing school inputs) depend not just on the production function but also on changes in effort induced by the policy change. Second, even if there are complementarities in the production function between inputs and effort, if teachers act like agents in standard economic models (with disutility from effort and no intrinsic motivation), then the optimal response to an increase in inputs may be to reduce effort, which may attenuate impacts on learning. Third, the introduction of financial incentives will typically raise the optimal amount of teacher effort when inputs increase (due to production function complementarities). While our results can potentially be explained by other models as well, this framework provides a parsimonious way to interpret our results as well as existing results in the literature.

We make several contributions to research and policy. First, we confirm and replicate results from several randomized evaluations of education in developing

[5] We test for complementarities only using the low-stakes tests administered by the research team and not the high-stakes tests administered by Twaweza, because the latter were not conducted in grant schools where there was no need to make any incentive payments. However, the difference between the "combination" group and the "incentives only" group (which estimates the effect of the grant *and* the complementarities) is similar for both the high-stakes and low-stakes tests, suggesting that the magnitude of the complementarities is similar across high-stakes and low-stakes tests (see section 3.5).

countries showing that augmenting school resources on their own seems to have very little impact on learning outcomes. These include Glewwe, Kremer, and Moulin (2009) in Kenya, Blimpo et al. (2015) in Gambia, Das et al. (2013) in India, Pradhan et al. (2014) in Indonesia, and Sabarwal, Evans, and Marshak (2014) in Sierra Leone. While the specific reasons for non-impact discussed in these papers vary (see discussion in Glewwe and Muralidharan (2016)), our results reinforce the point that the default "input-focused" approach to improving education in developing countries is unlikely (on it's own) to be very effective at improving learning outcomes.

Second, we replicate and validate the findings of Muralidharan and Sundararaman (2011b) on the positive effects of teacher performance pay on student learning in a different developing country setting. Specifically, the 0.21σ increase in test scores that we find after two years in Tanzanian primary schools is similar to the 0.22σ increase found after two years in India.[6] Our results are also consistent with those of Lavy (2002), Lavy (2009), Glewwe, Ilias, and Kremer (2010), Duflo, Dupas, and Kremer (2015), and Contreras and Rau (2012) who find that various forms of performance linked pay for teachers in low and middle-income countries improved student test scores. Overall, our results confirm that there is enough slack in teacher effort in developing countries that even modest amounts of performance-linked pay can improve learning outcomes.

Third, our most original contribution is to experimentally establish the existence of complementarities across policies to improve human capital, which (to the best of our knowledge) has not been shown to date. While several experimental studies in development economics have employed factorial or cross-cutting designs, these have typically *assumed away* complementarities to increase power in estimating the effects of the main treatments of interest (see Muralidharan, Romero, and Wuthrich (2018) for a

[6] We focus the comparison on the impacts on high-stakes tests because the study in India is also based on impacts on high-stakes tests. Both settings featured external testing with strict protocols to prevent cheating. However, finding no effect on low-stakes tests suggests that day-of-test effort by students and teachers may contribute non-trivially to measured test scores as shown by Levitt et al. (2016)

review). Other studies have evaluated variants of interventions that include basic and augmented versions of a program that feature the evaluation of variants A, and A + B; but not A, B, and A + B, which would be needed to test for complementarities (for instance, see Pradhan et al. (2014) and Kerwin and Thornton (2017)). The closest study that was explicitly designed to test for complementarities is Attanasio et al. (2015) which studies the effects of providing (1) nutrition supplements, (2) stimulation programs, and (3) both of them, on early childhood development in Colombia, and finds no evidence of complementarities across the two programs studied.[7]

Our results are relevant for the design of development interventions more generally, and for education policy in particular. They suggest that a binding constraint to the effective use of additional resources for service delivery may be the lack of adequate motivation and effort of front-line service providers (Chaudhury et al. 2006). Thus, implementing some form of teacher performance-pay may also raise the effectiveness of existing education inputs in developing countries (see Section 1.2.1 for a discussion of cost-effectiveness). More generally, our results also speak to the promise of similar policy approaches in the US, where several states are proposing to link parts of school financing to performance on state-wide tests - an approach that may generate complementarities of the sort we find (Collier 2016a; Mesecar and Soifer 2016; Calefati 2016).

[7] A second study that could in principle test for complementarities is Behrman et al. (2015) who study the impacts of providing (1) student incentives, (2) teacher incentives, and (3) both of them, on learning of high school students in Mexico. Yet, that study is not able to do so because the variant of student incentives provided were not the same across treatments (1) and (3), and so (3) was not the same as (1) + (2).

3.2 Theoretical framework

We present a simple model of how teachers choose effort. It shows how changes in inputs and incentives translate into changes in teacher effort and student learning outcomes. The model has three main goals.

First, it clarifies that the impact of an education intervention on learning outcomes will depend on both the production function *and* behavioral responses by teachers (and parents). In other words, experiments will typically identify the "policy effect" of an intervention and not the "production function" parameters. Second, it is only under the implicit (and often unstated) assumption that teachers are intrinsically motivated that increasing inputs should be expected to improve test scores. In contrast, if teachers behave like agents in standard economic models (with disutility of effort and no intrinsic utility from their job), then increasing inputs may lead to a *reduction* of effort and no change in learning. Finally, if there are complementarities between effort and inputs in the production function, then providing incentives to teachers may *raise* the optimal effort when inputs are increased, giving rise to policy complementarities between providing inputs and incentives that may be even stronger than the production function complementarities between inputs and teacher effort.

Formally, we model teachers' choice of effort (e) as solving the following problem:

$$\max_e U_i(e) = W + \lambda_i \Delta L - c_i(e) \tag{3.1}$$

subject to

$$W = S + b\Delta L \tag{3.1a}$$

$$\Delta L = f(e, I) \tag{3.1b}$$

$$\Delta L \geq \underline{\Delta L} \geq 0 \tag{3.1c}$$

where W is total earnings, which is equal to a base salary (S) plus a bonus ($b\Delta L$) proportional to gains in students' learning ΔL (b is typically zero in practice). λ_i is a measure of the teacher's intrinsic utility from improving student learning. Teacher effort, together with other inputs (I), translates into learning gains via f, which is strictly increasing in both arguments ($f_e > 0$ and $f_I > 0$), concave in each argument ($f_{ee} < 0$ and $f_{II} < 0$), and features complementarity between effort and inputs ($f_{eI} > 0$). Effort entails a cost, c_i, which is increasing and convex ($c_i'(\cdot) > 0$ and $c_i''(\cdot) > 0$). We allow λ_i and c_i to vary across teachers (indexed by i) to account for teacher heterogeneity. Finally, we assume that learning gains cannot be negative and have to be over a minimum level ($\underline{\Delta L}$). This can be interpreted as the minimum level of learning (including that taking place outside the school) required for teachers to not be sanctioned by parents or supervisors.[8]

Let $e_{min}(I)$ be the effort required to achieve $\underline{\Delta L}$ at a level of inputs equal to I (i.e., $f(e_{min}, I) = \underline{\Delta L}$). Let $e^*_{mc}(I)$ be the effort at which the marginal cost of effort is equal to its marginal benefit (i.e., $(\lambda_i + b) f_e(e^*_{mc}, I) = c_i'(e^*_{mc})$). Thus, the level of effort chosen will be $e^*(I) = \max(e_{min}(I), e^*_{mc}(I))$.

Figure 3.1 shows the optimal level of effort and learning gains associated with different levels of $\lambda_i + b$, and with different levels of inputs. In the absence of incentives or intrinsic motivation (i.e., $\lambda_i + b = 0$), it is Equation 3.1c that binds, and $e^*(I) = e_{min}(I)$. Thus, if $\lambda_i + b = 0$, then the marginal cost of effort is above the marginal benefit in equilibrium.[9] Effort does not change as b increases up to the point where the marginal benefit ($\lambda_i + b$) is equal to the marginal cost of providing effort. This corresponds to the flat region to the left of κ in Figure 3.1a at low levels of $\lambda_i + b$.

[8] $\Delta L \geq \underline{\Delta L} \geq 0$ can also be motivated by intrinsic motivation considerations with teachers experiencing disutility if outcomes are too low. This is a variant of Holmstrom and Milgrom (1991) where teachers have a minimum outcome threshold as opposed to a minimum effort threshold below which they experience disutility. In this case, $\underline{\Delta L}$ would also vary by teacher.

[9] If $\lambda_i > 0$ the qualitative results do not change as long as λ_i is low enough that Equation 3.1c binds, leading to $e^*(I) = e_{min}(I)$.

In the absence of incentives and for low values of λ_i (such that $b + \lambda_i$ is near zero), an increase in inputs will lead teachers to re-optimize and decrease the effort they exert. The intuition is straightforward: If inputs increase, teachers can achieve the required minimum $\underline{\Delta L}$ with lower effort. This is consistent with evidence from multiple settings showing that teachers in developing countries reduce effort when provided with more resources.[10] Since the binding constraint for effort continues to be Equation 3.1c, the increase in inputs would lead to a reduction of effort to the point that allows $\underline{\Delta L}$ to be achieved, and there would be no net gain in learning as seen in Figure 3.1b.

Thus, in the absence of incentives for improving learning outcomes, the relationship between extra inputs and improved test scores will depend on the distribution of intrinsic motivation (λ_i) in the population of teachers. In settings where λ_i is high for most teachers, improving school inputs may improve test scores.[11] Increasing inputs lowers the threshold (κ) that $\lambda_i + b$ needs to exceed for Equation 3.1c to not bind, and for effort to increase (because $f_{eI} > 0$). This is another channel through which increasing inputs could increase teacher effort and test scores (as seen in Figure 3.1a, where $\kappa_1 < \kappa_0$ when $I_1 > I_0$). However, in settings where λ_i is low for most teachers (such as many developing countries with high levels of teacher absence), this may be less likely (since $\lambda_i + b = 0$ may still be below κ_1).

If additional inputs are combined with performance-linked pay that increases b, then the distribution of $b + \lambda_i$ is shifted to the right, and for any given distribution of λ_i it is more likely that teachers are shifted to the right of κ_1 and find it optimal to

[10] For instance, Duflo, Dupas, and Kremer (2015) find that providing a randomly selected set of primary schools in Kenya with an extra contract teacher led to an *increase* in absence rates of teachers in treated schools. Muralidharan and Sundararaman (2013) find the same result in an experimental study of contract teachers in India. Finally, Muralidharan et al. (2017) show, using panel data from India, that reducing pupil-teacher ratios in public schools was correlated with an increase in teacher absence.

[11] For instance, Jackson, Johnson, and Persico (2016) find positive effects of school spending on education outcomes in the US, but this is a context where the default level of teacher effort may be higher than in developing countries.

increase effort.[12] Further, as discussed above, to the right of κ_1, the optimal amount of effort is higher at higher levels of inputs (i.e., $e_I^*(I_1) > e_I^*(I_0)$ if $b + \lambda_i > \kappa_1$). Thus, as long as Equation 3.1c is not binding, the complementarity in the production function ($f_{eI} > 0$) will also yield complementarities in the policy effects.

3.3 Context and Interventions

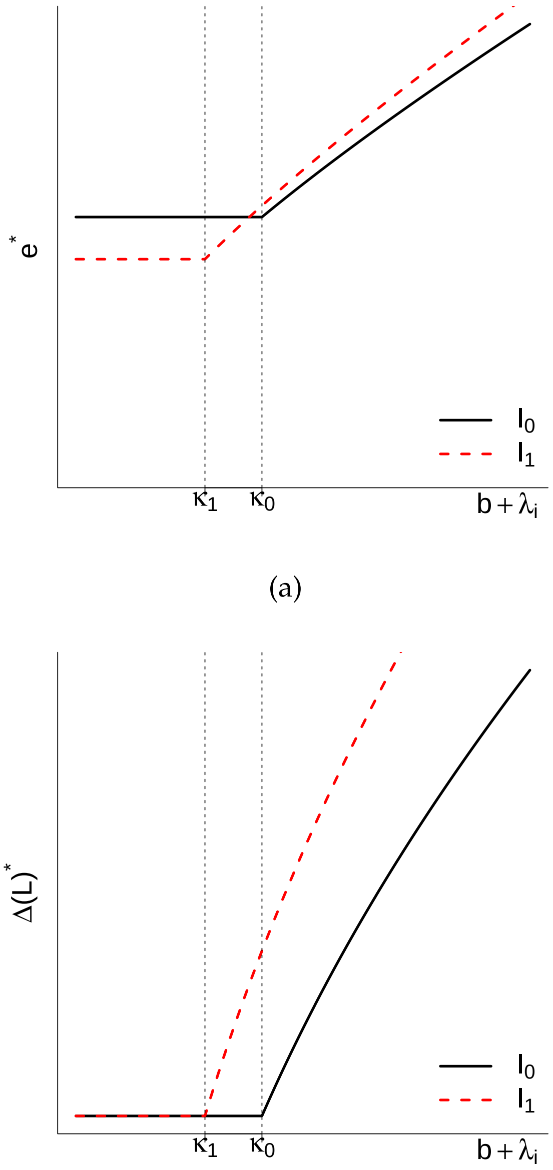
3.3.1 Context

Our study is set in Tanzania, which is the sixth largest African country by population, and home to over 50 million people. Partly due to the abolishment of school fees in public primary schools in 2001, Tanzania has made striking progress towards universal primary education with net enrollment growing from 52% in 2000 to over 94% in 2008 (Valente 2015). Yet, despite this increase in school enrollment, learning levels remain low. Recent nationwide learning assessments showed that less than one-third of grade 3 students were proficient at a grade 2 level in Kiswahili (the medium of instruction) literacy, or in basic numeracy. Proficiency in English (the medium of instruction in secondary schools) was especially limited, with less than 12% of grade 3 students able to read at a grade 2 level in English (Uwezo 2013; Jones et al. 2014).

Despite considerable public spending on education,[13] budgetary allocations to education (and actual funds received by schools) have not kept pace with the rapid

[12] While in theory it is possible that the provision of financial incentives for performance may crowd out intrinsic motivation (Deci and Ryan 1985; Fehr and Falk 2002), it is also possible that the opposite is true and that financial incentives can crowd in intrinsic motivation by reinforcing the value of the task (Mullainathan 2005). Empirical evidence from education in developing countries suggests that performance-based pay *increases* teachers' motivation (Muralidharan and Sundararaman 2011a). Further, Dal Bó, Finan, and Rossi (2013) find that increasing salaries for government jobs attracted higher ability workers with no adverse selection effects on their motivation levels. We assume therefore that λ_i and b are additively separable.

[13] About one-fifth of overall Tanzanian government expenditure is devoted to the education sector (WB 2014), over 40 percent of which is allocated to primary education (WB 2015c).



(a)

(b)

Figure 3.1: Effort and learning as a function of motivation, at different levels of inputs
Note: Figures 3.1a and 3.1b show how optimal effort (e^*) and optimal learning (ΔL^*) vary for different values of $b + \lambda_i$, across two levels of inputs ($I_1 > I_0$). In both figures $f(e, I) = \ln(e) + \ln(I) + e \cdot I$, $c_i(e) = e^2$, $I_0 = 1$, $I_1 = 1.2$, $\underline{\Delta L} = 0$, and $b + \lambda_i \in (0,1)$. κ_c is the threshold at which the constraint in Equation 3.1c is no longer binding for input level I_c, and therefore $e^*(I_c) = e^*_{mc}(I_c)$ to the right of κ_c.

increases in enrollment. As a result, inadequate school resources are a widely-posited reason for poor school quality. In 2012 only 3% of schools had sufficient infrastructure (clean water, adequate sanitation, and access to electricity) and in grades 1, 2, and 3 there was only one math textbook for every five children (World Bank 2012). Class sizes in primary schools average 74 students, with almost 50 students per teacher (World Bank 2012).

A second challenge for education quality is low teacher motivation and effort. A study conducted in 2010 found that nearly one in four teachers were absent from school on a given day, and over 50% of teachers who were present in school were absent from the classroom (World Bank 2012). The same study reported that on average, children receive only about 2 hours of instruction per day (less than half of the scheduled instructional time). Self-reported teacher motivation is also low: 47% of teachers surveyed in our data report that they would not choose teaching as a career if they could start over again.

3.3.2 Interventions and Implementation

Twaweza, an East African civil society organization that focuses on citizen agency and public service delivery, had played a leading role in independently measuring learning outcomes in Tanzania (Uwezo 2017). Having documented the challenge of low levels of learning, Twaweza conducted extensive consultations with local and global stakeholders and identified that the two main constraints to improving learning outcomes were likely to be inadequate school resources, and poor teacher motivation and effort.

Following this process, Twaweza formulated a program that aimed to alleviate these constraints and study their impact on learning outcomes. The program was called KiuFunza ("Thirst for learning" in Kiswahili) and was implemented in a representative sample of schools across Tanzania over two years (2013 and 2014). Twaweza also worked closely with both central and regional government officials to ensure smooth implementation of the program and evaluation. The interventions are described below:

Capitation grant (CG) program

Schools randomly selected for the capitation grants (CG) intervention received TZS 10,000 ($\sim$US$6.25 at the time of the study) per student from Twaweza. For context, GDP/capita in Tanzania in 2013 was $\sim$US$1,000 and the per-student grant value was $\sim$0.6% of GDP/capita, a sizeable amount. The value of this grant and the guidelines for their expenditure were similar to that of the government's own capitation grant program.[14] Typically, head teachers and members of the school board decided how to spend the grant funds, but schools had to maintain financial records of their transactions and were required to share revenue and expenditure information with the community (by displaying summary financial statements in a public area in the school).

Twaweza announced the grants early in the school year (March) during a series of meetings with school staff and community members, including parents and announced that the program would run for two years (2013 and 2014). Twaweza also distributed flyers and booklets that explained the program to parents, teachers, and community members. To minimize leakage, the funds were transferred directly into school bank accounts in two scheduled tranches: the first at the beginning of the second term (around April) and the second at the beginning of the third term (around August/September).[15]

Overall, Twaweza disbursed $\sim$ US$350,000/year to the 70 CG schools. The

[14] In practice, the average school received only around 60 percent of the stipulated grant value, and many received much less than that (World Bank 2012). Reasons included inadequate budgetary allocations, diversion of funds for other uses by local governments, and delays in disbursals. Thus, the marginal value of the CG program implemented by Twaweza was expected to be high.

[15] Twaweza also aimed to show that direct transfer of CG funds to school bank accounts would reduce leakage. This demonstration was successful as the Govt. of Tanzania decided to scale up this approach for its own capitation grant program as a result of the KiuFunza project.

size of the grants distributed to schools was $\sim$2-3 times the school-level spending per student (excluding teacher salaries and household spending), and the CG treatment represented a significant increase in the financial resources available to schools.[16]

Teacher performance pay (incentives) program

The teacher performance pay program provided cash bonuses to teachers based on the performance of their students on independent learning assessments conducted by Twaweza. Given Twaweza's emphasis on early grade learning, the program was limited to teachers in grades 1, 2, and 3 and focused on numeracy (mathematics) and literacy in English and Kiswahili. For each of these subjects, an eligible teacher earned a TZS 5,000 ($\sim$ US$3) bonus for each student who passed a simple externally administered, grade-appropriate assessment based on the national curriculum. Additionally, the head teacher was paid TZS 1,000 ($\sim$ US$0.6) for each subject test a student passed.[17]

The term used by Twaweza for the teacher-incentive program was "Cash on Delivery (CoD)" to reinforce the contrast between the approaches that underlay the two programs - with the CG program being one of unconditional school grants, and the teacher incentive program being one where payments were contingent on outcomes. The communication to schools and teachers emphasized that the aim of the CoD program was to motivate teachers and reward them for achieving better learning outcomes.

An advantage of the simple proficiency-based (or "threshold" based) incentive scheme used by Twaweza is its transparency and clarity. As pay-for-performance schemes are relatively novel in Tanzania, Twaweza prioritized having a bonus formula

[16] For example, if schools spent all of their grants on books, the funds would be sufficient to purchase about 4,000 textbooks per school ($\sim$ 4-5/student), given the average grant size of $\sim$ US$5,000 per school.

[17] Twaweza included head teachers in the incentive design to make them stakeholders in improving learning outcomes. It is also likely that any scaled up teacher incentive program would also feature bonuses for head-teachers along the lines implemented in the KiuFunza project.

that would be easy for teachers to understand. Bonuses based on passing basic tests of literacy and numeracy are also simpler to implement compared to more complex systems based on calculating student and teacher value added.

There are also important limitations to such a threshold-based design. It may encourage teachers to focus on students close to the passing threshold, neglecting students who are far below or far above the threshold (Neal and Schanzenbach 2010). In addition, such a design may be unfair to teachers who serve a large fraction of students from disadvantaged backgrounds, who may be further behind the passing standard. While Twaweza was aware of these limitations, they took a considered decision to keep the formula simple in the interest of transparency, simplicity of explaining to teachers, and ease of implementation.[18] Further, since the bonuses were based on achieving basic functional literacy and numeracy, they were not too concerned about students being so far behind the threshold that teachers would ignore them.

Twaweza announced the program to teachers in March 2013 and explained the details of the bonus calculations to the head teacher and teachers of the target grades (1-3) and subjects (math, Swahili, and English). Fliers with a description of the bonus structure and answers to frequently asked questions were handed out to teachers, and a booklet explaining the goals of the program were distributed to parents. A follow-up visit in July 2013 reinforced the details of the program and provided an opportunity for questions and feedback. Teachers understood the program: Over 90% of those participating in the program were able to correctly calculate the bonus level in a hypothetical scenario.

The high-stakes assessments that were used to determine the bonus payments were conducted at the end of the school year (with dates announced in advance), and consisted of three subject tests administered to all pupils in grades 1, 2 and 3.

[18] Even in the US, the early years of school accountability initiatives such as No Child Left Behind focused on measures based on *levels* of student learning rather than value-addition for similar reasons.

To ensure the integrity of the testing process, Twaweza created multiple versions of the high-stakes tests. These were allocated to students using random number tables. To prevent teachers from gaming the system by importing (or replacing) students, Twaweza only tested students enrolled at baseline (and took student photos at baseline to prevent identity fraud). Since each student enrolled at baseline had the potential to pass the exam, there would be no gains from preventing weaker students from taking the exam. All tests were conducted by and proctored by independent enumerators. Teacher bonuses were paid directly into their bank accounts or through mobile money transfers.

Combination (Combo) arm

Schools assigned to the combination arm received *both* the capitation grant and teacher incentive programs discussed above with identical implementation protocols.

3.4 Research Design

3.4.1 Sampling, and Randomization

We conducted the experiment in a nationally representative sample of 350 public schools across 10 districts in mainland Tanzania. We first randomly sampled 10 districts from mainland Tanzania, and then randomly sampled 35 schools within each of these districts to get a sample of 350 schools (Figure 3.2). Within each district, 7 schools were randomly assigned to receive capitation grants, 7 schools to receive teacher incentives, and 7 schools to receive both grants and incentives. The remaining 14 schools did not receive either program and served as our control group. Since the interventions were expensive, having a larger control group was a cost-effective way to increase power.

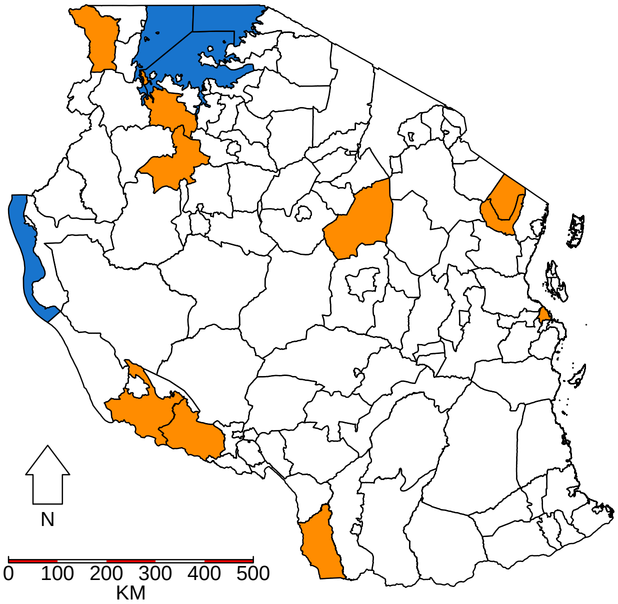


Figure 3.2: Districts in Tanzania from which schools were selected
Note: We drew a nationally representative sample of 350 schools from a random sample of 10 districts in Tanzania.

3.4.2 Data

Our analysis uses several pieces of data collected from schools, teachers, students, and households over the course of the study. Enumerators collected data on school facilities, input availability, management practices, and school income and expenditure.[19]. While most categories of school expenditure are difficult to map into specific grades, data on textbook expenditures was collected at the grade and subject level (since this is a substantial expenditure item, and can be mapped to the grade-level).

Enumerators also surveyed all teachers (about 1,500) who taught in focal grades

[19] Data on school expenditures were collected by reviewing receipts, accounting books, and other accounting records, following the expenditure tracking surveys developed and used by the World Bank (Reinikka and Smith 2004; Gurkan, Kaiser, and Voorbraak 2009)

(grades 1, 2, 3) and focal subjects (math, English and Swahili), and collected data on individual characteristics such as education and experience as well as effort measures such as teaching practices. They also conducted head teacher interviews.

For data on student learning outcomes, we sampled and tested 10 students from each focal grade (grades 1, 2 and 3) within each school, and follow them over the course of this study. Hence, we have a panel of 10,500 students. We refer to this as a low-stakes test as it is used purely for research purposes. From this set of students, we randomly sampled from each school five students from each of grades 2 and 3 to conduct household surveys. These 3,500 household surveys were designed to collect information on household characteristics, educational expenditures, and non-financial educational inputs at the household (such as helping with homework).[20]

We also use data from the high-stakes tests conducted by Twaweza that were used to determine teacher bonuses. These tests were taken by all students in grades 1, 2, and 3 in incentive and combo schools (where bonuses had to be paid). Twaweza did not conduct these tests in the CG schools, but they did conduct them in a sample of 40 control schools to enable the computation of treatment effects of the incentive programs on the high-stakes tests. However, we only have student level test-scores from the second year of the evaluation as the Twaweza teams only recorded aggregated pass rates (needed to calculate bonus payments) in the first year.

Figure 3.3 presents a timeline of the project, with implementation related activities listed below the line, and research related activities above the line. The baseline survey was conducted in February 2013, followed by an endline survey (with low-stakes testing) in October 2013. The high-stakes tests by Twaweza were conducted in November 2013. A similar calendar was followed in 2014. The trial registry record and the

[20] Because most of the survey questions focused on educational expenditures, including expenditures in the previous school year, we did not survey first-grade students in the first year of the study as they were typically not attending school in the previous year. In the second year of the study, the second graders (the initial cohort of first graders) were sampled for the household survey.

pre-analysis plan are available at: https://www.socialscienceregistry.org/trials/291.

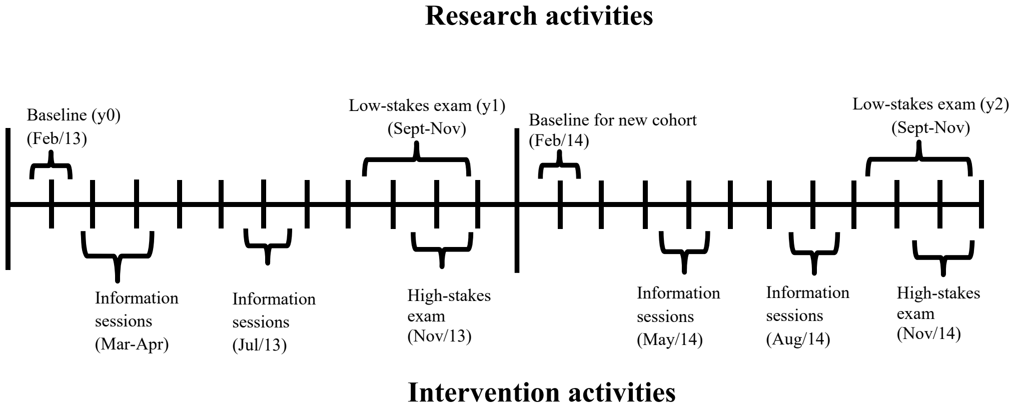

Research activities
Baseline (y0)
(Feb/13)
Low-stakes exam (y1)
(Sept-Nov)
Baseline for new cohort
(Feb/14)
Low-stakes exam (y2)
(Sept-Nov)
Information sessions (Mar-Apr)
Information sessions (Jul/13)
High-stakes exam (Nov/13)
Information sessions (May/14)
Information sessions (Aug/14)
High-stakes exam (Nov/14)
Intervention activities

Figure 3.3: Timeline

3.4.3 Validity

The randomization was successful and observable characteristics of students, households, schools, and teachers are balanced across our treatment arms; as are the normalized baseline test scores in each grade-subject (Table 3.1).

Table 3.1 also provides summary statistics on the (representative) study population. The student gender ratio is balanced, and the average student is 9 years old. The schools are mostly rural (85%), mean enrolment is ∼730, and class sizes are large – with an average of over 55 students per teacher (Panel C). Teachers in our sample were ∼2/3 female, ∼40 years old, had ∼15 years of experience, and ∼40% of them did not have a teaching certificate (Panel D).

Attrition on the low-stakes tests conducted by the research team is balanced across treatment arms and is low — we were able to track around 90% of students in both years (last two rows of Table 3.1: Panel A). On the high-stakes tests, we do find substantially higher student attendance in incentive and combo schools relative to the control group (Table C.11). This likely reflects the greater efforts put in by treatment schools to ensure high attendance on the day of testing (which was announced in advance) to maximize the value of the bonuses earned, which were paid on the basis of the *number* of students who passed the external test. We therefore present bounds of treatment effects when we use the high-stakes testing data, using the approach of Lee (2009).

3.4.4 Empirical Strategy

Our main estimating equation for school-level outcomes takes the form:

$$Y_{sdt} = \alpha_0 + \alpha_1 Incentives_s + \alpha_2 Grants_s + \alpha_3 Combo_s + \gamma_t + \gamma_d + X_s \alpha_4 + \varepsilon_{sdt}, \tag{3.2}$$

where Y_{sdt} is the outcome of interest in school s in district d at time t. $Incentives_s$ indicates whether school s received the teacher incentives program, $Grants_s$ is an indicator variable of whether school s received a capitation grant, and $Combo_s$ indicates whether schools s received both programs. γ_d, and γ_t are district and year fixed effects, and X_s is a set of school-level controls to increase precision. We use a similar specification to examine teacher-level outcomes such as teacher absence and pedagogical practices. All standard errors are clustered at the school-level.

We use a similar estimating equation to study effects on learning outcomes:

$$Z_{isdt} = \delta_0 + \delta_1 Incentives_s + \delta_2 Grant_s + \delta_3 Combo_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_g + X_i \delta_4 + X_s \delta_5 + \varepsilon_{isd}, \tag{3.3}$$

where Z_{isd} is the normalized test score of student i in school s in district d at time t (normalized with respect to the distribution of scores in the control group on the same test). We include normalized baseline test scores as controls to increase precision as well as stratification (district) fixed effects (γ_d). γ_g is a set of grade fixed effects, X_i is

Table 3.1: Summary statistics across treatment groups at baseline (February 2013)

	(1) Combo	(2) Grants	(3) Incentives	(4) Control	(5) p-value all equal
Panel A: Students (N=13,996)					
Male	0.50	0.49	0.50	0.50	0.99
	(0.01)	(0.01)	(0.01)	(0.01)	
Age	8.94	8.96	8.94	8.97	0.94
	(0.05)	(0.05)	(0.05)	(0.04)	
Swahili test score	0.05	-0.02	0.06	0.00	0.41
	(0.07)	(0.07)	(0.08)	(0.05)	
math test score	0.06	0.01	0.06	0.00	0.59
	(0.06)	(0.06)	(0.07)	(0.05)	
English test score	-0.02	-0.02	-0.00	0.00	0.91
	(0.04)	(0.05)	(0.05)	(0.04)	
Attrited in year 1	0.13	0.13	0.11	0.13	0.21
	(0.01)	(0.01)	(0.01)	(0.01)	
Attrited in year 2	0.10	0.10	0.10	0.10	0.95
	(0.01)	(0.01)	(0.01)	(0.01)	
Panel B: Households (N=7,001)					
HH size	6.23	6.26	6.41	6.26	0.19
	(0.12)	(0.12)	(0.13)	(0.08)	
Wealth index (PCA)	0.02	0.01	0.00	-0.02	0.99
	(0.16)	(0.16)	(0.17)	(0.12)	
Pre-treatment expenditure (TZS)	34,198.67	33,423.19	34,638.63	36,217.09	0.50
	(4,086.38)	(3,799.66)	(4,216.98)	(2,978.25)	
Panel C: Schools (N=350)					
Pupil-teacher ratio	54.78	58.78	55.51	60.20	0.50
	(2.63)	(3.09)	(2.53)	(3.75)	
Single shift	0.60	0.59	0.64	0.63	0.88
	(0.06)	(0.06)	(0.06)	(0.04)	
Infrastructure index (PCA)	-0.08	0.07	-0.12	0.06	0.50
	(0.13)	(0.14)	(0.16)	(0.08)	
Urban	0.16	0.13	0.17	0.15	0.85
	(0.04)	(0.04)	(0.05)	(0.03)	
Enrolled students	739.07	747.60	748.46	712.45	0.83
	(48.39)	(51.89)	(51.66)	(30.36)	

This table presents the mean and standard error of the mean (in parenthesis) for several characteristics of students in our sample (Panel A), households (Panel B), and schools (Panel C) across treatment groups. The student sample consists of all students tested by the research team. The sample consists of 30 students sampled in year one (10 from grade 1, 10 from grade 2, and 10 from grade 3) and 10 students sampled in year 2 (from the new grade 1 cohort). The attrition in year 1 is measured using only the original 30 students sampled per school. The attrition in year 2 is measured using the sample of 30 students that are enrolled in grades 1, 2 and 3 in that year. Column 4 shows the p-value from testing whether the mean is equal across all treatment groups ($H_0 :=$ mean is equal across groups). The household asset index is the first component of a Principal Component Analysis of the following assets: Mobile phone, watch/clock, refrigerator, motorbike, car, bicycle, television and radio. The school infrastructure index is the first component of a Principal Component Analysis of indicator variables for: Outer wall, staff room, playground, library, and kitchen. Standard errors are clustered at the school level for test of equality. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

a series of student characteristics (age, gender and grade), and X_s is a set of school and teacher characteristics.

We focus on test scores in math, English, and Kiswahili as our primary outcomes, and also study impacts on science (not a focal subject) to test if gains in focal subjects were achieved at the cost of other subjects (multi-tasking). To mitigate concerns about the potential for false positives due to multiple hypothesis testing, we also create a summary index of our focal subjects (math, English and Kiswahili), by taking the first component from a Principal Component Analysis (PCA) on the scores of the three subjects.

Since high-stakes tests were only conducted in incentive schools, combination schools, and a random set of 40 control schools, we cannot estimate the full comprehensive specification above. Furthermore, because the high-stakes exam is conducted only at the end of the year, we do not have baseline test scores or other student-level controls. Yet, most existing studies of teacher incentives have presented results based on high-stakes tests. We therefore present results using both high- and low-stakes tests.

For clarity of exposition and interpretation, we first present the impacts of the grant and incentive treatments individually (using only the intervention and the control group). We then present the impacts of all interventions estimated jointly, and test for complementarity: Specifically, we test $H_0 : \alpha_3 - \alpha_2 - \alpha_1 = 0$ and $H_0 : \delta_3 - \delta_2 - \delta_1 = 0$.

3.5 Results

3.5.1 Capitation Grant Program

How Were Grants Spent?

Table 3.2 presents descriptive statistics on how schools receiving the capitation grant program spent their funds. Textbooks and classroom teaching aids (like maps, charts, blackboards, chalk, etc.) were the largest category of spending, jointly accounting for $\sim$65% of average spending over the two years. Administrative costs, including wages of non-teaching staff (e.g., cooks, janitors, and security guards) accounted for ∼27% of spending. Smaller fractions (∼7%) were allocated to student support programs such as meal programs, and very little (∼1%) was spent on construction and repairs. There were essentially no funds allocated to teachers, as stipulated by the program rules.[21]

Schools also saved some of the grant funds (∼20% and ∼40% of grant value in the first and second year). Since schools knew that the CG program would end after two years, and government funding streams are uncertain (both in terms of timing and amount), we interpret this as "precautionary saving" and/or "consumption smoothing" behavior by schools (Sabarwal, Evans, and Marshak 2014). The possibility of outright theft was minimized by the careful review of expenditures conducted by the Twaweza team (and the prior announcements that such audits would take place).

Did Grants Offset other Spending?

Table 3.3 examines the extent to which receiving the CG program led to changes in other sources of income and spending. Column 1 summarizes the total extra spending from the capitation grant in grant schools. Schools that received Twaweza CG grants saw a reduction in school expenditure from other sources (Column 2). Aggregating across both years, schools receiving the CG program saw a reduction in other school spending of ∼2,400 TZS per child, which is around a third of the additional spending enabled by the CG program (Panel C - Columns 1 and 2).

Since average school spending per child in the control group was ∼5,200 TZS, spending the full grant value of 10,000 TZS would have tripled the school-level spending per child. After accounting for savings and offsetting reductions in school spending,

[21] Since teacher salaries are paid directly by the government, the capitation grant rules do not allow these funds to be used for teacher salaries. The Twaweza CG program had the same guidelines.

Table 3.2: How are schools spending the grants?

	(1)	(2)	(3)
		TZS per student	
	Year 1	Year 2	Average
Admin.	1,773.07	2,069.72	1,912.14
	(148.29)	(199.23)	(126.52)
Students	622.45	456.27	533.80
	(94.69)	(82.08)	(64.16)
Textbooks	3,858.69	1,315.83	2,585.75
	(257.56)	(172.39)	(154.05)
Teaching aids	1,761.43	2,132.32	1,947.61
	(126.53)	(190.00)	(118.45)
Teachers	0.00	3.36	1.68
	(0.00)	(3.36)	(1.68)
Construction	60.35	69.76	65.49
	(36.58)	(61.16)	(35.33)
Total Expenditure	8,075.99	6,047.26	7,046.46
	(318.42)	(352.57)	(238.98)
Unspent funds	1,924.01	3,952.74	2,953.54
	(318.42)	(352.57)	(238.98)
Total Value of CG	10,000.00	10,000.00	10,000.00
	(0.00)	(0.00)	(0.00)

Mean grant expenditure per student of school grants. *Admin:* Administrative cost (including staff wages), rent and utilities, and general maintenance and repairs. *Student:* Food, scholarships and materials (notebooks, pens, etc.). *Textbooks:* Textbooks. *Teaching aids:* Classroom furnishings, maps, charts, blackboards, chalk, practice exams, etc. *Teachers:* Salaries, bonuses and teacher training. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

there was still a significant net increase in total school spending per child of ∼4,700 TZS - almost double the expenditure relative to the control group (Panel C - Column 3).

Next, we examine changes in household spending. Column 4 shows the household offsets and Column 5 shows the total net per-child spending, accounting for both school and household spending. Consistent with the results documented by Das et al. (2013), we see an insignificant reduction in household spending by $\sim$1,000 TZS per child in the first year, and a larger significant reduction of $\sim$2,200 TZS per child in the second year (p=0.07). These spending cuts were from assorted fees, textbooks, and food (Table C.3).[22] Taken together, the reductions in school and household spending attenuated the impact of the Twaweza grant on per-student spending, but did not fully offset it. On net, CG schools saw a significant average increase in per-student spending of $\sim$3,100 TZS/year (Panel C, Column 5), a 60% increase over mean school-spending per student, enough to buy 3 textbooks per student per year.

[22] Households spend $\sim$5 times more per child than schools. Nearly 70% of this spending is on uniforms, tutoring, and food - which are typically not covered by the school (see Table C.3 for details).

Table 3.3: Effect of grants on school, household, and total expenditure

	(1)	(2)	(3)	(4)	(5)
			TZS per student		
	Grant exp.	Other school exp.	Total school [(1)+(2)]	Household exp.	Total exp. [(3)+(4)]
Panel A: Year 1					
Grants (α_1)	8,070.68***	-2,407.92***	5,662.75***	-1,014.96	4,647.79***
	(314.09)	(813.88)	(848.58)	(1,579.79)	(1,724.64)
N. of obs.	210	210	210	210	210
Mean control	0.00	5,959.67	5,959.67	28,821.01	34,780.68
Panel B: Year 2					
Grants (α_1)	6,033.08***	-2,317.74**	3,715.34***	-2,164.18*	1,585.75
	(336.95)	(1,096.16)	(1,122.60)	(1,201.53)	(1,548.42)
N. of obs.	209	209	209	210	209
Mean control	0.00	4,524.03	4,524.03	27,362.34	31,886.37
Panel C: Year 1 + Year 2					
Grants (α_1)	7,059.29***	-2,367.94***	4,688.04***	-1,589.57	3,133.33**
	(230.64)	(688.89)	(724.91)	(1,053.64)	(1,241.09)
N. of obs.	419	419	419	420	419
Mean control	0.00	5,241.85	5,241.85	28,091.68	33,333.53

Results from estimating Equation 3.2 for grant expenditure per child, other school expenditure per child, total school expenditure per child, and household reported expenditure in education. Column (1) shows grant expenditure as the dependent variable. Column (2) shows other school expenditure. Column (3) shows total school expenditure. Column (4) shows household data on expenditure in education. Column (5) shows total expenditure (total school expenditure + household expenditure). Panel C regressions are done including data for both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.4: Effect of grants on test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Year 1				Year 2	
	Math	Swahili	English	Combined (PCA)	Math	Swahili	English	Combined (PCA)
Grants (α_1)	-0.05	-0.01	-0.02	-0.03	0.01	-0.00	0.02	0.01
	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439

Results from estimating Equation 3.3 for different subjects at both follow-ups. Control variables include student characteristics (age, gender, grade and lag test scores) and school characteristics (PTR, Infrastructure PCA index, indicator for whether the school is in an urban or rural location, a PCA index of how close is the school to different facilities, and an indicator for whether the school is single shift or not). Clustered standard errors, by school, in parentheses. See Table C.4 for a version without school and household controls.

Did Grants Improve Learning?

Despite the significant and meaningful increases in per-pupil funding discussed above, schools receiving the extra capitation grants (CG) did not improve student test scores in math, English or Swahili in either year of our study, with point estimates close to zero (Table 3.4). Our estimates are precise enough to rule out (at a 95% level) effect sizes larger than 0.11σ on a composite measure of human capital in the second year and 0.03σ in the first year. Offsets are unlikely to be the main reason for our results, as we do not see any impacts of the grant on test scores even in the first year, when the net increase in spending per student in the CG schools was three times greater than in the second year (Table 3.3), Column 5). Overall, our results are consistent with and add to a large body of research that finds that merely increasing school resources rarely improves student learning outcomes in developing countries (including Glewwe, Kremer, and Moulin (2009) in Kenya, Blimpo et al. (2015) in Gambia, Das et al. (2013) in India, Pradhan et al. (2014) in Indonesia, and Sabarwal, Evans, and Marshak (2014) in Sierra Leone).

3.5.2 Teacher incentives

Since most studies of teacher incentives to date use high-stakes test scores for both evaluation and teacher bonus payments, we start by doing the same, and compare student performance on Twaweza's high-stakes end of year tests across treatment and control schools. Table 3.5 (Panel A) presents the impact of the teacher incentive program on pass rates on the Twaweza test (the metric on which the incentives were paid), and we see consistent evidence of a positive impact on pass rates.

At the end of two years, students in incentive schools were 37%, 17%, and 70% more likely to pass the Twaweza tests in math, Swahili, and English (all significant). These correspond to a 7.7, 7.3, and 2.1 percentage-point increase in the passing rate relative to control group means of 21%, 44%, and 3% in these subjects. Pass rates were also higher on all three subjects after the first year (though not significant in English).

Expressing the treatment effects in terms of normalized test scores, we find that students in incentive schools scored 0.17σ, 0.12σ, 0.12σ higher on math, Swahili, and English (all significant). Using a composite measure of human capital (based on the first principal component across the 3 subjects), we find that students in incentive schools scored 0.21σ higher (Table 3.5: Panel B). Since we do find differential attendance rates between treatment and control groups on the high-stakes tests (Table C.11), we estimate bounds on the treatment effects using the approach in Lee (2009) and find that the composite treatment effect is still positive and significant (Table C.12).

However, on the low-stakes tests administered by the research team, the effects are more modest and typically not significant (Table 3.5: Panel C). The composite treatment effect at the end of the first year is 0.06σ (p=0.09), and at the end of two years it is 0.03σ (not significant). The difference in estimated treatment effects of the teacher incentive program across the two types of tests can be explained by several reasons including cheating, timing, and testing day effort. We discuss each of these possibilities below.

As mentioned in Section 3.3.2, Twaweza employed strict security protocols for the high-stakes test, including having multiple versions of the test paper that were randomized across students in the same class, and having independent proctors present for every test. So, the possibility of cheating was minimized.

A second explanation is differences in the timing of the test. On average, the low-stakes tests were conducted about a month before the high-stakes test in both years (Figure 3.3). Since schools often conduct reviews and practice exams in this period, it is also possible that the superior performance on the high-stakes tests reflected this additional preparation (which would have had to be more intense in the incentive schools).

A final possibility is differences in student effort and testing conditions across the two sets of tests. During the low-stakes test, only a small number of students were tested (based on the random sample generated by the research team) while the rest of the school functioned as if it were a regular school day. On the other hand, Twaweza intervention testing was conducted in a more visible manner, where all other non-academic school activities were canceled to allow all grade 1, 2, and 3 students to take the test in as quiet an environment as possible. Further, in most cases the Twaweza test served as the end-of-school-year test for students in these schools. Finally, qualitative interviews suggest that teachers were more likely to have emphasized the importance of this test to students (since bonus payments depended on performance on these tests). Hence, students and teachers were likely to have been more motivated by the Twaweza exams.

Taken together, we conjecture that the main reason for the differences in treatment effects is the differences in student effort and testing conditions across the two sets of tests. Note also that the estimated difference in the two sets of tests of 0.10-0.15σ, is exactly in line with recent experimental estimates that quantify the role of day of test student effort on measured test scores Levitt et al. (2016).

The demonstration that test-taking effort is a salient component of measured test scores by Levitt et al. (2016) creates a conundrum for education researchers as to what the appropriate measure of human capital should be for assessing education interventions. On one hand, low-stakes tests may provide a better estimate of a true measure of human capital that does not depend on external stimuli to performance. On the other hand, effort is costly, and there is no reason to expect students to demonstrate their true potential under low-stakes testing, in which case, an 'incentivized' testing procedure may be a better measure of true human capital.

We therefore present both sets of results for completeness, and use the results from the high-stakes tests for cost-effectiveness calculations because existing research on teacher performance pay has typically used high-stakes tests for measuring program impact.

3.5.3 Combination of Capitation Grant and Teacher Incentives

As in the case of the incentive program, we start by presenting results from the Twaweza-implemented high stakes tests and then show impacts on the low-stakes tests. We also include comparisons with the other treatments (incentives treatment for high-stakes tests, and all treatments for low-stakes tests), and test for complementarities.

At the end of two years, students in "Combo" schools were 49%, 31%, and 116% more likely to pass the Twaweza-administered high-stakes test in math, Swahili, and English, with all three results being strongly significant at the 1% level (Table 3.6: Panel A). These correspond to a 10.3, 13.6, and 3.5 percentage-point increase relative to the control means of 21%, 44%, and 3%, and are very substantial increases. Pass rates were

Table 3.5: Effect of incentives on test scores: high- and low-stakes exams

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Swahili	English	Combined (PCA)	Math	Swahili	English	Combined (PCA)
Panel A: Pass rate, high-stakes								
Incentives	5.94^{***}	6.87^{*}	1.28		7.70^{***}	7.28^{**}	2.10^{**}	
	(1.95)	(3.61)	(1.00)		(1.84)	(3.35)	(0.81)	
N. of obs.	327	327	327	.	327	327	327	.
Control mean	20.06	36.76	3.73	.	20.99	43.97	3.01	.
Panel B: Z-scores, high-stakes								
Incentives (β_2)	.	.	.	.	0.17^{***}	0.12^{**}	0.12^{**}	0.21^{***}
					(0.05)	(0.05)	(0.05)	(0.07)
N. of obs.	.	.	.	.	19,256	19,256	19,256	19,256
Panel C: Z-scores, low-stakes								
Incentives (α_2)	0.06	0.05	0.06	0.06^{*}	0.07^{*}	0.01	0.00	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)
N. of obs.	5,496	5,496	5,496	5,496	5,653	5,653	5,653	5,653

Results from estimating Equation 3.3 for different subjects after two years. Control variables include student characteristics (age, gender, grade and lag test scores), school characteristics (PTR, Infrastructure PCA index, indicator for whether the school is in an urban or rural location, a PCA index of how close is the school to different facilities, and an indicator for whether the school is single shift or not), and household characteristics (household size, a PCA wealth index, and education expenditure prior to the intervention). Clustered standard errors, by school, in parentheses. See Table C.4 for a version without school and household controls.

also higher on all three subjects after the first year (though not significant in English). Point estimates of pass rates in the "Combo" treatment are always higher than those in the "Incentive" treatment, but not always significantly so.

Turning to normalized test scores, we find that students in "Combo" schools scored 0.25σ, 0.23σ, 0.22σ higher on math, Swahili, and English (p<0.01 in all cases), and scored 0.36σ higher on a composite measure of human capital (Table 3.6: Panel B). Due to the differential attendance rates between "Combo" and control groups on the high-stakes tests (Table C.11), we estimate bounds on the treatment effects using the approach in Lee (2009) and find that the composite treatment effect is still positive and significant (Table C.12).

These gains are also seen on the low-stakes tests. After two years, students in "Combo" schools scored 0.20σ, 0.21σ, 0.18σ higher on math, Swahili, and English ($p < 0.01$ in all cases), and scored 0.23σ higher on a composite measure of human capital (Table 3.6: Panel C).[23] They also do significantly better on all subjects at the end of one year. Thus, regardless of whether we use the high-stakes tests (conducted by the implementation team from Twaweza) or the low-stakes tests (conducted by the research team), we find that students in schools that received both programs had significantly higher test scores than those in control schools.

Using the low-stakes tests (that were conducted in *all* schools), we also find strong evidence of complementarities between the grant and incentive programs. Specifically, after two years, the impact under the "Combo" program is *significantly greater than the sum* of the impacts of the "Grant" and "Incentive" programs on their own, with this difference being significant for every subject, and also for the composite measure of

[23] The results in Panel C of Table 3.6 include students who were only treated for one year (e.g., third graders in the first year of the program and first graders during the second year), and students who were treated in both years (e.g., first and second graders during the first year of the program). Appendix Table C.5 shows the results focusing on the panel of students who were exposed the interventions in both years. We find very similar results among this group.

learning (last row of Table 3.6: Panel C). The point estimate for the complementarity is also positive for all subjects after one year, but not always significant.

Finally, while the high-stakes tests cannot be used to test for complementarities (because they were not conducted in the "Grant" schools), we see suggestive evidence of similar complementarities here as well using two different approaches. First, if we assume that the impact of the CG program on it's own is zero (based on the low-stakes test scores), then we can interpret the difference on the high-stakes tests between the "Combo" and the "Incentives" groups as an estimate of the complementarities and test that these are different from zero. We see that this is in fact the case (bottom row of Table 3.6: Panel B). A second approach is to not make this assumption and instead compare the difference between the "Combo" and the "Incentives" groups (which reflects the impact of the "Grant" *and* the "complementarities") on both the high-stakes and low-stakes tests and we see that we cannot reject that this difference is zero (last row of Panel D in Table 3.6). In other words, the results on the high-stakes tests are consistent with the complementarities estimated on the low-stakes tests.

3.5.4 Multi-tasking and Diversion

An important concern with teacher performance-pay schemes is the risk that such programs will encourage teachers to focus on incentivized subjects at the cost of other subjects or learning domains. On the other hand, if programs to incentivize math and reading are able to improve literacy and numeracy skills, they may promote student learning even in other non-incentivized subjects. We test for these possibilities by looking at impacts on science, a non-incentivized subject that was included in our battery of low-stakes student assessments. We find no evidence of negative impacts on learning in science (see Table 3.7). Rather, science scores actually increased in

Table 3.6: Effect of grants, incentives, and their interaction on test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Swahili	English	Combined (PCA)	Math	Swahili	English	Combined (PCA)
Panel A: Pass rate, high-stakes								
Incentives (γ_2)	5.94***	6.87*	1.28		7.70***	7.28**	2.10**	
	(1.95)	(3.61)	(1.00)		(1.84)	(3.35)	(0.81)	
Combo (γ_3)	8.99***	11.70***	1.58		10.30***	13.64***	3.49***	
	(2.05)	(3.59)	(0.99)		(1.97)	(3.27)	(1.06)	
N. of obs.	327	327	327	.	327	327	327	.
Control mean	20.06	36.76	3.73	.	20.99	43.97	3.01	.
$\gamma_3 - \gamma_2$	3	4.8*	.3	.	2.6	6.4**	1.4	.
p-value ($\gamma_3 - \gamma_2 = 0$)	.1	.071	.69	.	.17	.018	.17	.
Panel B: Z-scores, high-stakes								
Incentives (β_2)	.	.	.	.	0.17***	0.12**	0.12**	0.21***
					(0.05)	(0.05)	(0.05)	(0.07)
Combo (β_3)	.	.	.	.	0.25***	0.23***	0.22***	0.36***
					(0.05)	(0.06)	(0.06)	(0.08)
N. of obs.	.	.	.	.	46,886	46,882	46,882	46,882
$\beta_4 := \beta_3 - \beta_2$	.	.	.	.	0.081**	0.11**	0.099*	0.15**
p-value ($\beta_4 = 0$)	.	.	.	.	0.046	0.012	0.060	0.015
Panel C: Z-scores, low-stakes								
Grants (α_1)	-0.05	-0.01	-0.02	-0.03	0.01	-0.00	0.02	0.01
	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
Incentives (α_2)	0.06	0.05	0.06	0.06*	0.07*	0.01	0.00	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)
Combo (α_3)	0.10**	0.10***	0.10**	0.12***	0.20***	0.21***	0.18***	0.23***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	0.100*	0.063	0.068	0.089	0.12*	0.20***	0.16*	0.18***
p-value ($\alpha_4 = 0$)	0.093	0.27	0.28	0.11	0.084	0.0044	0.050	0.0070
Panel D: Difference								
$\beta_2 - \alpha_2$	.	.	.	.	0.086	0.11	0.12	0.17
p-value($\beta_2 - \alpha_2 = 0$)	.	.	.	.	0.14	0.048	0.071	0.016
$\beta_3 - \alpha_3$	.	.	.	.	0.035	0.014	0.030	0.12
p-value($\beta_3 - \alpha_3 = 0$)	.	.	.	.	0.53	0.81	0.63	0.082
$\beta_4 - \alpha_4$	.	.	.	.	-0.057	-0.088	-0.11	-0.056
p-value($\beta_4 - \alpha_4 = 0$)	.	.	.	.	0.55	0.16	0.44	0.61

Results from estimating Equation 3.3 for different subjects at both follow-ups. Control variables include student characteristics (age, gender, grade and lag test scores), school characteristics (PTR, Infrastructure PCA index, indicator for whether the school is in an urban or rural location, a PCA index of how close is the school to different facilities, and an indicator for whether the school is single shift or not), and household characteristics (household size, a PCA wealth index, and education expenditure prior to the intervention). Clustered standard errors, by school, in parentheses. See Table C.4 for a version without school and household controls.

combination schools by just under 0.1σ, while science scores in grant and incentive schools did not significantly change. Further, we find evidence of complementarities between grants and incentives in science learning in the second year. This mirrors the pattern of estimated complementarities found in our main specifications. As the program improved literacy and numeracy in combination schools, this suggests that improvements in those foundational skills facilitated science learning in combination schools.

Because school grants could be spent across all grades, the lack of positive treatment effects in our focal study grades (grades 1, 2, and 3) could reflect schools' propensity to invest more in upper grades. As the grade 7 exit exam determines a school's reputational quality, schools may be better off investing in later grades rather than earlier ones. We examine the impact of our interventions on performance on the Primary School Leaving Examination (PSLE) taken by students in Grade 7 (a non-incentivized grade) in columns 3 to 6 of Table 3.7. We show the proportion who pass the exam and are thus able to transition to secondary school, as well as the average school test score in both 2013 and 2014 (our two program years). We do not see any evidence that any of our interventions significantly affected performance on the national exit exam, both in terms of average scores or pass rates. We also do not see any evidence of complementarities between interventions in the grade 7 outcomes. An important caveat is that there is an increase in the number of test takers in the combination group in both years. To account for potential changes in student composition, we construct Lee (2009) bounds in Appendix Table C.8, where we assume the marginal students were the lowest performers. As our results from this exercise are almost identical to the full sample, we argue that there is no impact of the interventions on learning in higher grades.

Table 3.7: Spillovers into other grades and subjects

	(1)	(2)	(3)	(4)	(5)	(6)
	Science		Grade 7 PSLE 2013		Grade 7 PSLE 2014	
	Year 1	Year 2	Pass	Score	Pass	Score
Grants (α_1)	0.02	-0.04	-0.03	-0.03	-0.02	-0.05
	(0.05)	(0.06)	(0.03)	(0.05)	(0.03)	(0.05)
Incentives (α_2)	0.01	-0.01	-0.02	-0.02	-0.00	-0.02
	(0.05)	(0.05)	(0.03)	(0.04)	(0.03)	(0.05)
Combo (α_3)	0.09	0.09*	0.02	0.05	0.01	0.04
	(0.05)	(0.05)	(0.03)	(0.05)	(0.03)	(0.05)
N. of obs.	9,142	9,439	26,836	26,836	25,162	25,162
Mean control group			0.52	2.60	0.57	2.70
$\alpha_4 = \alpha_3 - \alpha_2 - \alpha_1$	0.058	0.13*	0.066	0.10	0.039	0.11
p-value ($H_0 : \alpha_4 = 0$)	0.48	0.096	0.11	0.11	0.35	0.12

Columns (1) and (2) estimate Equation 3.3 for science in focal grades (Grd 1 - Grd 3) using data for both follow-ups, and therefore coefficients represent the average treatment effect across both years. Columns (3)-(6) use data from the national exit examination as dependent variables: pass rates and average test scores. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.5.5 Potential mechanisms

Differential spending We explore the possibility that combination schools allocated their funds differently than grant schools. We first examine the impact of our interventions on the same broad spending categories analyzed in Table 3.3. We report the results separately by year (Panel A and B), but also report the averages across both years (Panel C) in Table 3.8. In the first year of the program, expenditure patterns between grant schools and combination schools are quite similar. In contrast, in the second year of the program, parents in grant schools cut back their spending, whereas there are no parental offsets in combination schools (p-value 0.11). This is consistent with increases in a number of dimensions of (unobservable) teacher (and head teacher) effort in combination schools. In particular, teachers (and head teachers) could lobby and encourage parents to continue to financially support their children. This differential effort response between combination and grant schools is consistent with the theoretical predictions of our model.

As our previous expenditure results analyzed broad funding categories, the results could mask differential investment across grades. We focus on textbook expenditures in Table 3.9, as textbooks are assignable to grades and account for a significant fraction of school spending. In particular, we examine whether combination schools invest relatively more in incentivized grades compared to schools that only receive grants. We present the results on expenditure pooling in Grades 4 to 7 in Column 1 and Grades 1 to 3 (the incentivized or focal grades) in Column 2. We present the differences between columns in Column 3. Textbook expenditures increased across all grade groups in both grant and combination schools. Grant schools spent more on textbooks in higher grades relative to lower grades, while combination schools spent approximately the same amount across all grades (Column 3). When we formally test the differences in relative spending across the treatments, we find that combination schools spent 543 TZS more per student on textbooks in incentivized grades (relative to non-incentivized grades) compared to schools that only received the grants (p-value of the difference is 0.05). Moreover, when we test for complementarities in expenditure, combination schools invested relatively more in incentivized grades (p-value is 0.1). This differential investment suggests that combination schools spent more of their capitation grant resources to support teachers who are eligible for the bonus program, perhaps in response to lobbying efforts among those teachers. As we do not see differences in self-reported (by the head teacher) lobbying efforts by teachers, our results could also be driven by head teachers in combination schools internalizing their potential individual payoff from providing additional resources to incentivized teachers.

Table 3.8: Effect of grants, incentives, and their interaction on expenditure

	(1)	(2)	(3)	(4)	(5)
	Grant exp.	Other school exp.	Total school	Household exp.	Total exp.
			[(1)+(2)]		[(3)+(4)]
Panel A: Year 1					
Grants (α_1)	8,070.68***	-2,407.92***	5,662.75***	-1,014.96	4,647.79***
	(314.09)	(813.88)	(848.58)	(1,579.79)	(1,724.64)
Incentives (α_2)	-6.77	-10.05	-16.82	-977.78	-994.60
	(63.15)	(642.21)	(638.81)	(1,294.84)	(1,439.10)
Combo (α_3)	8,329.38***	-1,412.22	6,917.16***	-1,382.23	5,534.93***
	(241.13)	(932.79)	(919.07)	(1,153.27)	(1,564.93)
N. of obs.	350	350	350	350	350
Mean control	0.00	5,959.67	5,959.67	28,821.01	34,780.68
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	265.47	1,005.76	1,271.23	610.51	1,881.74
p-value ($H_0 : \alpha_4 = 0$)	0.50	0.44	0.33	0.77	0.45
$\alpha_3 - \alpha_1$	258.70	995.70	1,254.41	-367.27	887.14
p-value ($H_0 : \alpha_3 - \alpha_1 = 0$)	0.51	0.39	0.28	0.83	0.67
Panel B: Year 2					
Grants (α_1)	6,033.08***	-2,317.74**	3,715.34***	-2,164.18*	1,585.75
	(336.95)	(1,096.16)	(1,122.60)	(1,201.53)	(1,548.42)
Incentives (α_2)	22.70	-1,166.46	-1,143.75	235.40	-907.97
	(98.63)	(818.24)	(830.33)	(1,214.01)	(1,422.09)
Combo (α_3)	5,620.07***	-1,896.28**	3,723.79***	-75.59	3,646.85**
	(320.69)	(928.05)	(989.27)	(1,151.27)	(1,520.20)
N. of obs.	349	349	349	350	349
Mean control	0.00	4,524.03	4,524.03	27,362.34	31,886.37
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-435.71	1,587.92	1,152.20	1,853.19	2,969.07
p-value ($H_0 : \alpha_4 = 0$)	0.35	0.15	0.33	0.30	0.16
$\alpha_3 - \alpha_1$	-413.01	421.46	8.45	2,088.59	2,061.10
p-value ($H_0 : \alpha_3 - \alpha_1 = 0$)	0.37	0.56	0.99	0.11	0.18

Results from Estimating Equation 3.2 for grant expenditure per child, other school expenditure per child, total school expenditure per child, and household reported expenditure on education. Column (1) shows grant expenditure as the dependent variable. Column (2) shows other school expenditure. Column (3) shows total school expenditure. Column (4) shows household data on expenditure in education. Column (5) shows total expenditure (total school expenditure + household expenditure). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.9: Effect of grants, incentives, and their interaction on textbook expenditure by grade

	(1)	(2)	(3)
	Grades 4-7	Grades 1-3	Difference
			[(2)-(1)]
Grants (α_1)	1,743.61***	1,259.14***	-484.47***
	(224.77)	(183.70)	(159.30)
Incentives (α_2)	-131.56	-50.42	81.13
	(105.69)	(71.51)	(92.99)
Combo (α_3)	1,504.34***	1,563.35***	59.01
	(194.64)	(202.35)	(228.66)
N. of obs.	2,780	2,100	4,880
Mean control	846.26	498.74	-347.52
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-107.71	354.64	462.35
p-value ($H_0 : \alpha_4 = 0$)	0.72	0.19	0.10
$\alpha_3 - \alpha_1$	-239.27	304.21	543.48
p-value ($H_0 : \alpha_3 - \alpha_1$=0)	0.40	0.25	0.045

Results from estimating Equation 3.2 on textbook expenditure for grades 4-7 (Column 1), grades 1-3 (Column 2), and the difference between them (Column 3). Expenditure for children in grades 4-7 are show in Column 1, expenditure for children enrolled in grades 1-3 are shown in Column 2, and the difference in Column 3. The sample of children only includes children living in the same household and attending the same schools as the sampled student. The regression includes data for both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Teacher behavioral responses We explore teacher behavioral responses to the interventions in Table 3.10. We primarily focus on broad measures of teacher effort and pedagogical techniques using a combination of self-reported and observed data. Although teacher absence rates are generally high in Tanzania, our interventions do not significantly affect absence. We conducted our surprise visits in the middle of the school year, and it is possible that teachers concentrated their efforts closer to exam time at the end of the year. Alternatively, teachers may have responded on the intensive margin, increasing their effort when they were in the classroom. Based on self-reported data from teachers, teachers in incentive schools increased the number of tests in their classrooms by 1.25, roughly a 10 percent increase (Column 2). Additionally, teachers in combination schools were 5 percentage points (or almost 50 percent) more likely to provide extra tutoring to students (Column 3), and were also 6 percentage points (or 7 percent) more likely to provide remedial support to students (Column 4). Taken together, the results in Columns 2 to 4 suggest that teacher behavior did respond to the incentives. They also suggest that teachers in combination schools were able to better support their students, especially those falling behind, through remedial support, where we find evidence of complementarities in teacher behavior. We also examine teacher self reports on teaching inputs in Column 5. We create a binary variable which indicates whether instructional inputs are above average. As expected, teachers in both grant and combination schools reported increases in classroom inputs, providing some reassurance about the quality of teacher self-reported data.

3.5.6 Heterogeneity

To gain additional insights into potential mechanisms, we explore heterogeneous treatment effects by various dimensions of student, teacher, and school characteristics in Table 3.11. These results are estimated using Equation 3.2, and adding interactions of the treatment with different variables. For brevity, we simply report the interaction coefficients and focus on the index of test scores. The first three columns focus on student-level heterogeneity. Boys generally benefit less from the incentive and combination treatment than girls (see Column 1). Teacher incentives can help reduce

Table 3.10: Effect of grants, incentives, and their interaction on teacher behavior

	(1)	(2)	(3)	(4)	(5)
		Self-reported			
	Attendance	Tests	Tutoring	Remedial	Inputs
Grants (α_1)	0.04	-0.23	0.02	-0.03	0.08***
	(0.03)	(0.79)	(0.03)	(0.03)	(0.02)
Incentives (α_2)	-0.01	1.43**	0.04	-0.05	-0.00
	(0.03)	(0.71)	(0.03)	(0.03)	(0.03)
Combo (α_3)	0.02	-0.07	0.06**	0.06**	0.08***
	(0.03)	(0.63)	(0.02)	(0.03)	(0.02)
N. of obs.	1,865	1,853	1,865	1,865	1,865
Mean of dep. var.	0.78	9.50	0.093	0.84	0.93
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-0.0094	-1.27	0.0037	0.14	0.0017
p-value ($H_0 : \alpha_4 = 0$)	0.84	0.23	0.93	0.0027***	0.96

Results from estimating treatment effects on teacher behavior. Column (1) shows teacher attendance independently measured by enumerators during a surprise visit in the middle of the school year. Column (2) shows the number of tests per period as the dependent variable. Column (3) shows a dummy variable that indicates whether the teacher provided any extra tutoring to students as the dependent variable. Column (4) shows a dummy variable that indicates whether the teacher provided remedial teaching to students as the dependent variable. Column (5) shows a dummy variable equal to one if the teacher indicates teaching inputs are "above average" as the dependent variable. All regressions are done including data for both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

gender gaps, although it is not clear why.[24] Older children also benefit less from the combination. Finally, better-prepared students, as measured by lagged test scores, benefited less from the combination and grant treatments (Column 3). This could reflect inequality aversion among teachers or school administrators.

Columns 4, 5, and 6 explore heterogeneity by teacher characteristics. To proxy for teacher motivation, we use data on teacher self-reports on their likelihood to pursue teaching careers if they had an opportunity to redo their career choice. We also use self-reported earnings and use the digit recall cognitive test to measure teachers' working memory. Overall, we do not find any significant interactions with these measures. As there is limited variation in teacher education in Tanzania, with almost all teachers holding a teaching certificate, we do not explore that dimension of teacher heterogeneity.

Columns 7, 8, and 9 explore heterogeneity by school characteristics. Schools with better baseline facilities (an index measure of facilities) improve more when they are provided teacher incentives (Column 7). This is consistent with our experimental findings on the complementarities of resources and incentives. We do not see any differential treatment effects by pupil teacher ratio (Column 8). As a growing number of studies have highlighted the importance of school management in the education production function (see Bloom et al. (2015) and Crawfurd (2017)), we explore the heterogeneity in treatment effect by an index of self-reported managerial ability of the head teacher. Combination schools where school heads had higher managerial ability saw greater increases in test scores (Column 9). Managerial ability may be particularly important in combination schools: Head teachers had to oversee two programs at their schools and sort the demand for scarce resources by all teachers in the school.

[24] Anderson (2008) reviews several early childhood interventions with larger effects for girls.

Table 3.11: Heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Student			Teacher			School		
	Male	Age	Lagged score	Motivation	Salary	Memory	Facilities	PTR	Management
Grants*Covariate	0.02	0.00	-0.06**	0.04	0.00	0.02	0.08	0.00	0.07
	(0.04)	(0.01)	(0.03)	(0.18)	(0.00)	(0.08)	(0.07)	(0.00)	(0.08)
Incentives*Covariate	-0.07*	-0.00	-0.01	0.03	-0.00	0.06	0.14**	-0.00	-0.07
	(0.04)	(0.01)	(0.02)	(0.12)	(0.00)	(0.07)	(0.07)	(0.00)	(0.06)
Combo*Covariate	-0.10**	-0.03*	-0.06**	-0.12	0.00	0.00	0.09	-0.00	0.15**
	(0.04)	(0.01)	(0.03)	(0.13)	(0.00)	(0.07)	(0.07)	(0.00)	(0.06)
N. of obs.	18,581	18,581	18,581	18,581	18,581	18,581	18,581	18,581	18,206

The independent variable is the standardized test score. Each regression has a different covariate interacted with the treatment dummies. The column title indicates the covariate interacted. Panel A has the following covariates at the student level: The standardized test score at baseline; Gender, a dummy equal to one if the student is male; and the age in years. Panel B has the following covariates as the school level: a dummy for whether the PCA index of facilities is above the median; the pupil-teacher ratio; and a dummy equal to one if the PCA index for managerial ability of the principal is above the median. Panel C has the following covariates at the teacher level: a dummy measuring motivation that is equal to one if the teacher reported that he/she would choose teaching as a career if he/she could start over; the annual salary; and a dummy equal to one if the PCA index of a memory test is above the median. The teacher covariates are averaged across teachers in both years. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.6 Conclusion

In this paper we report findings from a large education RCT aimed at improving learning in early grades. Consistent with the existing literature, merely increasing school resources does little to improve learning outcomes. A simple incentive program yields limited positive impacts on learning. However, test scores in schools that received both programs were significantly higher. Moreover, we find strong evidence of complementarities between inputs and incentives. The increases in learning (in the combo schools) were concentrated among students near the passing threshold. This is further evidence of the importance of incentive design in promoting student learning. In the presence of large complementarities, programs that are rolled out in isolation may yield limited returns as they fail to address multiple binding constraints, which may be especially important for developing country settings. The failure of many programs or interventions in education, such as textbook programs, to alleviate multiple constraints may help explain why many rigorous evaluations of education interventions often find limited impact. There may be large gains if policymakers and researchers developed and evaluated multifaceted interventions that address several constraints, scaling up the combination of programs that exhibit large complementarities.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Mbiti, Isaac; Muralidharan, Karthik; Romero, Mauricio; Schipper,Youdi; Manda, Constantine; Rajani, Rakesh. “Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania”. The dissertation author was a primary investigator and author of this material.

Appendix A

Outsourcing Service Delivery in a Fragile State: Experimental Evidence from Liberia

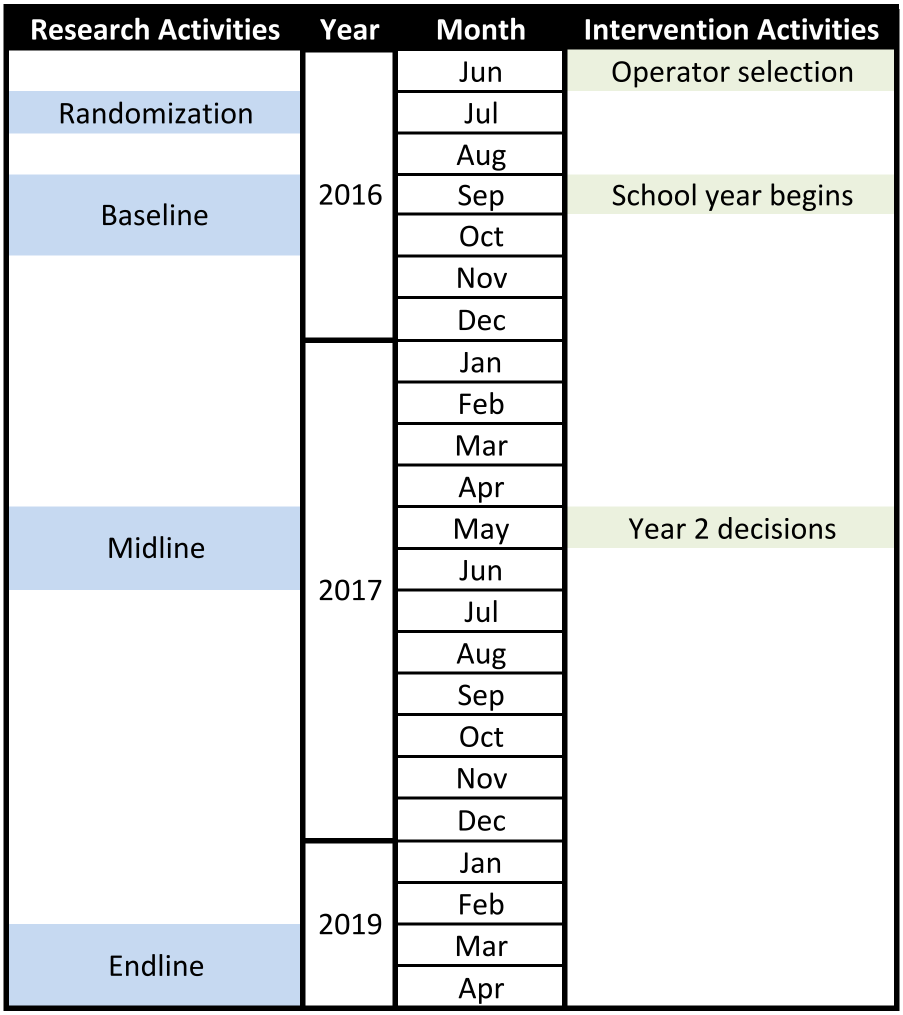
A.1 Additional tables and figures

Table A.1: External validity: Differences in characteristics of schools in the RCT (treatment and control) and other public schools (based on EMIS data)

	(1) RCT (Treatment and control)	(2) Other public schools	(3) Difference
Students: ECE	142.68	112.71	29.97***
	(73.68)	(66.46)	(5.77)
Students: Primary	151.55	132.38	19.16*
	(130.78)	(143.57)	(10.18)
Students	291.91	236.24	55.67***
	(154.45)	(170.34)	(12.15)
Classrooms per 100 students	1.17	0.80	0.37***
	(1.63)	(1.80)	(0.13)
Teachers per 100 students	3.04	3.62	-0.58**
	(1.40)	(12.79)	(0.28)
Textbooks per 100 students	99.21	102.33	-3.12
	(96.34)	(168.91)	(7.88)
Chairs per 100 students	20.71	14.13	6.58***
	(28.32)	(51.09)	(2.38)
Food from Gov or NGO	0.36	0.30	0.06
	(0.48)	(0.46)	(0.04)
Solid building	0.36	0.28	0.08*
	(0.48)	(0.45)	(0.04)
Water pump	0.62	0.45	0.17***
	(0.49)	(0.50)	(0.04)
Latrine/toilet	0.85	0.71	0.14***
	(0.33)	(0.45)	(0.03)
Observations	185	2,420	2,605

This table presents the mean and standard error of the mean (in parentheses) for schools in the RCT (Column 1) and other public schools (Column 2), as well as the difference in means across both groups (Column 3). The sample of RCT schools is the original treatment and control allocation. ECE = Early childhood education. MOE= Ministry of Education. Authors' calculations based on 2015/2016 EMIS data.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



Research Activities	Year	Month	Intervention Activities
	2016	Jun	Operator selection
Randomization		Jul	
		Aug	
Baseline		Sep	School year begins
		Oct	
		Nov	
		Dec	
	2017	Jan	
		Feb	
		Mar	
		Apr	
Midline		May	Year 2 decisions
		Jun	
		Jul	
		Aug	
		Sep	
		Oct	
		Nov	
		Dec	
	2019	Jan	
		Feb	
Endline		Mar	
		Apr	

Figure A.1: Timeline

Note: Bridge signed its MOU with the Government of Liberia in March 2016, and thus started preparing for the program earlier than other providers.

Table A.2: Number of schools by provider

	(1)	(2)	(3)	(4)	(5)	(6)
	Assigned	Non-compliant	Replacement	Outside sample	Managed	% compliant
BRAC	20	0	0	0	20	100%
Bridge	23	0	0	2	25	100%
YMCA	4	0	0	0	4	100%
MtM	6	2	2	0	6	67%
Omega	19	2	0	0	17	89%
Rising	5	1	0	1	5	80%
Stella	4	4	0	0	0	0%
St.Child	12	2	2	0	12	83%

Note: The table shows the number of schools originally assigned to treatment (Column 1) and the schools that either did not meet Ministry of Education criteria or were rejected by providers (Column 2). The Ministry of Education provided replacement schools for those that did not meet the criteria, presenting each provider with a new list of paired schools and informing them, as before, that they would operate one of each pair (but not which one). Replacement schools are shown in Column 3. Column 4 contains non-randomly assigned schools given to some providers. Column 5 shows the final number of schools managed by each provider and is equal to the sum of the first four columns. Finally, the last column shows the percentage of schools actually managed by the provider that are in our main sample.

Table A.3: Balance table: Differences in characteristics of treatment and control schools, pre-treatment year (2015/2016, EMIS data)

	(1) Treatment	(2) Control	(3) Difference	(4) Difference (F.E)
Students: ECE	148.51	136.72	11.79	11.03
	(76.83)	(70.24)	(10.91)	(9.74)
Students: Primary	159.05	143.96	15.10	15.68
	(163.34)	(86.57)	(19.19)	(16.12)
Students	305.97	277.71	28.26	27.56
	(178.49)	(124.98)	(22.64)	(19.46)
Classrooms per 100 students	1.21	1.13	0.09	0.08
	(1.62)	(1.65)	(0.24)	(0.23)
Teachers per 100 students	3.08	2.99	0.09	0.09
	(1.49)	(1.30)	(0.21)	(0.18)
Textbooks per 100 students	102.69	95.69	7.00	7.45
	(97.66)	(95.40)	(14.19)	(13.74)
Chairs per 100 students	18.74	22.70	-3.96	-4.12
	(23.06)	(32.81)	(4.17)	(3.82)
Food from Gov or NGO	0.36	0.36	-0.01	-0.01
	(0.48)	(0.48)	(0.08)	(0.05)
Solid building	0.39	0.33	0.06	0.06
	(0.49)	(0.47)	(0.07)	(0.06)
Water pump	0.56	0.67	-0.11	-0.12*
	(0.50)	(0.47)	(0.07)	(0.06)
Latrine/toilet	0.85	0.86	-0.01	-0.01
	(0.35)	(0.32)	(0.05)	(0.05)
Observations	93	92	185	185

This table presents the mean and standard error of the mean (in parenthesis) for the control (Column 1) and treatment (Column 2), as well as the difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 4. The sample is the final treatment and control allocation. Authors' calculations based on EMIS data.

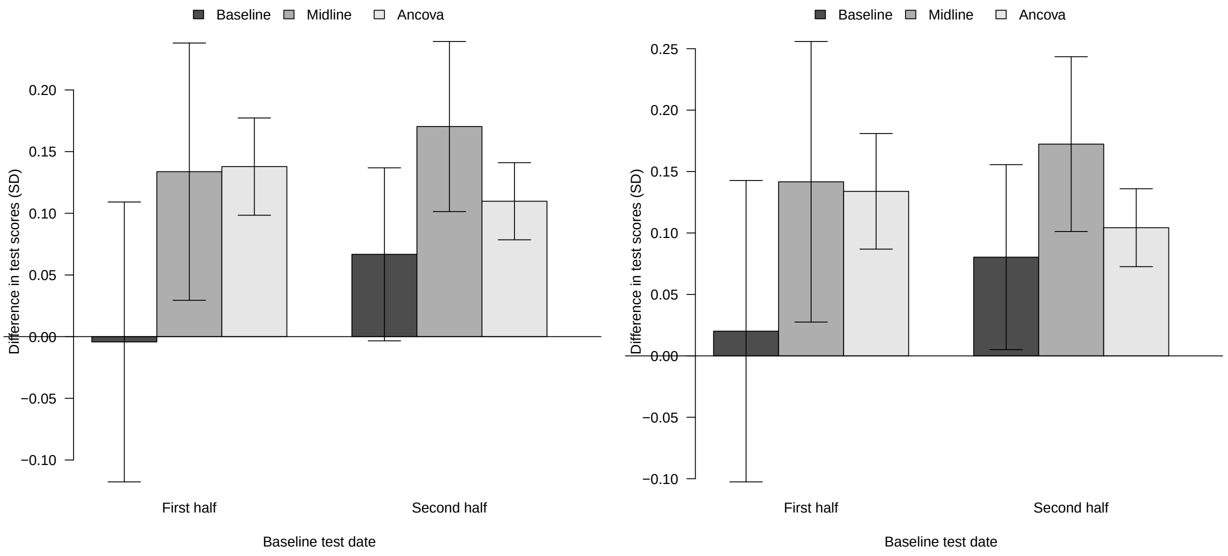
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: ITT treatment effects on learning

	First wave (1-2 months after treatment)		Second wave (9-10 months after treatment)			
	Difference	Difference (F.E.)	Difference	Difference (F.E.)	Difference (F.E. + Controls)	Difference (ANCOVA)
	(1)	(2)	(3)	(4)	(5)	(6)
English	0.05	0.09^{*}	0.17^{**}	0.17^{***}	0.18^{***}	0.13^{***}
	(0.08)	(0.05)	(0.08)	(0.04)	(0.03)	(0.02)
Math	0.08	0.08^{*}	0.17^{***}	0.19^{***}	0.18^{***}	0.14^{***}
	(0.07)	(0.04)	(0.07)	(0.04)	(0.03)	(0.02)
Abstract	0.04	0.05	0.05	0.05	0.05	0.03
	(0.06)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)
Composite	0.07	0.08^{*}	0.17^{**}	0.19^{***}	0.19^{***}	0.14^{***}
	(0.07)	(0.05)	(0.07)	(0.04)	(0.03)	(0.02)
New			0.17^{**}	0.20^{***}	0.19^{***}	0.16^{***}
			(0.07)	(0.04)	(0.04)	(0.03)
Conceptual			0.12^{**}	0.13^{***}	0.12^{***}	0.10^{***}
			(0.05)	(0.04)	(0.04)	(0.04)
Observations	3,496	3,496	3,492	3,492	3,492	3,492

Columns 1-2 use baseline data and show the difference between treatment and control (Column 1), and the difference taking into account the randomization design — i.e., including "pair" fixed effects — (Column 2). Columns 3-6 use May/June 2017 data and show the difference between treatment and control (Column 3) in test scores, the difference taking into account the randomization design — i.e., including "pair" fixed effects — (Column 4), the difference taking into account other student and school controls (Column 5), and the difference using an ANCOVA style specification which controls for baseline test scores (Column 6).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) Math

(b) English

Figure A.2: Treatment effects by date tested during the first round of data collection
Note: The panel on the left shows results for math test scores, while the panel on the right shows English test scores.

Table A.5: Heterogeneity by student characteristics

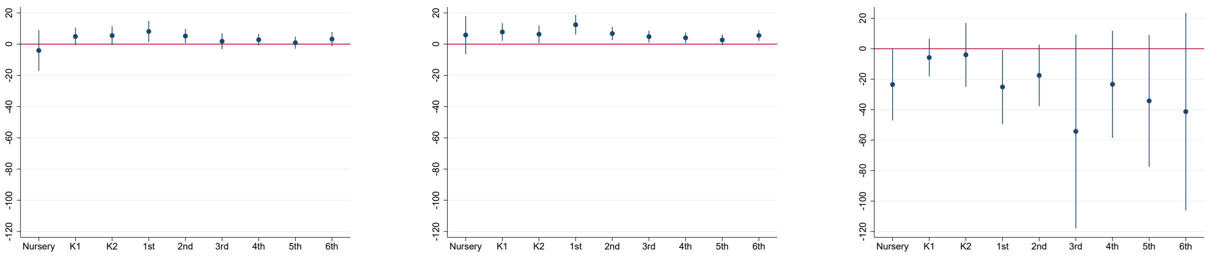
	Male	Top wealth quartile	Bottom wealth quartile	Grade
	(1)	(2)	(3)	(4)
Treatment	0.20***	0.18***	0.17***	0.16
	(0.047)	(0.035)	(0.035)	(0.10)
Treatment × covariate	-0.021	0.030	0.061	0.0050
	(0.068)	(0.066)	(0.050)	(0.020)
No. of obs.	3,492	3,492	3,492	3,492

Each column shows the interaction of a different covariate with treatment. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: ITT and ToT effect

	Difference (Controls)			ANCOVA		
	Math (1)	English (2)	Abstract (3)	Math (4)	English (5)	Abstract (6)
Panel A: ITT						
Treatment	0.18***	0.18***	0.046	0.14***	0.13***	0.031
	(0.034)	(0.030)	(0.037)	(0.023)	(0.021)	(0.036)
No. of obs.	3,492	3,492	3,492	3,492	3,492	3,492
Panel B: ToT						
Treatment	0.23***	0.22***	0.058	0.18***	0.17***	0.040
	(0.041)	(0.038)	(0.047)	(0.028)	(0.026)	(0.045)
No. of obs.	3,492	3,492	3,492	3,492	3,492	3,492

The treatment-on-the-treated effect is estimated using the assigned treatment as an instrument for whether the student is in fact enrolled in a PSL school during the 2016/2017 academic year. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) All schools

(b) Non-constrained schools

(c) Constrained schools

Figure A.3: Treatment effect on enrollment by grade

Note: These figures show differences in enrollment (2016/2017 compared to the 2015/2016 academic year) by grade. The dots represent point estimates, while the bars represent 95% confidence intervals. Panel A.3a shows the effect across all schools. Panel A.3b shows the effect in non-constrained school-grades, and Panel A.3c shows the effect in constrained school-grades.

Table A.7: Treatment effects across various measures of difference in student ability

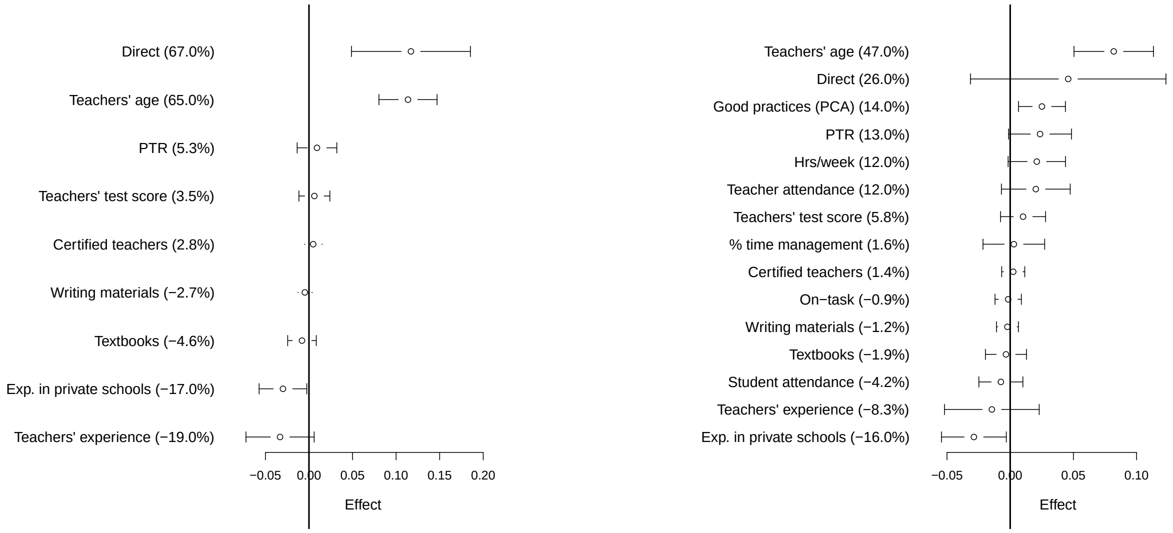
	Difference	Difference (F.E.)	Difference (F.E. + Controls)
	(1)	(2)	(3)
Panel A: Base IRT model			
English	0.17**	0.17***	0.18***
	(0.08)	(0.04)	(0.03)
Math	0.17***	0.19***	0.18***
	(0.07)	(0.04)	(0.03)
Panel B: IRT model per grade			
English	0.21**	0.23***	0.25***
	(0.09)	(0.05)	(0.04)
Math	0.22***	0.25***	0.26***
	(0.08)	(0.05)	(0.04)
Panel C: Base PCA			
English	0.16**	0.17***	0.16***
	(0.08)	(0.04)	(0.03)
Math	0.18***	0.19***	0.24***
	(0.06)	(0.05)	(0.04)
Panel D: PCA per grade			
English	0.17*	0.18***	0.20***
	(0.09)	(0.05)	(0.04)
Math	0.21***	0.24***	0.25***
	(0.07)	(0.05)	(0.05)
Panel E: % correct answers			
English	2.99**	3.00***	2.97***
	(1.40)	(0.75)	(0.55)
Math	3.88***	4.14***	4.24***
	(1.32)	(0.83)	(0.71)
Observations	3,492	3,492	3,492

Column 1 shows the simple difference between treatment and control; Column 2 shows the difference taking into account the randomization design — i.e., including "pair" fixed effects; and Column 3 shows the difference taking into account other student and school controls. Panel A uses our default IRT model and normalizes test scores using the same mean and standard deviation across all grades. Panel B estimates a different IRT model for each grade. Panel C estimates students' ability as the first component from a principal component analysis (PCA), and normalizes test scores using a common mean and standard deviation across all grades. Panel D performs a different principal component analysis for each grade. Panel E calculates the percentage of correct responses.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: Student selection

	(1) Same school	(2) Same school	(3) Same school
Treatment	0.061	0.012	0.021
	(0.082)	(0.026)	(0.019)
Treatment × Age	-0.0042		
	(0.0064)		
Treatment × Male		-0.011	
		(0.028)	
Treatment × Asset Index (PCA)			-0.0061
			(0.011)
No. of obs.	3,487	3,487	3,428

Standard errors are clustered at the school level. The sample is the original treatment and control allocation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) Inputs

(b) Inputs & Management

Figure A.4: Direct and causal mediation effects

Note: This figure provides the direct effect ($\beta_{1.5}$) and the mediation effects ($\beta_{1.4} \times \theta_{1.5}$) for all the possible mediators. The point estimates within the same panel are comparable to each other. Point estimates and 90% confidence intervals are plotted. Panel A.4a shows treatment effects allowing only changes in inputs as mediators. Panel A.4b shows treatment effects allowing changes in inputs and in the use of inputs as mediators.

Table A.9: ITT treatment effects, by whether class size caps are binding without including adjectent grades

	(1) Δ enrollment	(2) % same school	(3) % in school	(4) Test scores
Constrained=0 × Treatment	2.96***	3.83***	1.53**	0.10**
	(1.08)	(1.43)	(0.67)	(0.039)
Constrained=1 × Treatment	17.3**	-12.5**	-13.4***	0.36***
	(7.53)	(5.84)	(3.53)	(0.14)
No. of obs.	1,256	2,773	2,636	2,641
Mean control (Unconstrained)	-0.43	82.57	94.00	0.08
Mean control (Constrained)	-9.03	80.95	100.00	-0.33
$\alpha_0 =$ Constrained - Unconstrained	14.30	-16.34	-14.95	0.26
p-value ($H_0 : \alpha_0 = 0$)	0.07	0.01	0.00	0.07

This table mirrors Table 1.5, but adjacent grades are not included in the calculation of the constrained indicator. Column 1 uses school-grade level data. Columns 2 - 4 use student level data. The independent variable in Column 4 is the composite test score. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. There were 216 constrained classes at baseline (holding 35% of students), and 1,448 unconstrained classes at baseline (holding 65% of students).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.10: Intensive margin effect on teacher attendance and classroom observation with Lee bounds

	(1)	(2)	(3)
	Difference	Difference (F.E)	90% CI
		(F.E)	(bounds)
Panel A: Spot check (N = 929)			
% on schools campus	15.75***	14.17***	2.51
	(4.45)	(3.75)	28.11
% in classroom	9.91**	9.96**	-1.34
	(4.78)	(3.86)	24.44
B: Classroom observation (N = 143)			
Active instruction (% class time)	7.98	7.62	-4.75
	(4.86)	(4.75)	19.92
Passive instruction (% class time)	3.44	4.72	-4.93
	(2.95)	(3.23)	9.62
Classroom management (% class time)	10.16***	10.33***	0.77
	(2.85)	(3.32)	16.99
Teacher off-task (% class time)	-21.58***	-22.66***	-40.24
	(5.92)	(6.26)	-10.32
Student off-task (% class time)	-2.54	-5.19	-16.05
	(5.26)	(4.88)	12.63
Panel C: Inputs (N = 143)			
Number of seats	0.06	0.58	-7.22
	(2.21)	(1.90)	5.36
% with students sitting on the floor	-1.82	-1.51	-7.48
	(2.94)	(2.61)	2.76
% with chalk	17.51***	16.58***	9.47
	(5.29)	(5.50)	27.85
% of students with textbooks	19.48***	22.60***	-1.21
	(6.33)	(6.32)	34.87
% of students with pens/pencils	8.88**	8.16**	1.36
	(4.19)	(4.10)	20.98

This table presents the difference between treatment and control (Column 1), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2. Column 3 shows the 90% confidence interval using Lee (2009) bounds. Panel A provides results from the spot check using the EMIS data (2015/2016) on teachers as a baseline, and treating teachers who no longer teach at school as attriters. Panel B provides the classroom observation information without imputing values for schools not in session during our visit, and treating the missing information as attrition. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.11: Treatment effect on schools' good practices

	(1) Control	(2) Difference	(3) Difference (F.E)
Maintains an enrollment log	0.80	0.10*	0.10*
	(0.40)	(0.05)	(0.05)
Log contains student name	0.82	0.08	0.08
	(0.39)	(0.05)	(0.05)
Log contains student grade	0.84	0.10**	0.10**
	(0.37)	(0.05)	(0.05)
Log contains student age	0.64	0.00	-0.00
	(0.48)	(0.07)	(0.07)
Log contains student gender	0.83	0.07	0.06
	(0.38)	(0.05)	(0.05)
Log contains student contact information	0.13	0.13**	0.13**
	(0.34)	(0.06)	(0.06)
Enrollment log is clean and neat	0.26	0.13*	0.13*
	(0.44)	(0.07)	(0.07)
Maintains official schedule	0.89	0.09**	0.09***
	(0.31)	(0.04)	(0.03)
Official schedule is posted	0.70	0.14**	0.14**
	(0.46)	(0.06)	(0.06)
Has a PTA	0.98	0.01	0.01
	(0.15)	(0.02)	(0.02)
Principal has PTA head's number at hand	0.26	0.15**	0.15**
	(0.44)	(0.07)	(0.06)
Maintains expenditure records	0.09	0.05	0.05
	(0.28)	(0.05)	(0.05)
Maintains a written budget	0.22	0.04	0.04
	(0.41)	(0.06)	(0.06)
Observations	92	185	185

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) group, as well as the difference between treatment and control (Column 2), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 3. Standard errors are clustered at the school level. The sample is the original treatment and control allocation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.12: Treatment effect on household expenditure

	(1) Treatment	(2) Control	(3) Difference	(4) Difference (F.E)
Fees (USD/year)	5.72	8.04	-2.32**	-2.89***
	(10.22)	(9.73)	(0.96)	(0.61)
Tutoring (USD/year)	0.35	0.38	-0.04	-0.03
	(1.22)	(1.34)	(0.09)	(0.08)
Textbooks (USD/year)	0.61	0.86	-0.25**	-0.22**
	(1.44)	(1.65)	(0.12)	(0.09)
Copy books (USD/year)	1.02	1.09	-0.07	-0.08
	(1.96)	(1.94)	(0.15)	(0.13)
Pencils (USD/year)	3.23	2.95	0.28	0.20
	(3.05)	(2.88)	(0.31)	(0.16)
Uniform (USD/year)	9.24	11.45	-2.21***	-1.95***
	(6.31)	(5.18)	(0.63)	(0.42)
Food (USD/year)	42.94	46.43	-3.50	-1.66
	(70.95)	(76.05)	(6.90)	(3.93)
Other (USD/year)	3.42	3.06	0.36	0.32
	(4.56)	(4.28)	(0.34)	(0.27)
Observations	595	520	1,115	1,115

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) and treatment (Column 2) groups, as well as the difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 4. Standard errors are clustered at the school level. The sample is the original treatment and control allocation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.13: Treatment effect on household engagement

	(1) Treatment	(2) Control	(3) Difference	(4) Difference (F.E)
Attended school meeting	0.76	0.77	-0.01	0.03
	(0.43)	(0.42)	(0.04)	(0.02)
Made cash donation	0.12	0.11	0.02	-0.00
	(0.33)	(0.31)	(0.02)	(0.02)
Made in-kind donation	0.03	0.04	-0.01	-0.02
	(0.17)	(0.20)	(0.01)	(0.01)
Donated work	0.13	0.15	-0.01	-0.00
	(0.34)	(0.35)	(0.03)	(0.02)
Helped with homework	0.58	0.61	-0.03	-0.04
	(0.49)	(0.49)	(0.04)	(0.03)
Observations	619	543	1,162	1,162

This table presents the mean and standard error of the mean (in parenthesis) for the control (Column 1) and treatment (Column 2) groups, as well as the difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 4. Standard errors are clustered at the school level. The sample is the original treatment and control allocation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

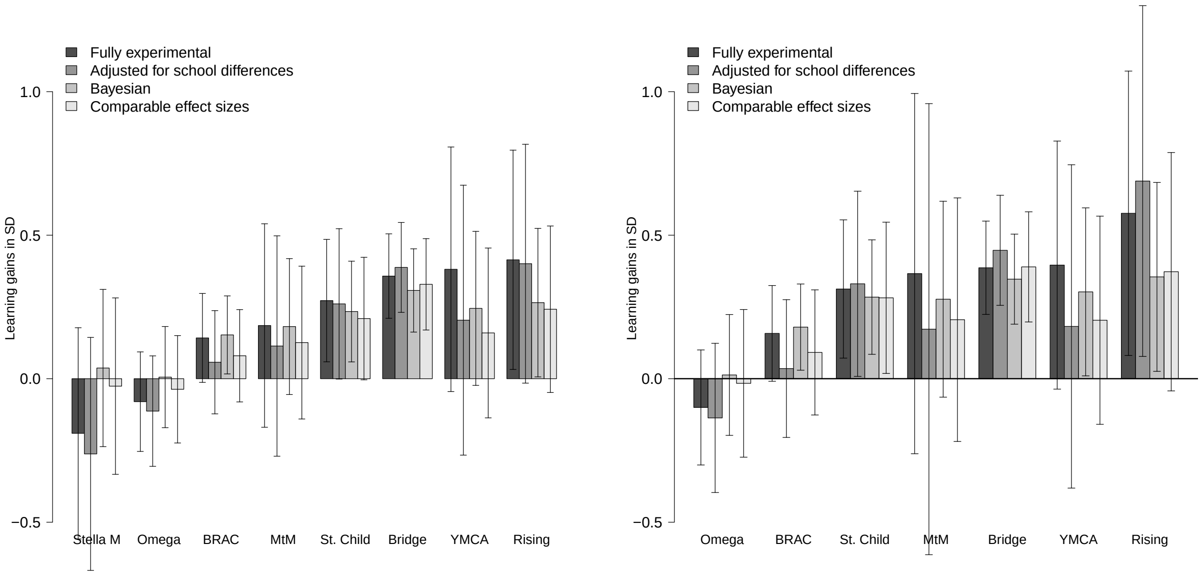
Table A.14: Control variables

Student controls	Question	Questionnaire
Wealth index	A1-A7	Student
Age	B1	Student
Gender	B2	Student
Grade (2015/2016)	B6a	Student
School controls		
Enrollment (2015/2016)	C1	Principal
Infrastructure quality (2015/2016)	L1-L3	Principal
Travel time to nearest bank	L6	Principal
Rurality	L7	Principal
NGO programs in 2015/2016	M1-M4	Principal
Donations in 2015/2016	N1A-N3b_a_5	Principal

Table A.15: Mediated treatment effects, when the effect of mediators on learning is estimating using only control schools

	% of total treatment effect	
	(1)	(2)
Direct	79.0%	66.0%
PTR	6.1%	6.2%
Teachers' age	70.0%	67.0%
Teachers' experience	-55.0%	-49.0%
Certified teachers		2.5%
Exp. in private schools		6.3%
Teachers' test score		2.0%
Textbooks		0.4%
Writing materials		-1.9%

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) Intention-to-treat (ITT) effect

(b) Treatment-on-the-treated effect (ToT)

Figure A.5: Treatment effects by provider

Note: These figures show the raw, fully experimental treatment effects, the effects after adjusting for differences in school characteristics (before the Bayesian hierarchical model), the effects after applying a Bayesian hierarchical model (but without adjusting for school differences), and the comparable treatment effects after adjusting for differences in school characteristics and applying a Bayesian hierarchical model. Figure A.5a shows the intention-to-treat (ITT) effect, while Figure A.5b shows the treatment-on-the-treated (ToT) effect. The ToT effects are larger than the ITT effects due to providers replacing schools that did not meet the eligibility criteria, providers refusing schools, or students leaving PSL schools. Stella Maris had full non-compliance at the school level and therefore there is no ToT effect for this provider.

Table A.16: Raw (fully experimental) treatment effects by provider

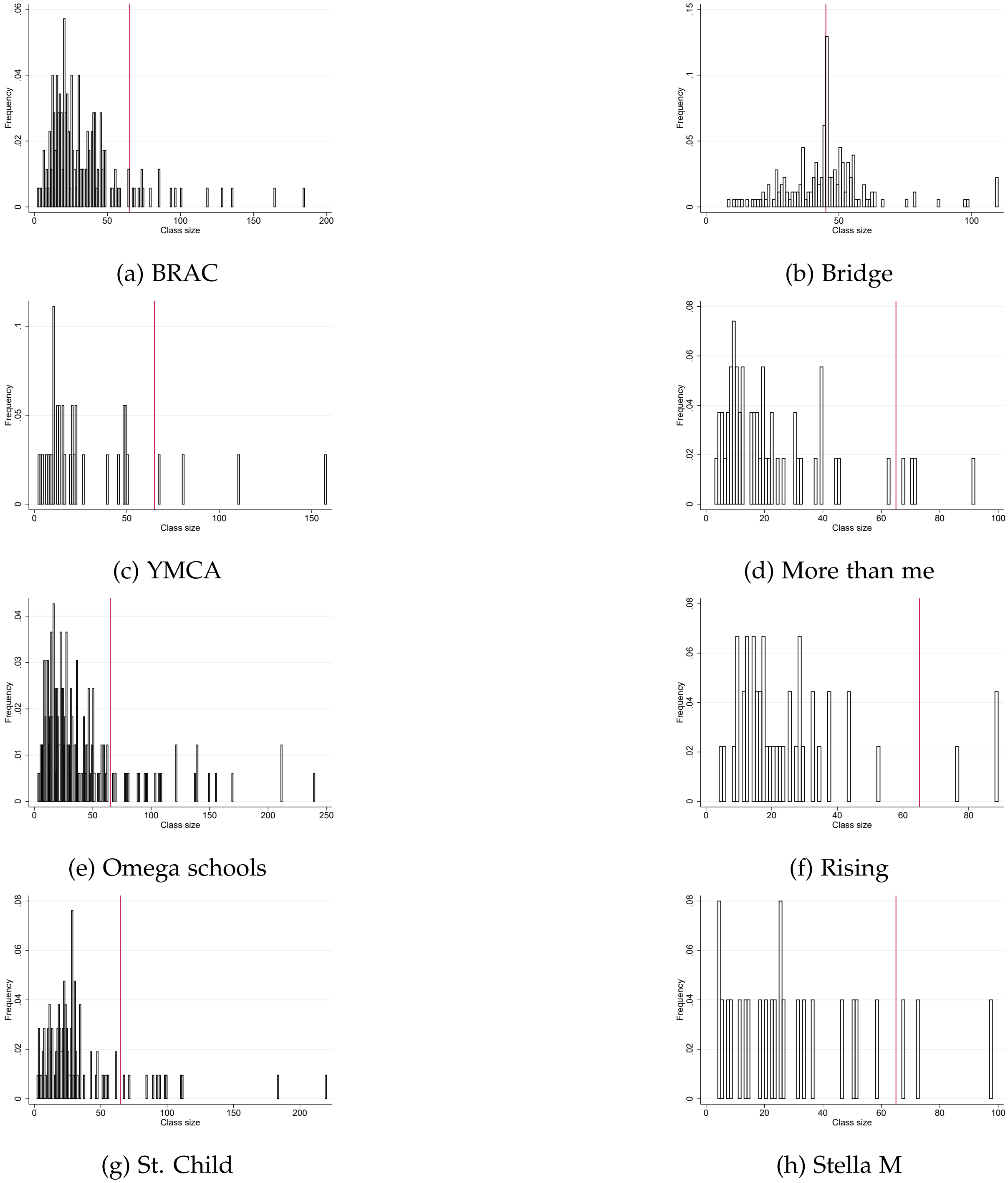
	(1) BRAC	(2) Bridge	(3) YMCA	(4) MtM	(5) Omega	(6) Rising	(7) St. Child	(8) Stella M
Panel A: Student test scores								
English (standard deviations)	0.19**	0.28***	0.19	-0.07	0.35	0.23*	-0.23	0.58**
	(0.10)	(0.09)	(0.22)	(0.11)	(0.24)	(0.13)	(0.23)	(0.26)
Math (standard deviations)	0.10	0.39***	0.19	-0.06	0.41*	0.28**	-0.17	0.26
	(0.09)	(0.09)	(0.21)	(0.10)	(0.23)	(0.13)	(0.22)	(0.26)
Composite (standard deviations)	0.14	0.36***	0.19	-0.08	0.41*	0.27**	-0.19	0.38
	(0.09)	(0.09)	(0.22)	(0.11)	(0.23)	(0.13)	(0.22)	(0.26)
Panel B: Changes to the pool of teachers								
% teachers dismissed	-6.83	49.98***	15.86	-9.12	-5.73	-2.63	-3.51	20.96
	(6.51)	(6.36)	(11.90)	(6.89)	(12.88)	(8.59)	(14.52)	(14.52)
% new teachers	39.62***	62.95***	69.54***	24.34*	24.38	40.94**	-21.93	62.20**
	(12.29)	(12.02)	(22.46)	(13.01)	(24.31)	(16.21)	(27.41)	(27.41)
Age in years (teachers)	-5.03***	-10.92***	-11.20***	-5.46***	-10.75***	-5.79**	-4.53	3.25
	(1.93)	(2.01)	(3.52)	(2.03)	(3.82)	(2.54)	(4.30)	(4.30)
Panel C: Enrollment and access								
Δ enrollment	38.02	-13.26	-25.98	51.27	19.31	44.86	-15.92	45.38
	(34.33)	(33.60)	(62.76)	(35.26)	(67.84)	(45.21)	(76.59)	(76.53)
Δ enrollment (constrained grades)	0.00	-23.85**	0.00	0.28	0.00	32.15	0.00	0.00
	(0.00)	(11.19)	(0.00)	(37.16)	(0.00)	(61.95)	(0.00)	(0.00)
Student attendance (%)	20.09**	5.25	37.81**	18.01*	28.76	19.56*	9.71	13.53
	(9.02)	(9.05)	(16.50)	(9.53)	(17.82)	(11.88)	(23.32)	(20.11)
% students still attending any school	1.22	5.21	-3.11	4.73	2.78	3.57	5.96	4.49
	(4.45)	(4.21)	(10.17)	(4.98)	(10.96)	(6.09)	(10.56)	(12.20)
% students still attending same school	0.78	4.41**	0.62	1.60	3.73	-0.83	1.03	-0.80
	(2.20)	(2.08)	(5.03)	(2.46)	(5.42)	(3.01)	(5.22)	(6.03)
Observations	40	45	8	12	38	10	24	8

Table presents the raw treatment effect for each provider on different outcomes. The estimates for each provider are *not* comparable to each other without further assumptions; we do not include a test of equality. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.17: Descriptive statistics by provider and treatment

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Provider	Treatment	Schools	Teachers				Enrollment			Constrained classes		
			15/16	16/17	Left	New	Classes	15/16	16/17	Constrained classes	15/16	16/17
BRAC	0	20	141	148	41	48	180	5,694	5,107	10	780	703
BRAC	1	20	141	209	33	101	180	5,684	5,872	11	1,130	1,138
Bridge	0	22	177	174	38	35	198	7,110	6,610	61	3,969	3,648
Bridge	1	23	236	212	174	150	207	9,788	8,282	72	6,909	3,475
YMCA	0	4	20	22	1	3	36	729	727	2	142	120
YMCA	1	4	27	40	6	19	36	908	1,068	2	217	238
MtM	0	6	52	41	21	10	54	1,140	1,312	2	155	167
MtM	1	6	46	64	20	38	54	1,145	1,223	2	171	159
Omega	0	19	132	130	33	31	171	4,895	5,200	12	1,255	1,232
Omega	1	19	151	196	26	71	171	5,764	6,841	19	1,953	2,446
Rising	0	5	47	43	23	19	45	1,209	1,308	2	202	185
Rising	1	5	36	47	11	22	45	918	1,134	1	87	89
St. Child	0	12	88	68	29	9	108	3,094	2,794	7	738	557
St. Child	1	12	81	100	22	41	108	3,351	3,506	9	877	797
Stella M	0	4	20	20	8	8	36	765	683	1	73	45
Stella M	1	4	31	27	9	5	36	958	978	3	213	192

Teachers in 2015/2016 are taken from the EMIS data; teachers in 2016/2017 are taken from our first-year follow-up data. “Left” refers to teachers in the 2015/2016 EMIS data who are not working at the school at the end of 2016/2017. “New” are the teachers working at the school at the end of the 2016/2017 who are not in the 2015/2016 EMIS data. “Constrained classes” are those with more students in 2015/2016 than the class size cap.



(a) BRAC

(b) Bridge

(c) YMCA

(d) More than me

(e) Omega schools

(f) Rising

(g) St. Child

(h) Stella M

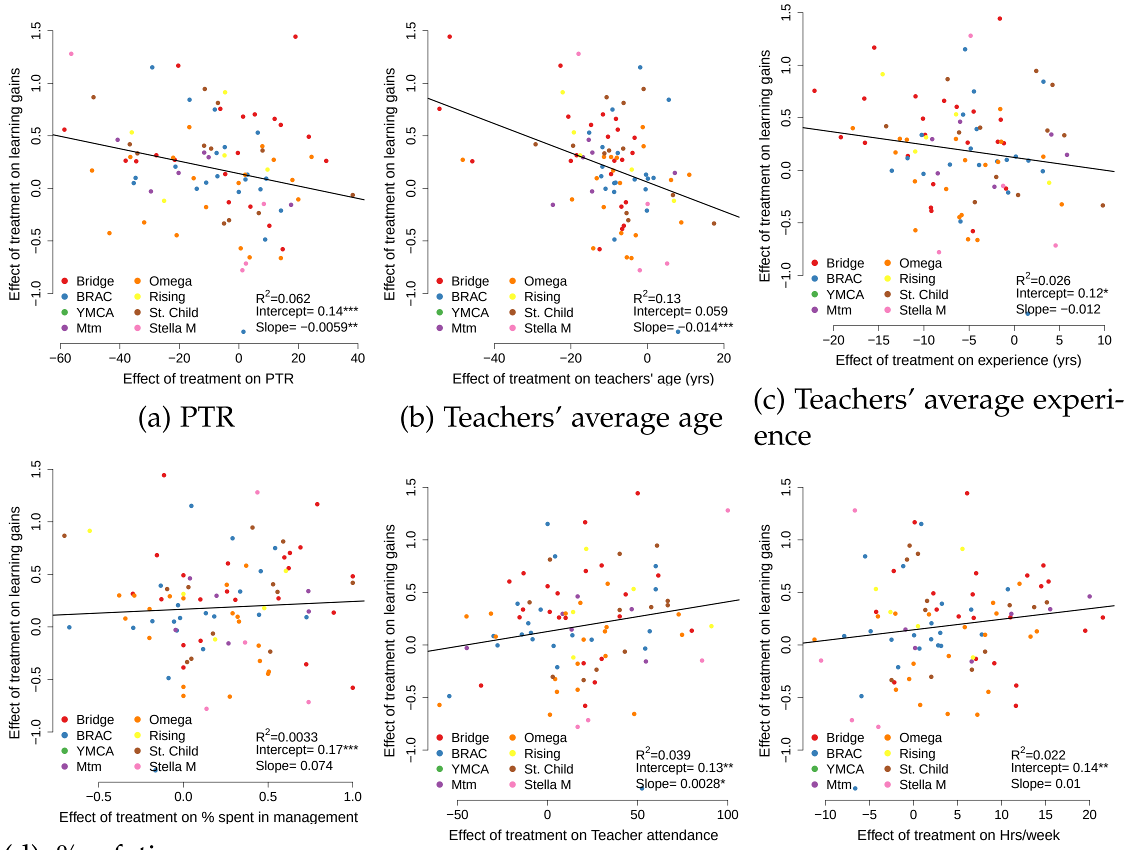
Figure A.6: Class sizes and class caps

Note: These figures show the distribution of class sizes in treatment schools during the 2016/2017 academic year, as well as the class cap for each provider. The cap for all providers is 65 students, except for Bridge that has a cap of 45.

A.2 Treatment effects at the matched-pair level

We can estimate the treatment effect for all 93 matched-pairs in our sample. We do this for learning outcomes, as well as for intermediate outcomes (e.g., teacher attendance). As an exploratory analysis, we plot the treatment effects for learning outcomes and for intermediate outcomes in Figure A.7.[1] Table A.18 shows the correlation between different treatment effects. The slope of the OLS line between two variables (y and x) is equal to $Cor(x,y)\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$, and therefore there is a direct relationship between the slope of the fitted lines in Figure A.7 and the correlations in Table A.18.

[1] We use the same intermediate outcomes determined by "Double Lasso" in Section 1.4 as high predictors of learning gains.



(a) PTR

(b) Teachers' average age

(c) Teachers' average experience

(d) % of time on management

(e) Teacher attendance

(f) Hours per week

Figure A.7: Correlation between treatment effects at the matched-pair level for different outcomes

Note: Each dot represents a matched-pair. The y-axis is the treatment effect on learning outcomes. The x-axis is the treatment effect on the intermediate outcomes determined by "Double Lasso" in Section1.4. In Figure A.7a the x-axis is the effect on the pupil-teacher ratio (PTR); in Figure A.7b is the effect on the average age of teachers; in Figure A.7c is the effect on the average experience of teachers; in Figure A.7d is the effect on the proportion of time the principal spends on management activities; in Figure A.7e is the effect on teacher attendance; and in Figure A.7f is the effect on the hours per week of instructional time according to the official time schedule.

Table A.18: Correlation between treatment effects at the matched-pair level

Variable:	Learning	PTR	Age	Experience	Management	Attendance	Hours/Week
Learning	1						
Age	-0.37***	1					
PTR	-0.25**	0.025	1				
Attendance	0.20*	0.056	-0.034	1			
Experience	-0.16	0.47***	-0.054	0.12	1		
Hours/Week	0.15	-0.19*	0.11	-0.00049	-0.18*	1	
Management	0.057	-0.020	-0.071	0.34***	0.19*	0.084	1

Each number represents the correlation between treatment effects at the matched-pair level. Learning refers to treatment effects on learning outcomes, PTR is the pupil-teacher-ratio, Age is the average age of teachers, Experience is the average experience of teachers, Management is the proportion of time the principal spends on management activities, Attendance is teachers' attendance, and Hours/Week is the hours per week of instructional time according to the official time schedule. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.3 Tracking and attrition

A potential issue with our sampling strategy is differential attrition at each round of data collection. In the first round, enumerators were instructed to sample 20 students from the 2015/2016 enrollment logs, track them, and test them. However, if a student had moved to another village, had died, or was impossible to track, the enumerators were instructed to sample another student. Thus, even at the first round an endogenous sampling problem arises if treatment makes students easier or harder to track in combination with enumerator shrinkage. To mitigate this issue, enumerators participated in additional training on tracking and its importance and were provided with a generous amount of tracking time. Students were tracked to their homes and tested there when not available at school. As Table A.19 shows, we have no reason to believe that this issue arose. The effort required to track students was different between treatment and control (it is easier to track students at the school), yet the total number of students sampled, to obtain a sample of 20 students, is balanced between treatment and control (see Table A.19).

A.4 Test design

Most modules follow the Early Grade Reading Assessment (EGRA), Early Grade Mathematics Assessment (EGMA), Uwezo, and Trends in International Mathematics and Science Study (TIMSS) assessments. For the first wave of data collection the test contained a module for each of the following skills: object identification (like the Peabody Picture Vocabulary Test), letter reading (adapted from EGRA), word reading (adapted from EGRA), a preposition module, reading comprehension (adapted from Uwezo), listening comprehension (adapted from EGRA), counting (adapted from Uwezo), number

Table A.19: Tracking and sampling in the first wave of data collection

	(1) Treatment	(2) Control	(3) Difference	(4) Difference (F.E)
Number of students sampled	24.8	24.6	0.13	0.035
	(5.74)	(5.10)	(0.81)	(0.81)
Found at the school	18.2	16.7	1.49^{***}	1.555^{***}
	(2.30)	(4.70)	(0.55)	(0.54)
Found at home	1.73	2.91	-1.18^{**}	-1.223^{**}
	(2.12)	(3.97)	(0.48)	(0.47)
Interviewed	19.8	19.5	0.30	0.320
	(0.83)	(2.18)	(0.25)	(0.26)
Observations	88	90	178	171

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) and treatment (Column 2) groups, as well as the difference between treatment and control (Column 3), and the difference taking into account the randomization design (i.e., including "pair" fixed effects) in Column 4. The table shows the average number of students we sampled (and tried to track), the number of students we were able to track at the assigned school or at home, and the total number of students we tracked and found during the first round of data collection. Standard errors are clustered at the school level. The sample is the original treatment and control allocation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

discrimination (adapted from Uwezo), number identification (adapted from EGMA), addition (adapted from Uwezo and EGMA), subtraction (adapted from Uwezo and EGMA), multiplication (adapted from Uwezo and EGMA), division (adapted from Uwezo and EGMA), shape identification, fractions, and word problems in mathematics.

For the second round of data collection the test did not include the following modules: Prepositions, shape identification, and fractions. These modules were excluded given the low variation in responses in the first wave of data collection and to make space for new modules. Instead, the test included letter, word and number dictation, and a verb and a pronoun module. Additionally, we included some "conceptual" questions from TIMSS released items (items M031317 and M031316) that do not resemble the format of standard textbook exercises but rather test knowledge in an unfamiliar way. The number identification module remained exactly the same across rounds of data collection (to provide us with absolute learning curves on these two items), while every other module was different. In addition, the word and number identification modules were identical to the EGRA/EGMA assessments used in Liberia previously (for comparability with other impact evaluations taking place in Liberia, most notably USAID's reading program Piper and Korda 2011 and the LTTP program King et al. 2015), but during the first round of data collection they were different. Two of the reading comprehension questions were taken from the Pre-Pirls released items (L11L01C and L11L02M) and one of the word problems was taken from TIMSS released items (M031183). Finally, we added a Raven's style module to measure the students' abstract thinking abilities.

A.5 Bayesian hierarchical model

Figure A.8 shows the distribution of treatment effects across all 93 matched-pairs in our sample. This gives us an idea of what the scale for τ should be.

Figures A.9 and A.10 show the posterior treatment effects and standard errors for different values of τ. Assuming $\tau = 0$ is equivalent to imposing that the treatment effect is the same across all providers (and thus that the average treatment effect is the best estimator for all providers). Larger values of τ correspond to minimal pooling. Figure A.11 shows the posterior distribution of τ in the case of a flat prior.

Table A.20 shows the posterior treatment effect and standard error across different priors, as suggested by Gelman (2006).

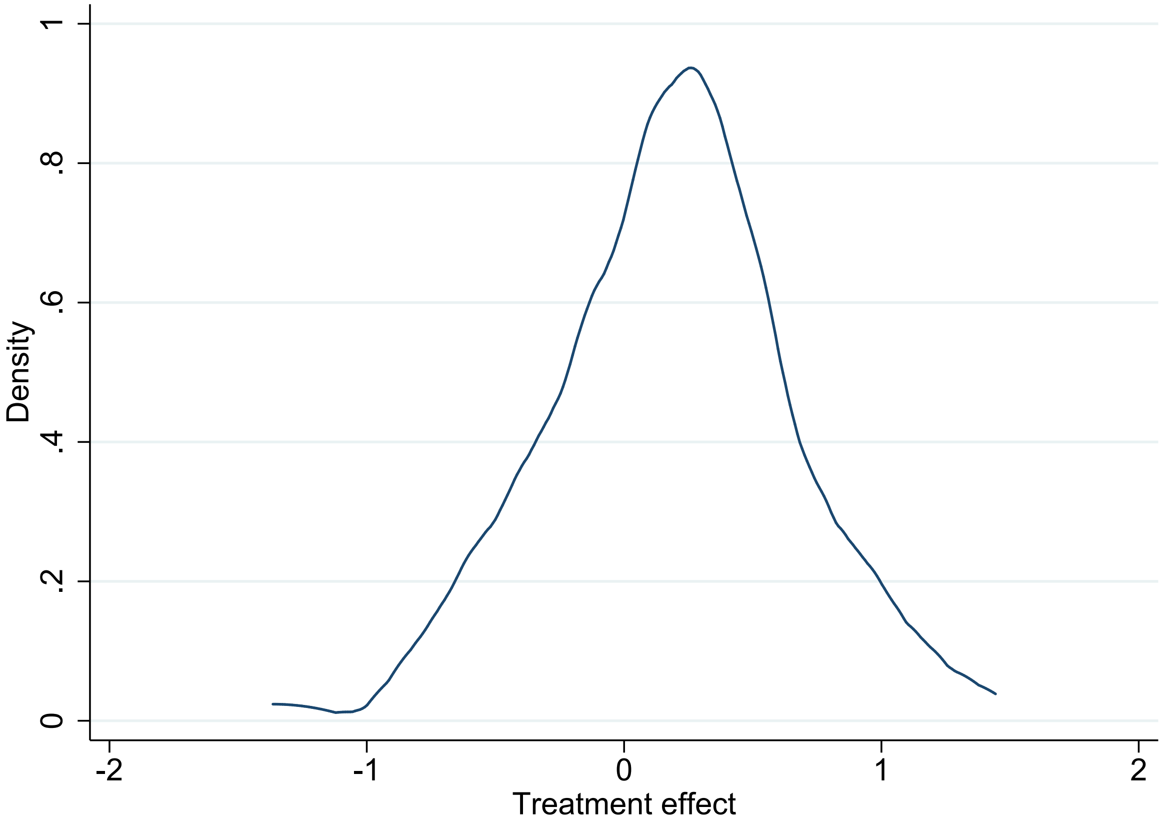


Figure A.8: Treatment effect distribution across all 93 matched-pairs

A.6 Satisfaction and support for the PSL program

For a government program to be politically viable, it needs the support of those affected by it. The PSL program has met with resistance from teachers' unions and provoked criticism from international organizations and the media.[2] Data we collected independently on levels of support for and satisfaction with the PSL program among students, parents, and teachers are shown in Table A.21.

There are three main messages from the data in this table. First, students are

[2] The Liberian government's announcement of the PSL program generated international coverage, from the BBC to the New York Times, focused on outsourcing and privatization The New York Times 2016; BBC Africa 2016; Vox World 2016; Foreign Policy 2016; Mail & Guardian Africa 2016b, 2016a, and even condemnation from a UN Special Rapporteur that Liberia was abrogating its responsibilities under international law OHCHR 2016.

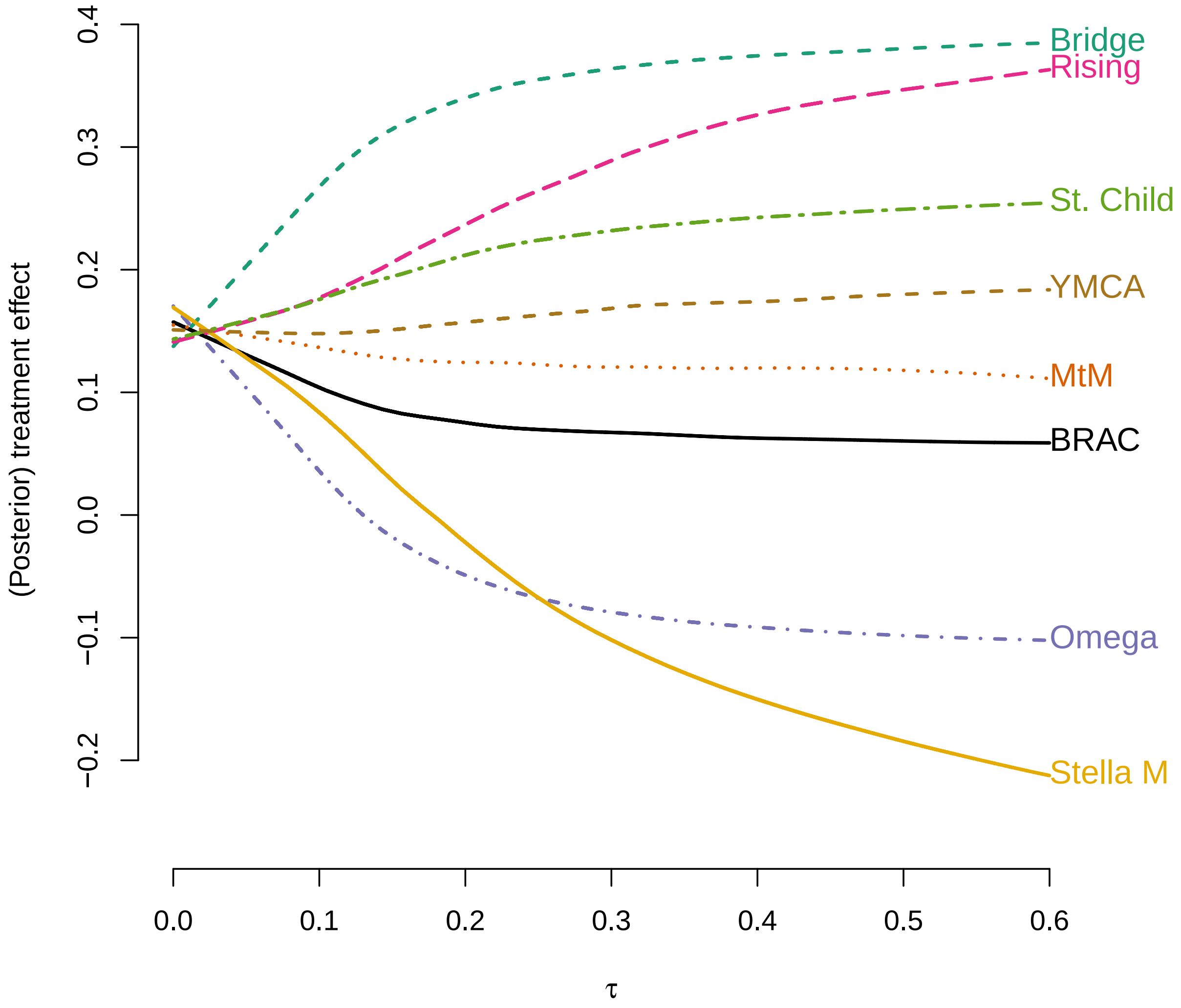


Figure A.9: Posterior treatment effects by provider for different values of τ

happier in PSL schools than in traditional public schools (measured by whether they think going to school is fun). Second, households with children in PSL schools (enrolled in 2015/2016) are 7.4 percentage points (p-value .022) more likely to be satisfied with the education their children are receiving. Additionally, most households, even in the control group, would prefer that providers manage more schools the following year

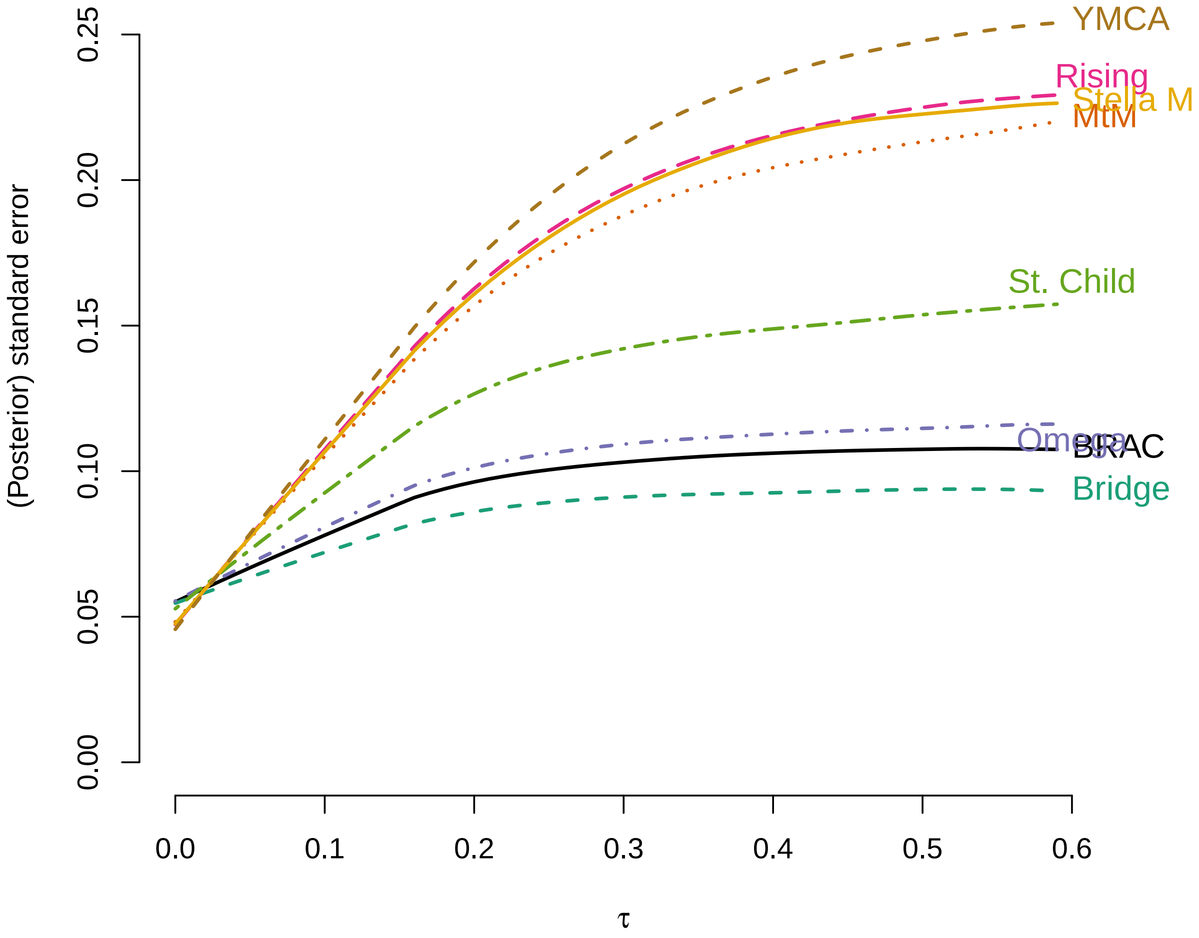


Figure A.10: Posterior standard errors by provider for different values of τ

(87% of households overall) and would rather send their children to a school managed by a provider than to a traditional public school (72% of households overall). Third, despite any (statistically significant) difference in the satisfaction of teachers across treatment and control schools, most teachers, even in control schools, would rather work in a school managed by a provider (64% of teachers overall) and would prefer

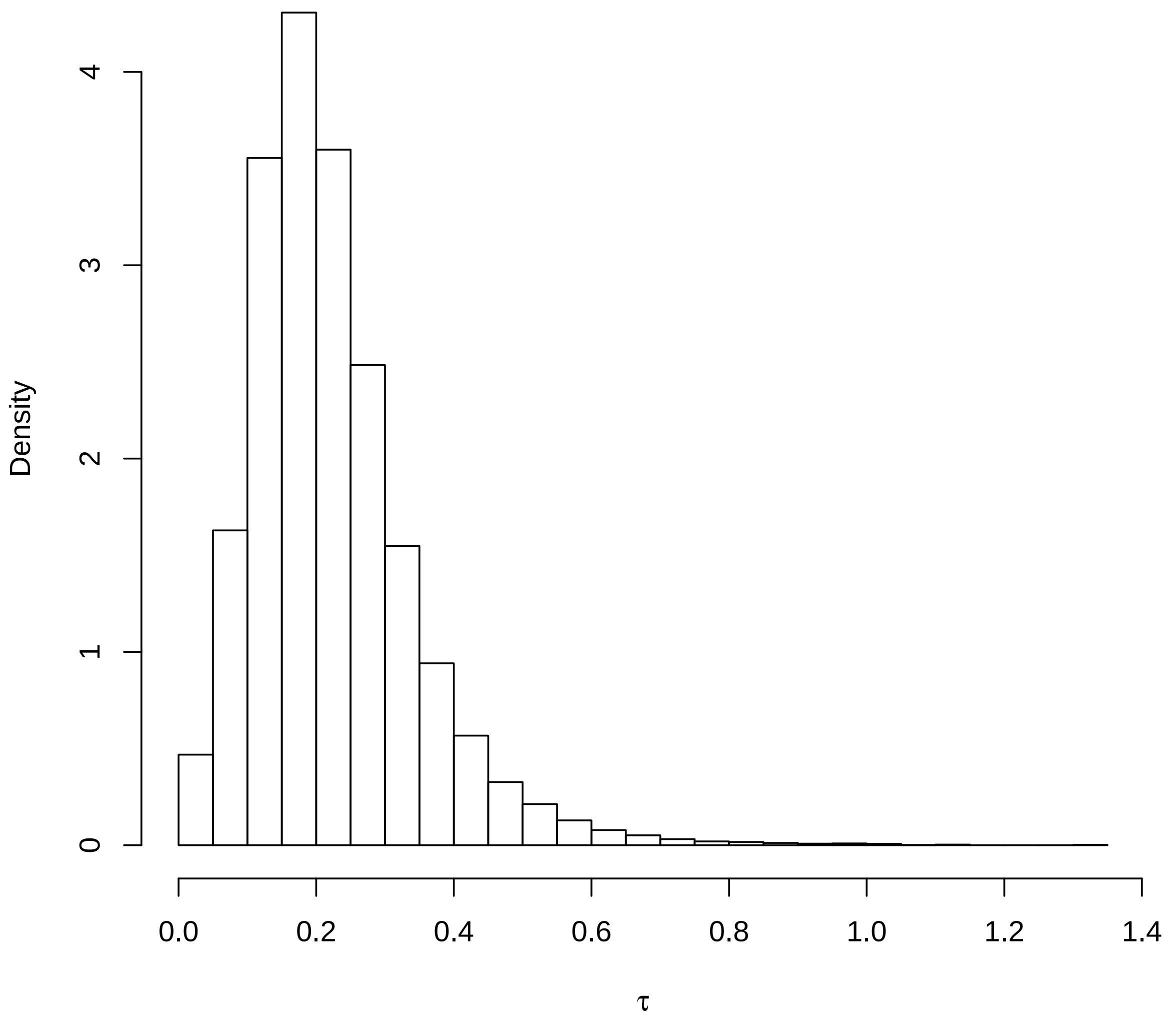


Figure A.11: Posterior distribution of τ

that providers managed more schools the following year (85% of teachers overall).

Table A.20: Posterior treatment effects and standard errors for different priors

	(1) BRAC	(2) Bridge	(3) YMCA	(4) MtM	(5) Omega	(6) Rising	(7) St.Child	(8) Stella
Flat prior	0.080	0.329^{***}	0.126	-0.037	0.242	0.210	-0.026	0.159
	(0.098)	(0.097)	(0.162)	(0.114)	(0.176)	(0.130)	(0.187)	(0.180)
Cauchy(0,25)	0.080	0.329^{***}	0.127	-0.037	0.241	0.209	-0.025	0.160
	(0.098)	(0.097)	(0.162)	(0.114)	(0.176)	(0.130)	(0.186)	(0.180)
Half-normal	0.081	0.327^{***}	0.127	-0.035	0.241	0.208	-0.023	0.160
	(0.097)	(0.097)	(0.161)	(0.114)	(0.175)	(0.128)	(0.186)	(0.178)
Half-t(4)	0.080	0.327^{***}	0.127	-0.035	0.239	0.208	-0.022	0.160
	(0.098)	(0.097)	(0.160)	(0.114)	(0.175)	(0.128)	(0.184)	(0.178)

This table presents the treatment effect and the standard error for each provider across different priors. The Cauchy prior has a location parameter of zero and a scale of 25. The half-normal is a folded standard normal distribution. The half-t is a folded t student distribution with 4 degrees of freedom. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.7 What “managing” a school means in practice

In this section we show data from the teacher survey on provider activities in each school. Our pair-matched design allowed us to ask provider-specific questions of teachers at control schools; their responses are shown in Tables A.22-A.24. First, no provider visited a control school on a a regular basis, nor did they provide control schools with inputs. However, only 62% of treatment schools received provider visits on a regular basis (recall that there is non-compliance in our sample). Managing a school does seem to entail a wide range of activities. Teachers report that providers provided hard inputs (textbooks, copybooks, tablets, and repairs) and soft inputs (training and community meetings). The two most likely activities during the last visit from the provider entailed either checking attendance and school records and/or observing teaching practices.

Table A.21: Student, household and teacher satisfaction and opinions

	(1) Control	(2) Difference	(3) Difference (F.E)
Panel A: Students (N = 3,492)			
School is fun (%)	52.37	5.94***	5.90**
	(15.52)	(2.28)	(2.45)
Panel B: Households (N = 185)			
% satisfied with school	67.46	7.42**	7.44**
	(23.95)	(3.20)	(3.23)
% have heard of PSL	14.35	3.46	3.44
	(16.12)	(2.33)	(2.22)
% have heard of provider	23.93	33.00***	33.08***
	(24.41)	(4.10)	(3.66)
% want provider managing more schools	81.69	8.94*	11.18**
	(34.79)	(4.88)	(4.83)
% preferring to send child to PSL school	61.96	16.87***	16.73**
	(42.13)	(6.09)	(6.92)
Panel C: Teachers (N = 185)			
% would choose teaching as a career	88.23	2.51	1.99
	(17.81)	(2.32)	(2.56)
% work a second job	23.77	-7.50**	-7.45**
	(25.80)	(3.45)	(3.74)
Job satisfaction index (PCA)	-0.14	0.18	0.21
	(0.86)	(0.13)	(0.14)
% have heard of PSL	28.43	36.38***	35.19***
	(27.01)	(4.50)	(4.03)
% have heard of operator	39.76	54.23***	54.76***
	(36.46)	(4.53)	(4.28)
% would rather work at provider school	43.12	27.87***	21.93***
	(36.80)	(6.00)	(5.98)
% want provider managing more schools	81.15	4.65	1.46
	(31.66)	(4.97)	(5.15)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) groups, as well as the difference between treatment and control (Column 2), and the difference taking into account the randomization design (i.e., including "'pair"' fixed effects) in Column 3. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.22: Teachers' general opinion of PSL

	(1) Control	(2) Difference	(3) Difference (F.E)
Heard of PSL	0.28	0.36^{***}	0.35^{***}
	(0.45)	(0.04)	(0.03)
Heard of provider	0.40	0.54^{***}	0.55^{***}
	(0.49)	(0.05)	(0.03)
Provider staff visits at least once a week	0.00	0.64^{***}	0.62^{***}
	(0.00)	(0.04)	(0.04)
Provider support rating (0-100)	15.08	52.22^{***}	53.48^{***}
	(30.50)	(3.88)	(3.64)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) group, as well as the difference between treatment and control (Column 2), and the difference taking into account the randomization design (i.e., including "'pair"' fixed effects) in Column 3. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. N = 1,097. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.8 Standard deviation and equivalent years of schooling

Figure A.12 shows how many standard deviations are equal to an extra year of schooling in different countries, using different exams and testing different underlying populations. The height of each bar is equal to the estimate of $\beta_1 + \beta_2$ from the following equation $Z_i = \beta_0 + \beta_1 Grade_i + \beta_2 age_i + \beta_3 male_i + \varepsilon_i$. This is slightly different from the methodology used by Evans and Popova (2016). The graph also shows the 90% confidence interval of $\beta_1 + \beta_2$. For each data set we used a vertically linked 2LP IRT model to estimate comparable scores across grades.[3] The main message from this figure is: Reporting results in standard deviations can be misleading. What a standard

[3] The Global Reading Network (https://globalreadingnetwork.net) provided the EGRA/EGMA data. The Young Live data can be downloaded from the UK Data service webpage. Abhijeet Singh kindly provided the complementary files needed to vertically link the questions for Young Lives.

Table A.23: What do providers provide?, according to teachers

	(1) Control	(2) Difference	(3) Difference (F.E)
Teacher guides (or teacher manuals)	0.02	0.72^{***}	0.77^{***}
	(0.13)	(0.03)	(0.03)
Textbooks	0.03	0.85^{***}	0.87^{***}
	(0.17)	(0.02)	(0.03)
Copybooks	0.01	0.56^{***}	0.46^{***}
	(0.11)	(0.05)	(0.05)
Paper	0.01	0.68^{***}	0.69^{***}
	(0.11)	(0.04)	(0.04)
Teacher training	0.02	0.77^{***}	0.81^{***}
	(0.15)	(0.03)	(0.03)
School repairs	0.01	0.32^{***}	0.37^{***}
	(0.11)	(0.04)	(0.03)
Organization of community meetings	0.02	0.60^{***}	0.65^{***}
	(0.13)	(0.04)	(0.03)
Food programs	0.02	0.01	0.01
	(0.13)	(0.02)	(0.01)
Computers, tablets, electronics	0.01	0.44^{***}	0.58^{***}
	(0.11)	(0.06)	(0.05)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) group, as well as the difference between treatment and control (Column 2), and the difference taking into account the randomization design (i.e., including "'pair"' fixed effects) in Column 3. Standard errors are clustered at the school level. The sample is the original treatment and control allocation.N = 803. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

deviation means in practice (compared to business as usual) varies depending on the questions in the exam, the population tested, and the country.

A.9 Absolute learning levels

The test has some questions that are identical to those of other assessments, which allows us to compare absolute levels of learning: Two math questions taken

Table A.24: What did providers do during their last visit, according to teachers

	(1) Control	(2) Difference	(3) Difference (F.E)
Check attendance and collect records	0.10	0.40***	0.28***
	(0.30)	(0.06)	(0.06)
Observe teaching practices and give suggestions	0.13	0.50***	0.45***
	(0.34)	(0.06)	(0.06)
Provide/deliver educational materials	0.01	0.25***	0.22***
	(0.11)	(0.03)	(0.04)
Ask students questions to test learning	0.09	0.21***	0.10**
	(0.28)	(0.06)	(0.05)
Monitor other school-based government programs	0.01	0.07***	0.09***
	(0.11)	(0.02)	(0.03)
Meet with principal	0.30	0.11	0.08
	(0.46)	(0.08)	(0.08)
Meet with PTA committee	0.01	0.10***	0.10**
	(0.11)	(0.02)	(0.04)
Monitor health/sanitation issues	0.00	0.07***	0.06***
	(0.00)	(0.02)	(0.02)

This table presents the mean and standard error of the mean (in parentheses) for the control (Column 1) group, as well as the difference between treatment and control (Column 2), and the difference taking into account the randomization design (i.e., including ‘“pair”’ fixed effects) in Column 3. Standard errors are clustered at the school level. The sample is the original treatment and control allocation. N = 715. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

from TIMSS released items (M031317 and M031316), two reading comprehension questions taken PrePIRLS released items (L11L01C and L11L02M), and the number and word identification matrices used during the Liberia Teacher Training Program (LTTP) program evaluation in Liberia King et al. 2015.

Figure A.13 shows the average words per minute (wpm) and numbers per minute (npm) that students in different grades achieved at the 2013 LTTP program midline and at our own midline (for both treatment and control schools in both programs). Figures A.14 and A.15 show the results from 4th grade students (enrolled in 3rd grade in 2015/2016) in treatment and control schools in the TIMSS items, as well as the average for every country in 2011. Finally, Figure A.16 show the results from 4th grade students (enrolled in 3rd grade in 2015/2016) in treatment and control schools in the PrePIRLS items, as well as the average for every country in 2011.

Absolute learning levels are low. Despite the positive treatment effect of PSL,

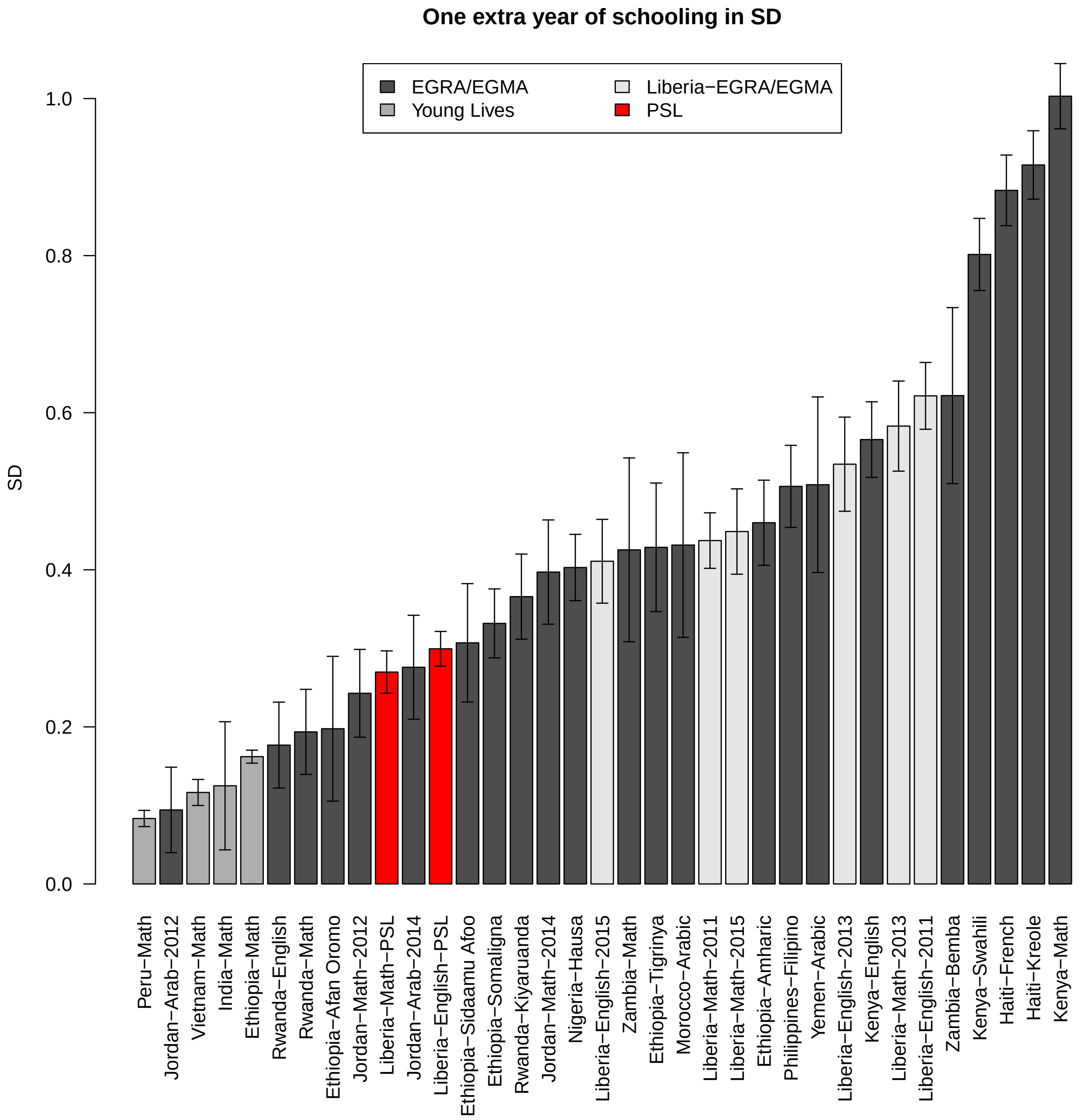


Figure A.12: International benchmark: How much do children learn per year?

students in treatment schools are still far behind their international peers. Based on results for the TIMSS or the PrePIRLS items, Liberia (both treatment and control schools) is at the very bottom of the ranking or close to it. This is especially worrisome in regard to English learning. Liberian students perform well below their peers in other countries, particularly when considering that PrePIRLS is specifically designed for countries where most children in the fourth grade are still developing fundamental reading skills (and thus, in most countries the PIRLS assessment is used).

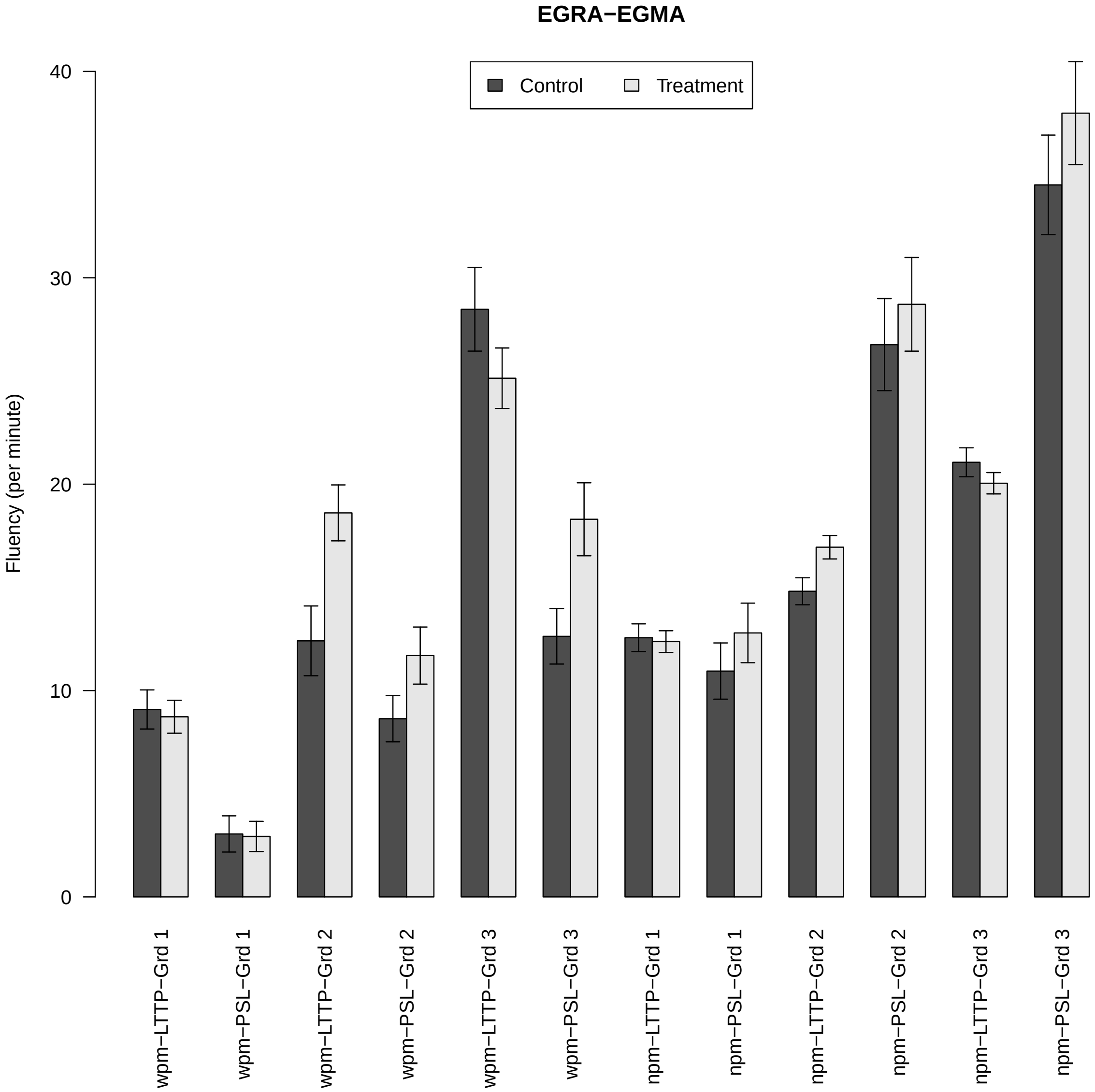


Figure A.13: PSL treatment effects on EGRA and EGMA vs. USAID's LTTP progra

Note: Figures show the average number of words per minute (wpm) and numbers per minute (npm) in the LTTP evaluation and the PSL evaluation for students in grades 1-3.

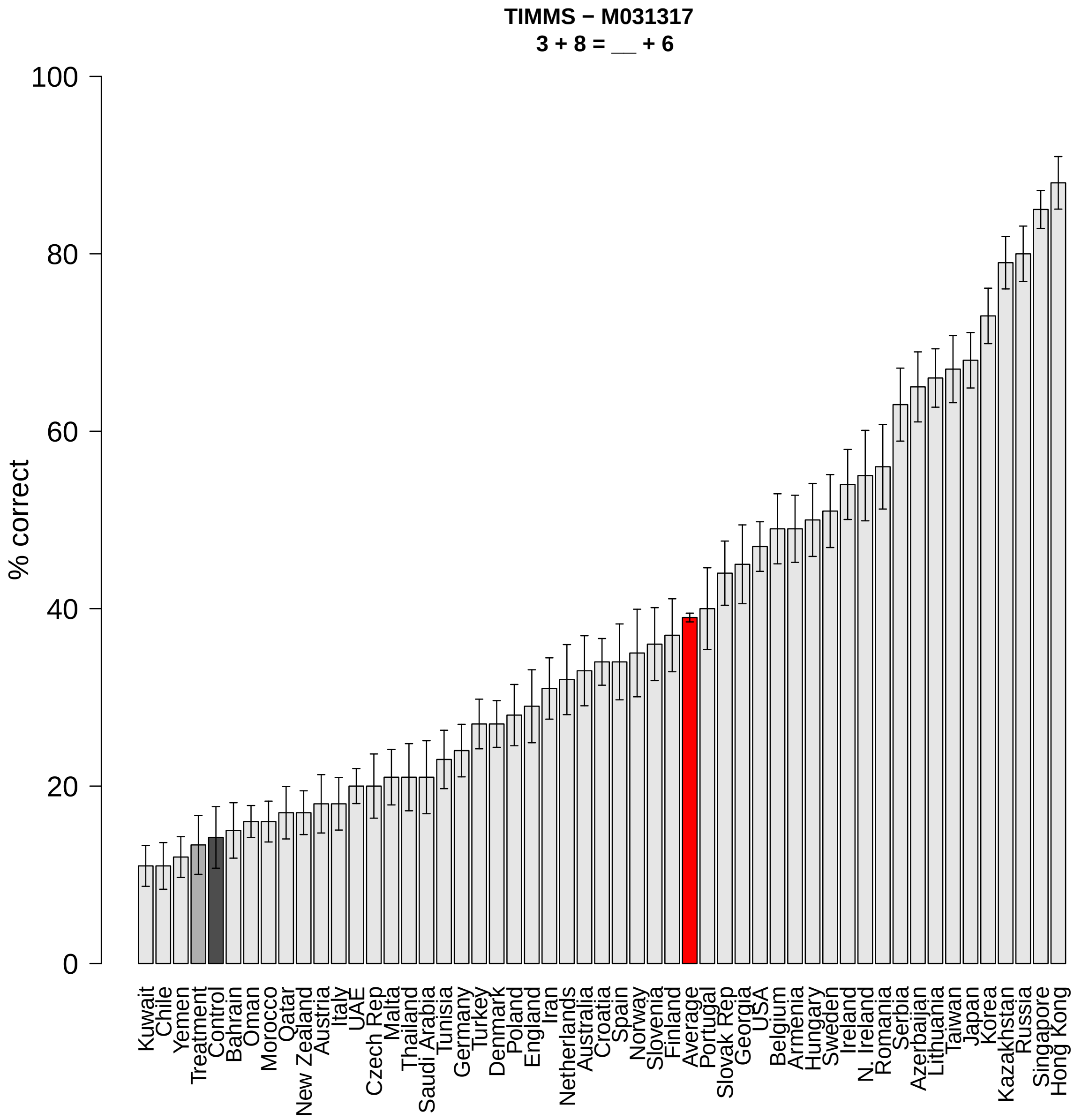


Figure A.14: International benchmark for mathematics proficiency (1 of 2)

Note: Figures show the proportion of students with correct responses to this question in the PSL evaluation (only students in grade 3 in 2015/2016), and in TIMSS assessments. This question is multiple-choice in TIMSS and open-ended in our assessment.

A.10 Comparisons across providers

The assignment of providers to schools was not random. Providers stated different preferences for locations and some volunteered to manage schools in more

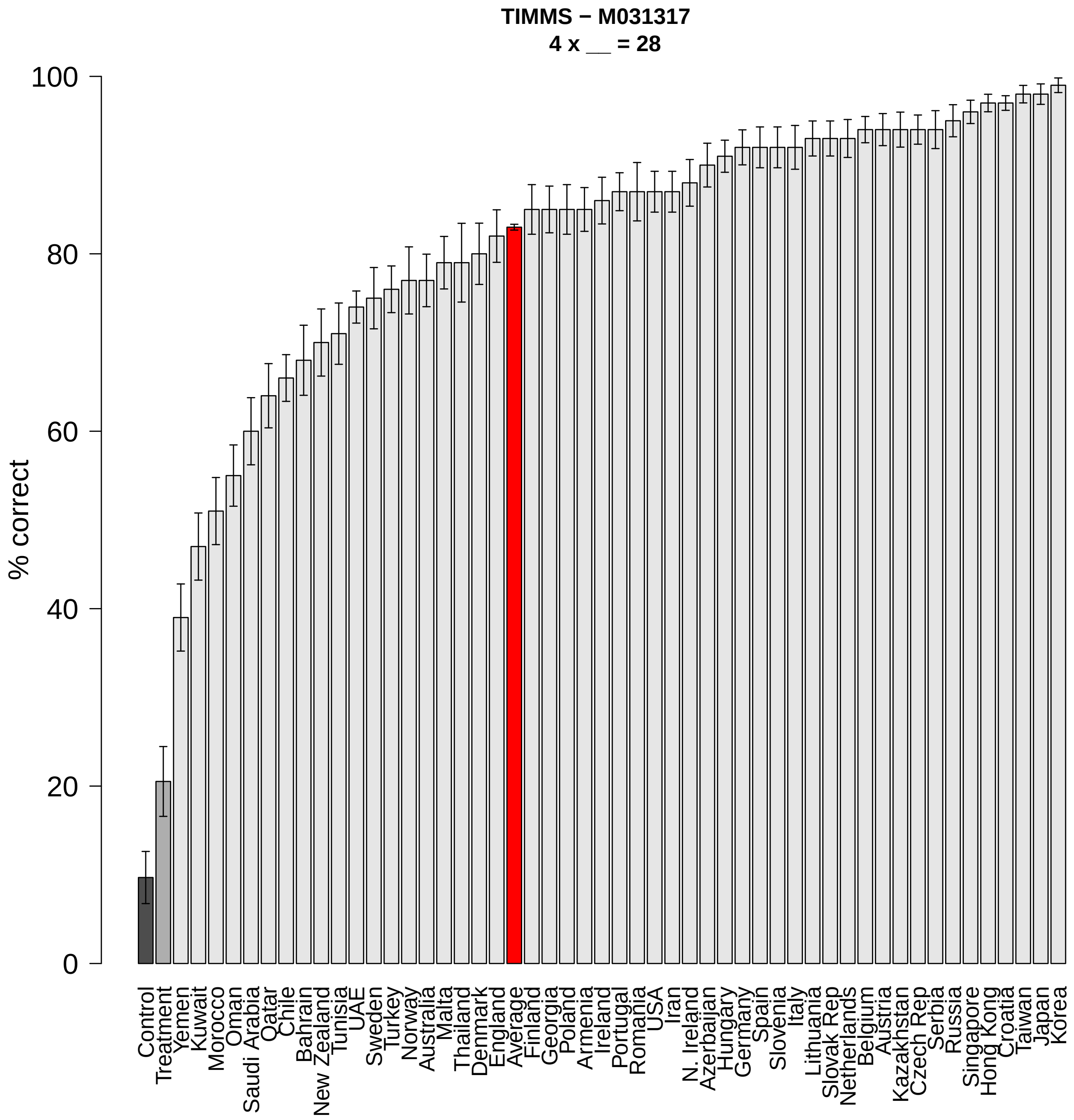


Figure A.15: International benchmark for mathematics proficiency (2 of 2)

Note: Figures show the proportion of students with correct responses to this question in the PSL evaluation (only students in grade 3 in 2015/2016), and in TIMSS assessments. This question is open-ended in TIMSS and in our assessment.

remote and marginalized areas. Thus, any heterogeneous effects by provider or by provider characteristics are not experimental. Figure A.25 shows the treatment and control schools allocated to each provider. Table A.17 shows the difference in school

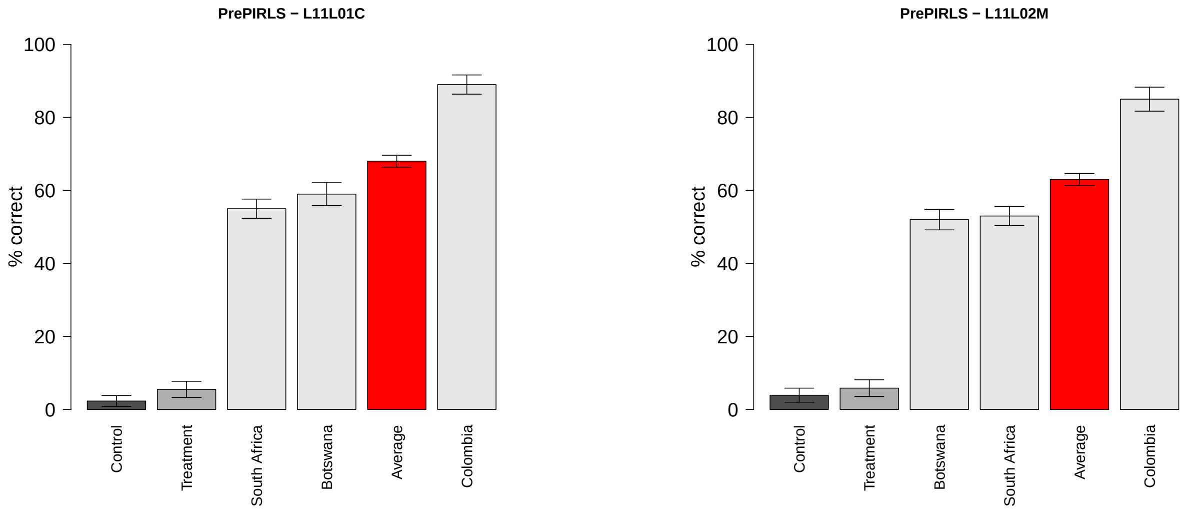


Figure A.16: International benchmark for reading proficiency
Note: Figures show the proportion of students with correct responses to this question in the PSL evaluation (only students in grade 3 in 2015/2016) and in PrePirls assessments. Question L11L01C is open-ended in TIMSS and in our assessment. Question L11L02M is multiple-choice in TIMSS and open-ended in our assessment.

characteristics (treatment and control) across providers.

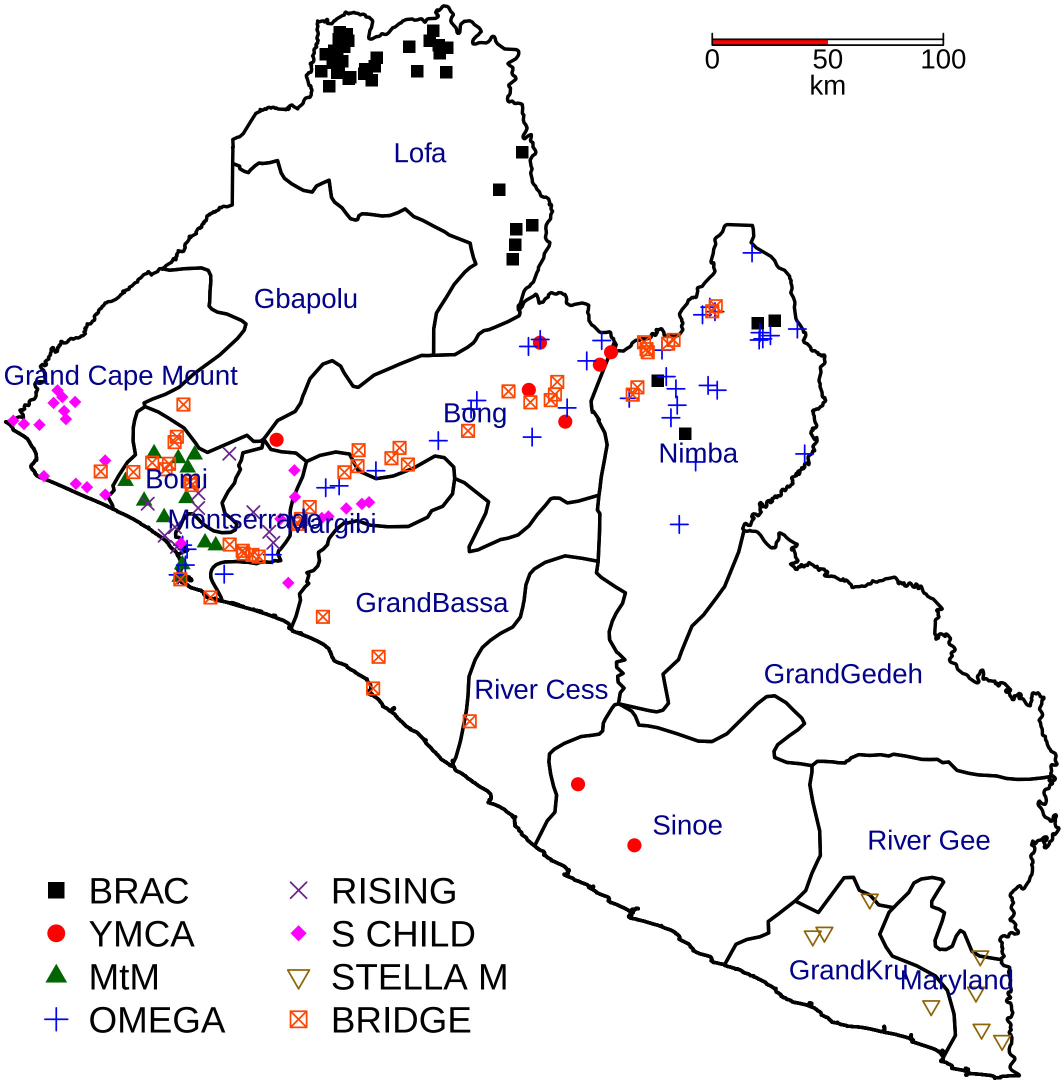


Figure A.17: Geographical distribution of providers

Table A.25: Pre-treatment EMIS characteristics of treatment schools by provider

	BRAC	BRIDGE	MtM	OMEGA	RISING	SCHILD	STELLAM	YMCA	Total
Students: ECE	126.14	178.50	106.78	158.37	123.67	154.86	115.17	115.43	146.94
	(12.18)	(18.27)	(11.04)	(9.55)	(18.21)	(11.62)	(13.80)	(21.66)	(6.04)
Students: Primary	152.20	225.08	140.33	115.14	120.00	109.36	99.00	110.43	148.28
	(11.72)	(35.58)	(43.47)	(7.96)	(14.47)	(7.57)	(16.13)	(20.35)	(9.68)
Students	278.34	403.58	247.11	273.51	243.67	264.23	214.17	225.86	295.22
	(19.59)	(39.60)	(46.23)	(13.21)	(26.78)	(14.53)	(29.01)	(32.47)	(11.97)
Classrooms per 100 students	0.97	1.28	2.16	0.56	1.90	1.11	0.00	1.45	1.07
	(0.26)	(0.20)	(0.95)	(0.20)	(0.66)	(0.33)	(0.00)	(0.66)	(0.12)
Teachers per 100 students	2.97	2.49	3.95	3.17	3.55	2.76	3.21	3.17	2.98
	(0.19)	(0.17)	(1.11)	(0.18)	(0.62)	(0.26)	(0.29)	(0.45)	(0.11)
Textbooks per 100 students	139.13	75.74	58.67	96.39	120.84	83.64	68.20	75.67	96.63
	(16.65)	(11.50)	(23.96)	(22.27)	(42.49)	(19.15)	(15.53)	(24.30)	(7.90)
Chairs per 100 students	6.19	25.42	38.68	15.56	34.82	23.20	15.49	41.69	20.33
	(2.23)	(3.30)	(11.89)	(2.94)	(9.86)	(7.27)	(11.59)	(16.75)	(2.04)
Food from Gov or NGO	0.03	0.39	0.67	0.31	0.78	0.64	0.67	0.00	0.36
	(0.03)	(0.08)	(0.17)	(0.08)	(0.15)	(0.10)	(0.21)	(0.00)	(0.04)
Solid building	0.26	0.61	0.33	0.14	0.67	0.41	0.00	0.71	0.37
	(0.07)	(0.08)	(0.17)	(0.06)	(0.17)	(0.11)	(0.00)	(0.18)	(0.04)
Water pump	0.31	0.64	0.56	0.71	0.89	0.73	0.83	0.71	0.62
	(0.08)	(0.08)	(0.18)	(0.08)	(0.11)	(0.10)	(0.17)	(0.18)	(0.04)
Latrine/toilet	0.78	0.87	0.81	0.88	0.89	0.91	0.93	0.86	0.86
	(0.07)	(0.06)	(0.13)	(0.05)	(0.08)	(0.06)	(0.07)	(0.14)	(0.03)
Observations	40	45	8	12	38	10	24	8	185

This table presents the mean and standard error of the mean (in parentheses) for several school characteristics across providers. The sample is the original treatment and control allocation. Source: EMIS data.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix B

Cross-Age Tutoring: Experimental Evidence from Kenya

Table B.1: Pupil and tutor test scores during T1DG16

	(1) English Tutoring	(2) Math Tutoring	(3) Difference	(4) Difference (F.E)
Panel A: Pupils				
English	0.000	-0.050	-0.047	-0.077
	(1.000)	(1.061)	(0.082)	(0.078)
Math	0.000	-0.060	-0.056	-0.086
	(1.000)	(1.060)	(0.085)	(0.087)
Science	0.000	-0.160	-0.164*	-0.180**
	(1.000)	(1.064)	(0.089)	(0.080)
S.S.	0.000	-0.130	-0.128	-0.147**
	(1.000)	(1.053)	(0.077)	(0.073)
Swahili	0.000	0.030	0.026	-0.001
	(1.000)	(1.053)	(0.078)	(0.076)
Panel C: Tutors				
English	0.000	0.040	0.041	0.027
	(0.999)	(1.030)	(0.050)	(0.048)
Math	0.000	0.050	0.046	0.030
	(0.999)	(0.999)	(0.050)	(0.044)
Science	0.000	0.030	0.027	0.011
	(0.999)	(1.051)	(0.044)	(0.040)
S.S.	0.000	0.050	0.051	0.043
	(0.999)	(1.027)	(0.047)	(0.044)
Swahili	0.000	-0.010	-0.010	-0.010
	(0.999)	(1.017)	(0.065)	(0.041)

Math, Language (English), Swahili, Science, and S.S. (Social Sciences) represent the standardized test scores (mean zero and standard deviation 1 in English tutoring schools).

Each row presents the mean for schools that receive English tutoring (Column 1), schools that receive math tutoring (Column 2), the difference between the two (Column 3), and the difference taking into account the randomization design (Column 4). In the first two columns the standard deviation is shown in parentheses, while in the third and fourth columns the standard error of the difference is in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.2: Heterogeneity: English test scores

	Tutee characteristics			Tutor characteristics			School characteristics		
	(1) Age	(2) Male	(3) Age joined Bridge	(4) Age	(5) Male	(6) Score in T3ET15	(7) PTR	(8) TTR	(9) Enrollment
Math tutoring × Covariate	0.030^{**}	-0.0055	0.022	0.035^{*}	-0.21	0.013	-0.000058	0.12^{***}	0.000036
	(0.015)	(0.029)	(0.014)	(0.019)	(0.21)	(0.034)	(0.0035)	(0.037)	(0.00045)
Observations	48597	48704	48597	48311	48311	39029	48704	48683	48704
Adjusted R^2	0.293	0.293	0.294	0.292	0.292	0.302	0.292	0.295	0.292

The independent variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does not include T2ET16 data. See Table B.6 for a version of this table that includes T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

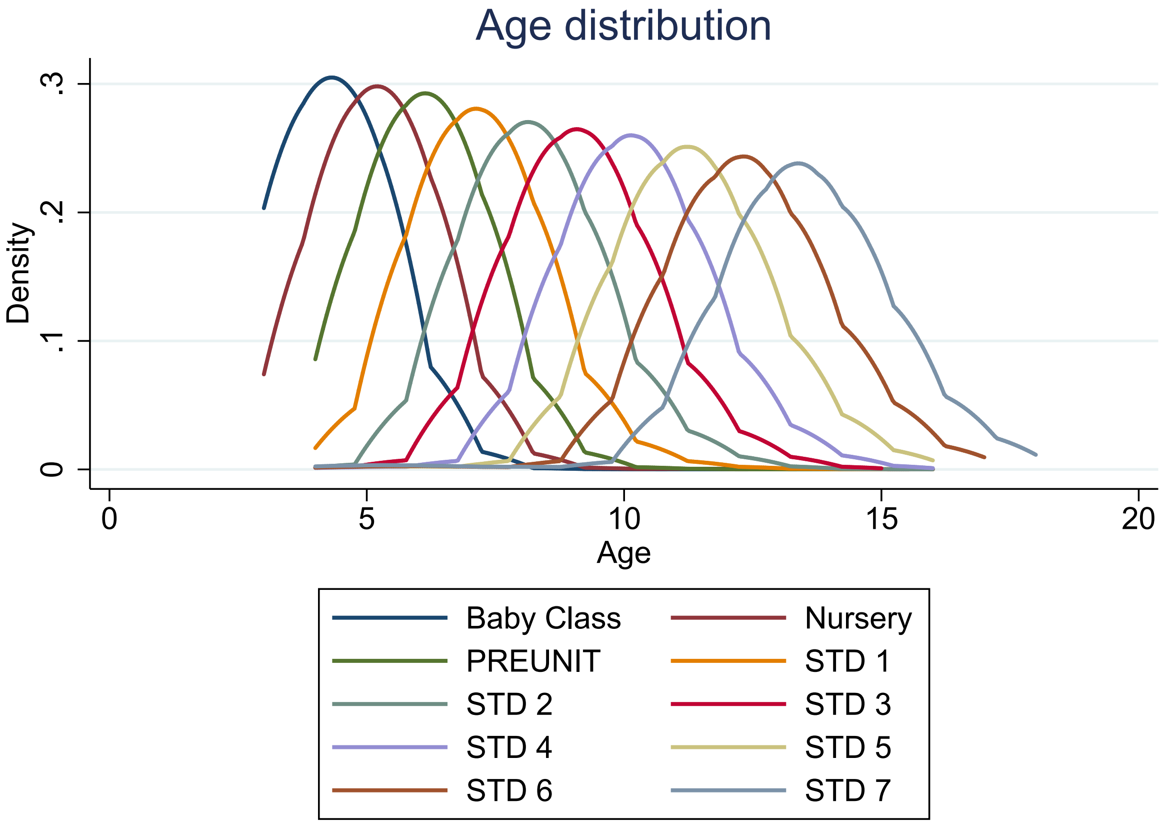


Figure B.1: Age distribution across grades

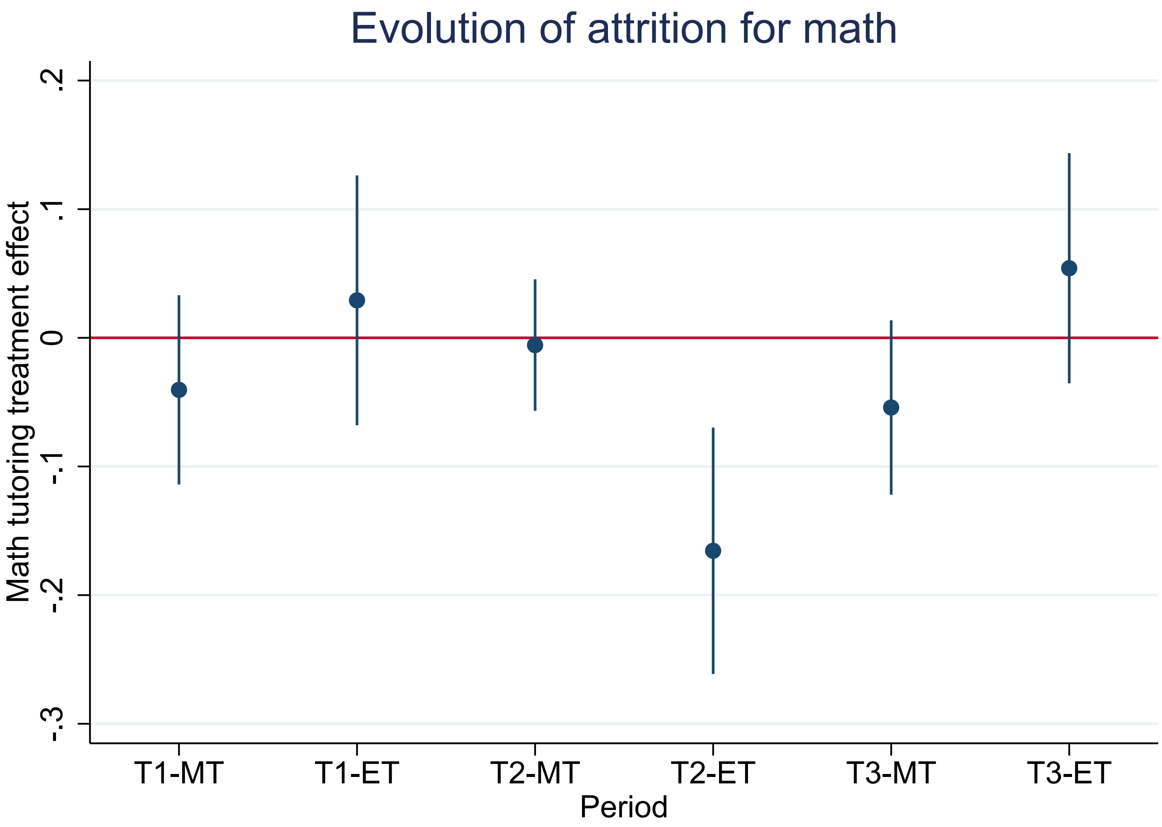


Figure B.2: Difference in testing rates across English and math tutoring in each period. Bars show 95% confidence intervals.

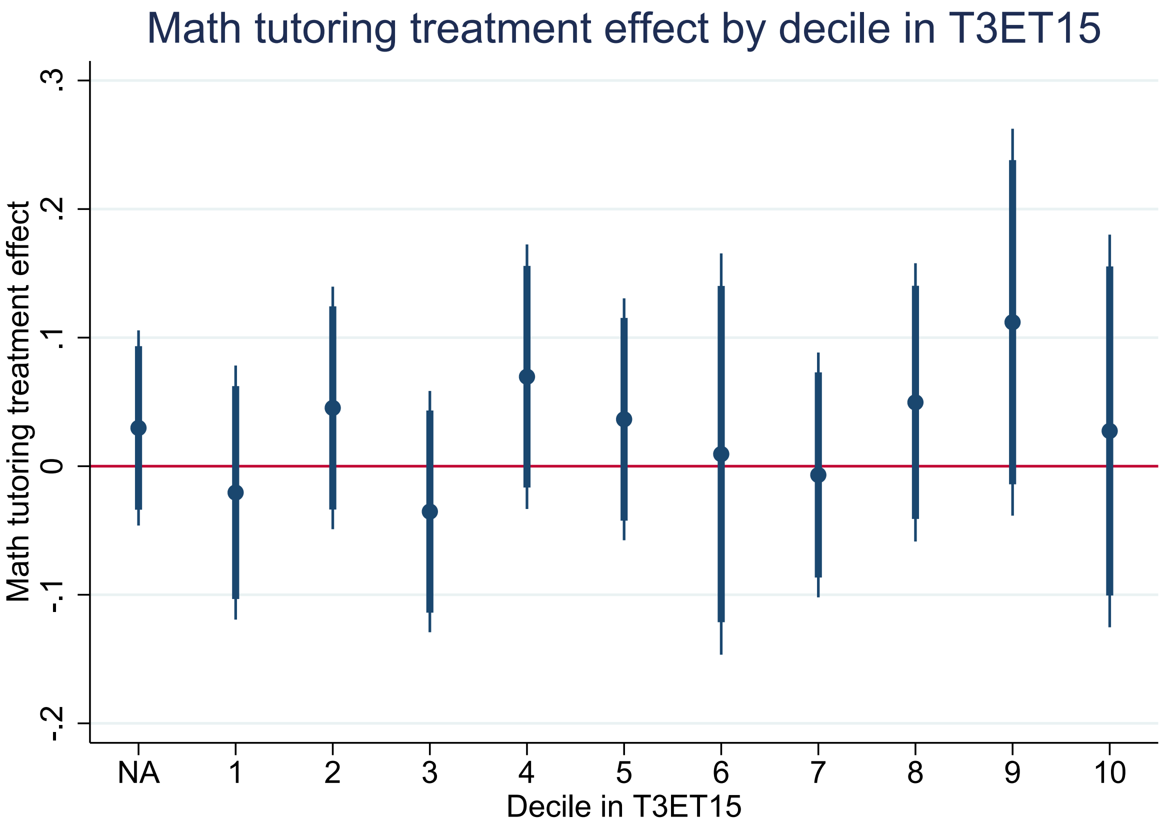


Figure B.3: Treatment effect of math tutoring on tutors' math test scores, relative to English tutoring, by ability decile in T3ET15. Bars represent 90% and 95% confidence intervals (thick lines and thin lines, respectively).

Table B.3: Differential attrition between treatment and control students

	Math	English	Swahili
Math tutoring	-0.031	-0.0026	-0.0067
	(0.023)	(0.024)	(0.028)
Mean English	0.61	0.58	0.59
N. of obs.	97742	97756	66149
Number of schools	187	187	187

This table shows the differential attrition between students in math tutoring schools compared to students in English tutoring schools. The estimation data set does include T2ET16 data. Clustered standard errors, by school, in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.4: Effect on tutees' test scores

	Math	English	Swahili
Math tutoring	0.057	-0.0038	0.017
	(0.034)	(0.034)	(0.048)
N. of obs.	56834	55937	37835
Number of schools	187	187	186

The independent variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does include T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, ***

Table B.5: Effect on tutees' test scores

	Math	English	Swahili
Math tutoring	0.038	-0.0097	-0.014
	(0.031)	(0.035)	(0.036)
N. of obs.	55066	53222	52560
Number of schools	187	187	187

The independent variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does include T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, ***

B.1 With T2ET16

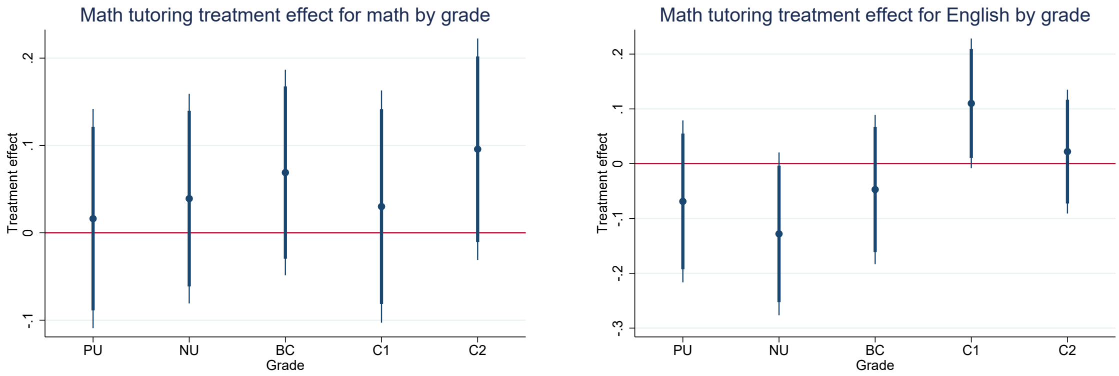


Figure B.4: Treatment effect of math tutoring, relative to English tutoring, on math (left panel) and English (right panel) test scores by grade. Bars represent 90% and 95% confidence intervals (thick lines and thin lines, respectively).

Table B.6: Heterogeneity

	Tutee characteristics			Tutor characteristics			School characteristics		
	(1) Age	(2) Male	(3) Age joined Bridge	(4) Age	(5) Male	(6) Score in T3ET15	(7) PTR	(8) TTR	(9) Enrollment
Panel A: Math									
Math tutoring × Covariate	0.018	-0.030	0.020	0.013	-0.11	-0.016	0.0028	0.040	0.00035
	(0.014)	(0.027)	(0.012)	(0.019)	(0.21)	(0.032)	(0.0039)	(0.028)	(0.00050)
Observations	57258	57390	57258	56966	56966	46346	57390	57363	57390
Adjusted R^2	0.222	0.220	0.222	0.221	0.221	0.232	0.221	0.221	0.221
Panel B: English									
Math tutoring × Covariate	0.030**	-0.0058	0.022	0.034*	-0.20	0.015	0.00089	0.13***	0.00013
	(0.015)	(0.029)	(0.013)	(0.018)	(0.21)	(0.033)	(0.0035)	(0.034)	(0.00044)
Observations	56363	56490	56363	56064	56064	45513	56490	56463	56490
Adjusted R^2	0.293	0.293	0.293	0.292	0.292	0.303	0.292	0.295	0.292

The independent variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does include T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, ***

Appendix C

Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania

Table C.1: Effect of grants, incentives, and their interaction on school expenditure

	(1)	(2)	(3)	(4)	(5)	(6)
				TZS per student		
	Total	Admin.	Student	Teaching Aids	Teacher	Construction
Panel A: Year 1						
Grants (α_1)	-2,407.92***	-189.74	198.39	-2,578.23***	-22.80	184.46
	(813.88)	(446.25)	(121.11)	(409.48)	(74.54)	(428.89)
Incentives (α_2)	-10.05	-265.49	29.90	-142.81	3.72	364.62
	(642.21)	(215.47)	(63.55)	(244.66)	(81.62)	(494.07)
Combo (α_3)	-1,412.22	-16.74	138.47	-1,325.72**	-13.10	-195.13
	(932.79)	(469.02)	(111.26)	(576.26)	(78.46)	(327.41)
N. of obs.	350	350	350	350	350	350
Mean control	5,959.67	2,083.48	274.83	2,745.51	180.83	675.02
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	1,005.76	438.49	-89.82	1,395.32	5.98	-744.21
p-value ($H_0 : \alpha_4 = 0$)	0.44	0.50	0.60	0.056	0.96	0.24
$\alpha_3 - \alpha_1$	995.70	173.00	-59.92	1,252.51	9.70	-379.59
p-value ($H_0 : \alpha_3 - \alpha_1 = 0$)	0.39	0.78	0.71	0.072	0.90	0.35
Panel B: Year 2						
Grants (α_1)	-2,317.74**	27.08	-1,267.61	-1,115.91***	35.26	3.45
	(1,096.16)	(514.15)	(900.20)	(210.24)	(77.55)	(294.17)
Incentives (α_2)	-1,166.46	-124.02	-813.77	-265.70**	-46.38	83.41
	(818.24)	(163.17)	(733.05)	(133.45)	(37.81)	(299.89)
Combo (α_3)	-1,896.28**	-112.95	-722.99	-666.12***	-7.45	-386.77**
	(928.05)	(193.77)	(876.74)	(181.96)	(57.10)	(189.52)
N. of obs.	349	349	349	349	349	349
Mean control	4,524.03	1,422.30	1,276.60	1,314.33	96.58	414.21
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	1,587.92	-16.01	1,358.40	715.49	3.67	-473.63
p-value ($H_0 : \alpha_4 = 0$)	0.15	0.97	0.11	0.010	0.97	0.23
$\alpha_3 - \alpha_1$	421.46	-140.03	544.62	449.79	-42.70	-390.22
p-value ($H_0 : \alpha_3 - \alpha_1 = 0$)	0.56	0.78	0.10	0.064	0.64	0.13

Results from estimating Equation 3.2 for expenditure per child. *Admin:* Administrative cost (including staff wages), rent and utilities, and general maintenance and repairs. *Student:* Food, scholarships and materials (notebooks, pens, etc.) *Teaching aids:* Classroom furnishings, textbooks, maps, charts, blackboards, practice exams, etc. *Teachers:* Salaries, bonuses and teacher training. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.2: Effect of grants, incentives, and their interaction on school income from different sources

	(1) Total	(2) Government CG	(3) Government Other	(4) Local Government	(5) NGOs	(6) Parents	(7) Other
Panel A: Year 1							
Grants (α_1)	468.30	612.70	195.24	-199.96	2.24	-217.53	-25.45
	(581.64)	(429.79)	(195.12)	(235.14)	(12.78)	(212.56)	(152.94)
Incentives (α_2)	-145.32	57.48	192.96	127.45	-7.70	-515.51**	94.65
	(618.82)	(343.24)	(209.40)	(337.31)	(5.64)	(202.47)	(207.25)
Combo (α_3)	36.92	227.78	457.09	-348.82	-8.63	-255.64	-125.01
	(623.61)	(381.56)	(348.02)	(222.74)	(6.72)	(213.15)	(106.82)
N. of obs.	350	342	349	349	349	339	348
Mean control	6,095.82	4,569.53	40.07	366.57	7.82	1,124.25	158.28
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-286.06	-442.40	68.89	-276.30	-3.17	477.39	-194.21
p-value ($H_0 : \alpha_4 = 0$)	0.75	0.47	0.88	0.43	0.79	0.069	0.42
$\alpha_3 - \alpha_1$	-431.38	-384.92	261.85	-148.85	-10.86	-38.12	-99.56
p-value ($H_0 : \alpha_3 - \alpha_1 = 0$)	0.53	0.44	0.52	0.26	0.33	0.84	0.42
Panel B: Year 2							
Grants (α_1)	-2,011.28**	-744.66*	-188.31	-6.68	177.58	-1,178.21	-51.61
	(983.53)	(415.19)	(150.09)	(19.54)	(142.44)	(902.89)	(35.07)
Incentives (α_2)	-1,126.86	-486.32	-121.16	45.30	4.01	-564.55	7.92
	(825.42)	(345.35)	(134.68)	(57.43)	(32.46)	(756.05)	(49.34)
Combo (α_3)	-910.15	-699.61*	385.51	-17.84	115.62	-817.95	148.87
	(1,116.60)	(414.84)	(553.78)	(19.27)	(134.51)	(895.09)	(97.86)
N. of obs.	349	346	349	349	349	349	349
Mean control	4,759.24	2,437.46	168.62	18.88	13.64	2,046.92	73.72
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	2,227.99	531.37	694.98	-56.46	-65.97	924.81	192.57
p-value ($H_0 : \alpha_4 = 0$)	0.049	0.24	0.22	0.33	0.74	0.29	0.092
$\alpha_3 - \alpha_1$	1,101.13	45.04	573.82	-11.16	-61.96	360.26	200.49
p-value ($H_0 : \alpha_3 - \alpha_1 = 0$)	0.15	0.89	0.31	0.36	0.76	0.33	0.040

Results from estimating Equation 3.2 for income per child from different sources. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.3: Effect of grants, incentives, and their interaction on household expenditure

	(1) Total expenditure	(2) Fees	(3) Textbooks	(4) Other books	(5) Supplies	(6) Uniforms	(7) Tutoring	(8) Transport	(9) Food	(10) Other
Panel A: Year 1										
Grants (α_1)	-1,014.96	-145.37	-33.05	-27.04	363.57	-334.43	-1,061.87	-143.55	542.56	-39.38
	(1,579.79)	(632.75)	(84.42)	(44.32)	(270.40)	(663.91)	(845.69)	(150.10)	(1,140.43)	(219.47)
Incentives (α_2)	-977.78	-11.27	7.73	-3.96	180.38	-287.47	-502.75	303.21	-240.27	-144.49
	(1,294.84)	(451.70)	(101.54)	(50.20)	(229.47)	(636.92)	(840.70)	(306.75)	(1,043.16)	(248.75)
Combo (α_3)	-1,382.23	-526.39	135.08	23.41	-52.45	-240.56	-708.35	86.01	-41.01	-210.18
	(1,153.27)	(391.13)	(82.78)	(56.94)	(253.33)	(640.66)	(874.28)	(270.39)	(779.80)	(217.14)
N. of obs.	350	350	350	350	350	350	350	350	350	350
Mean control	28,821.01	3,247.03	273.35	139.44	5,004.53	11,362.63	4,760.02	235.37	4,689.80	1,549.91
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	610.51	-369.75	160.41	54.40	-596.40	381.33	856.27	-73.66	-343.30	-26.31
p-value ($H_0 : \alpha_4 = 0$)	0.77	0.64	0.26	0.47	0.13	0.71	0.51	0.85	0.82	0.94
$\alpha_3 - \alpha_1$	-367.27	-381.02	168.14	50.44	-416.02	93.86	353.52	229.56	-583.57	-170.80
p-value ($H_0 : \alpha_3 - \alpha_1 = 0$)	0.83	0.58	0.084	0.36	0.20	0.91	0.72	0.38	0.62	0.45
Panel B: Year 2										
Grants (α_1)	-2,164.18*	-919.53*	-210.52**	46.71	-105.93	-427.54	-439.50	-70.46	-1,341.18**	-342.89*
	(1,201.53)	(550.69)	(100.77)	(65.39)	(246.27)	(638.46)	(693.04)	(301.90)	(624.04)	(204.00)
Incentives (α_2)	235.40	-147.95	-96.95	48.26	410.99	217.61	570.57	-445.89	-1,152.35**	-73.60
	(1,214.01)	(765.96)	(121.33)	(63.20)	(261.44)	(608.93)	(799.43)	(329.30)	(584.26)	(211.05)
Combo (α_3)	-75.59	-297.84	-145.61	85.07	175.34	320.83	-647.17	-420.25	-148.02	-101.52
	(1,151.27)	(605.34)	(92.38)	(61.37)	(253.04)	(589.29)	(749.68)	(316.05)	(872.65)	(184.35)
N. of obs.	350	350	350	350	350	350	350	350	350	350
Mean control	27,362.34	2,782.55	442.72	137.02	4,178.28	14,437.64	3,252.00	468.80	3,565.93	2,003.89
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	1,853.19	769.64	161.86	-9.90	-129.72	530.76	-778.24	96.10	2,345.52	314.98
p-value ($H_0 : \alpha_4 = 0$)	0.30	0.38	0.29	0.92	0.73	0.57	0.49	0.78	0.031	0.28
$\alpha_3 - \alpha_1$	2,088.59	621.69	64.91	38.37	281.27	748.37	-207.67	-349.79	1,193.17	241.38
p-value ($H_0 : \alpha_3 - \alpha_1 = 0$)	0.11	0.12	0.49	0.62	0.31	0.29	0.80	0.018	0.18	0.23

Results from estimating Equation 3.2 for household expenditure per child. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.4: Effect of grants, incentives, and their interaction on test scores without controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Swahili	English	Combined (PCA)	Math	Swahili	English	Combined (PCA)
Grants (α_1)	-0.05	-0.01	-0.03	-0.03	0.01	0.00	0.03	0.02
	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.05)	(0.06)	(0.05)
Incentives (α_2)	0.06	0.06	0.06	0.07^{*}	0.08^{*}	0.01	0.00	0.04
	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)	(0.04)
Combo (α_3)	0.10^{**}	0.11^{***}	0.10^{**}	0.12^{***}	0.21^{***}	0.22^{***}	0.19^{***}	0.24^{***}
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	0.096	0.059	0.065	0.085	0.12	0.20^{***}	0.16^{*}	0.18^{**}
p-value ($H_0 : \alpha_4 = 0$)	0.12	0.32	0.33	0.16	0.10	0.0068	0.054	0.011

Results from estimating Equation 3.3 for different subjects at both follow-ups. Control variables only include student characteristics (age, gender, grade and lag test scores). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.5: Effect of grants, incentives, and their interaction on test scores on a fix cohort of students

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Swahili	English	Combined (PCA)	Math	Swahili	English	Combined (PCA)
Grants (α_1)	-0.02	-0.04	-0.00	-0.02	0.06	0.01	0.03	0.04
	(0.05)	(0.05)	(0.05)	(0.04)	(0.06)	(0.06)	(0.06)	(0.05)
Incentives (α_2)	0.02	0.02	0.09*	0.05	0.09*	-0.02	0.01	0.03
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Combo (α_3)	0.12**	0.10**	0.13**	0.14***	0.25***	0.21***	0.18***	0.24***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.06)	(0.04)
N. of obs.	6,043	6,043	6,043	6,043	6,343	6,343	6,343	6,343
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	0.11	0.12*	0.046	0.11	0.096	0.21***	0.14	0.17**
p-value ($H_0 : \alpha_4 = 0$)	0.12	0.090	0.55	0.12	0.21	0.0081	0.12	0.026

Results from estimating Equation 3.3 for different subjects at both follow-ups. Sample only includes students treated over the two-year period (i.e., students in grade 1 and grade 2 at baseline 2013). Control variables include only student characteristics (age, gender, grade and lag test scores). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.6: Effect of incentives and the combination on test scores: low- and high-stakes exams

	(1)	(2)	(3)	(4)
	Year 2			
	Math	Swahili	English	Combined (PCA)
Panel A: Low-stakes				
Incentives (α_2)	0.08	-0.00	-0.03	0.02
	(0.06)	(0.06)	(0.06)	(0.05)
Combo (α_3)	0.21^{***}	0.20^{***}	0.15^{**}	0.21^{***}
	(0.06)	(0.06)	(0.07)	(0.06)
N. of obs.	4,860	4,860	4,860	4,860
Interaction $(\alpha_4) = \alpha_3 - \alpha_2$	0.13^{***}	0.20^{***}	0.18^{***}	0.19^{***}
p-value ($H_0 : \alpha_4 = 0$)	0.0070	0.00015	0.0031	0.000057
Panel B: High-stakes				
Incentives (β_2)	0.17^{***}	0.12^{**}	0.12^{**}	0.21^{***}
	(0.05)	(0.05)	(0.05)	(0.07)
Combo (β_3)	0.25^{***}	0.23^{***}	0.22^{***}	0.36^{***}
	(0.05)	(0.06)	(0.06)	(0.08)
N. of obs.	46,886	46,882	46,882	46,882
$\beta_4 := \beta_3 - \beta_2$	0.081^{**}	0.11^{**}	0.099^{*}	0.15^{**}
p-value ($\beta_4 = 0$)	0.046	0.012	0.060	0.015
Panel C: High-stakes – Low-stakes				
$\beta_2 - \alpha_2$	0.085	0.12	0.15	0.19
p-value($\beta_2 - \alpha_2 = 0$)	0.13	0.023	0.024	0.0020
$\beta_3 - \alpha_3$	0.039	0.034	0.074	0.15
p-value($\beta_3 - \alpha_3 = 0$)	0.46	0.53	0.28	0.013
$\beta_4 - \alpha_4$	-0.045	-0.084	-0.079	-0.037
p-value($\beta_4 - \alpha_4 = 0$)	0.40	0.065	0.20	0.53

Results from estimating Equation 3.3 for different subjects. This sample only includes control schools on which the high-stakes exam was conducted. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.7: Spillovers into other grades and subjects, including test takers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Science		Grade 7 PSLE 2013			Grade 7 PSLE 2014		
	Year 1	Year 2	Pass	Score	Test takers	Pass	Score	Test takers
Grants (α_1)	0.02	-0.04	-0.03	-0.03	4.12	-0.02	-0.05	3.27
	(0.05)	(0.06)	(0.03)	(0.05)	(5.01)	(0.03)	(0.05)	(4.89)
Incentives (α_2)	0.01	-0.01	-0.02	-0.02	6.32	-0.00	-0.02	4.35
	(0.05)	(0.05)	(0.03)	(0.04)	(5.02)	(0.03)	(0.05)	(4.89)
Combo (α_3)	0.09	0.09*	0.02	0.05	7.88	0.01	0.04	7.56
	(0.05)	(0.05)	(0.03)	(0.05)	(5.07)	(0.03)	(0.05)	(4.99)
N. of obs.	9,142	9,439	26,836	26,836	346	25,162	25,162	345
Mean control group			0.52	2.60	73.8	0.57	2.70	69.8
$\alpha_4 = \alpha_3 - \alpha_2 - \alpha_1$	0.058	0.13*	0.066	0.10	-2.56	0.039	0.11	-0.060
p-value ($H_0 : \alpha_4 = 0$)	0.48	0.096	0.11	0.11	0.74	0.35	0.12	0.99

Columns (1) and (2) show estimates of Equation 3.3 for science in focal grades (Grd 1 - Grd 3) using data for both follow-ups, and therefore coefficients represent the average treatment effect across both years. Columns (3)-(6) use data from the national exit examination as dependent variables: pass rates, average test scores, and number of test takers. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.8: Lee bounds on PSLE scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Science				Grade 7 PSLE 2013			
	Pass	Pass (Lee)	Score	Score (Lee)	Pass	Pass (Lee)	Score	Score (Lee bound)
Grants (α_1)	-0.03	-0.02	-0.03	-0.03	-0.02	-0.01	-0.05	-0.04
	(0.03)	(0.03)	(0.05)	(0.05)	(0.03)	(0.03)	(0.05)	(0.05)
Incentives (α_2)	-0.02	-0.02	-0.02	-0.02	-0.00	-0.00	-0.02	-0.02
	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.05)
Combo (α_3)	0.02	0.02	0.05	0.05	0.01	0.01	0.04	0.03
	(0.03)	(0.03)	(0.05)	(0.05)	(0.03)	(0.03)	(0.05)	(0.05)
N. of obs.	26,836	24,137	26,836	24,137	25,162	21,844	25,162	21,844
Mean control group	0.52	0.53	2.60	2.61	0.57	0.58	2.70	2.71
$\alpha_4 = \alpha_3 - \alpha_2 - \alpha_1$	0.066	0.065	0.10	0.097	0.039	0.028	0.11	0.096
p-value ($H_0 : \alpha_4 = 0$)	0.11	0.12	0.11	0.14	0.35	0.51	0.12	0.18

Odd-numbered columns show the results without adjusting for number of test takes. Even-numbered columns trim the data to drop the left-tail in each treatment arm so that the average number of test takes per school is the same across treatment arms. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.9: Teachers' behavioral responses: tutoring, tests, and remedial teaching

	(1)	(2)	(3)	(4)	(5)
		Self reported			
	Attendance	Tests	Tutoring	Remedial	Inputs
Panel A: Year 1					
Grants (α_1)	0.04	-0.10	0.01	-0.04	0.12***
	(0.04)	(0.88)	(0.03)	(0.03)	(0.03)
Incentives (α_2)	-0.00	2.95***	0.04	0.00	0.00
	(0.04)	(0.99)	(0.03)	(0.03)	(0.04)
Combo (α_3)	0.05	-0.11	0.03	0.06*	0.13***
	(0.03)	(0.93)	(0.03)	(0.03)	(0.03)
N. of obs.	1,007	999	1,007	1,007	1,007
Mean of dep. var.	0.80	9.52	0.11	0.88	0.91
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	0.015	-2.96	-0.023	0.097	0.0052
p-value ($H_0 : \alpha_4 = 0$)	0.79	0.024**	0.63	0.050**	0.91
Panel B: Year 2					
Grants (α_1)	0.05	-0.71	0.03	-0.02	0.04**
	(0.05)	(1.31)	(0.03)	(0.06)	(0.02)
Incentives (α_2)	-0.02	-0.16	0.03	-0.11*	-0.01
	(0.05)	(0.95)	(0.03)	(0.06)	(0.03)
Combo (α_3)	-0.01	-0.39	0.09***	0.05	0.03
	(0.05)	(0.93)	(0.03)	(0.05)	(0.02)
N. of obs.	858	854	858	858	858
Mean of dep. var.	0.74	9.77	0.073	0.79	0.95
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-0.036	0.48	0.032	0.18	0.0020
p-value ($H_0 : \alpha_4 = 0$)	0.61	0.76	0.56	0.022**	0.95

Results from estimating any treatment effects on teacher behavior. All data is self-reported. Column (1) has the number of tests per period as the dependent variable. Column (2) has a dummy variable that indicates whether the teacher provided any extra tutoring to students as the dependent variable. Column (3) uses a dummy variable equal to one if teacher indicates teaching inputs are "above average" as the dependent variable. All regressions are done including data for both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.10: Teachers' behvaioral responses: Time use

	Preparing class (mins)	Teaching (mins)	Extra classes (mins)	Socializing (mins)	Time at school (hrs)
Panel A: Year 1					
Grants (α_1)	-1.75	2.28	3.35	-0.66	0.08
	(3.39)	(6.04)	(3.17)	(1.98)	(0.14)
Incentives (α_2)	-0.11	-2.69	4.63	2.61	0.02
	(2.92)	(5.95)	(3.46)	(2.30)	(0.13)
Combo (α_3)	-2.53	-3.12	5.69*	2.43	-0.05
	(3.04)	(6.42)	(3.15)	(2.37)	(0.14)
N. of obs.	1,056	1,056	1,056	1,056	1,056
Mean of dep. var.	44.7	151.1	33.4	35.2	7.72
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-0.67	-2.71	-2.29	0.48	-0.14
p-value ($H_0 : \alpha_4 = 0$)	0.89	0.77	0.65	0.89	0.51
Panel B: Year 2					
Grants (α_1)	2.75	-0.54	2.35	1.62	0.09
	(3.40)	(8.00)	(4.04)	(3.55)	(0.17)
Incentives (α_2)	-0.33	-15.40*	1.04	3.32	0.15
	(3.02)	(8.50)	(4.01)	(4.95)	(0.16)
Combo (α_3)	-0.69	8.99	3.71	4.18	0.05
	(2.95)	(7.57)	(3.54)	(3.63)	(0.16)
N. of obs.	857	857	857	857	857
Mean of dep. var.	43.3	155.4	20.1	39.1	7.63
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-3.12	24.9	0.32	-0.75	-0.19
p-value ($H_0 : \alpha_4 = 0$)	0.51	0.041**	0.96	0.91	0.44

Results from estimating any treatment effects on teacher time use. All data is self-reported. Column (1) estimates the effect on the time (in minutes) spent preparing class, Column (2) on the time (in minutes) spent teaching regular classes, Column (3) on the time (in minutes) spent teaching extra classes, Column (4) on the time (in minutes) spent socializing, and Column (5) on the total number of hours spent at the school. All regressions are done including data for both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.11: Number of high-stakes test takers

	(1) Test Takers
Incentives (β_2)	23.22*** (5.23)
Combo (β_3)	26.88*** (5.23)
N. of obs.	540
Mean control group	67.34
$\alpha_3 = \alpha_2 - \alpha_1$	3.66
p-value($\alpha_3 = 0$)	0.41

The independent variable is the number of test takers during the high-stakes exam at the end of the second year. Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.12: Lee bounds for high-stakes exams

	(1) Math	(2) Swahili	(3) English	(4) Combined (PCA)
Incentives (β_2)	0.17***	0.12**	0.12**	0.21***
	(0.05)	(0.05)	(0.05)	(0.07)
Combo (β_3)	0.25***	0.23***	0.22***	0.36***
	(0.05)	(0.06)	(0.06)	(0.08)
N. of obs.	46,886	46,882	46,882	46,882
$\beta_4 = \beta_3 - \beta_2$	0.081**	0.11**	0.099*	0.15**
p-value ($H_0 : \beta_4 = 0$)	0.046	0.012	0.060	0.015
Lower 95% CI (β_2)	0.047	-0.0072	-0.0041	0.041
Higher 95% CI (β_2)	0.29	0.25	0.24	0.37
Lower 95% CI (β_3)	0.12	0.097	0.080	0.18
Higher 95% CI (β_3)	0.37	0.36	0.35	0.53
Lower 95% CI (β_4)	-0.0024	0.021	-0.0082	0.024
Higher 95% CI (β_4)	0.16	0.20	0.21	0.27

The independent variable is the standardized test score for different subjects. For each subject we present Lee (2009) bounds for all the treatment estimates (i.e., trimming the left/right tail of the distribution in Incentive and Combination schools so that the number of test takes is the same as the number in control schools). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Bibliography

Abeberese, Ama Baafra, Todd J Kumler, and Leigh L Linden. 2014. "Improving reading skills by encouraging children to read in school: A randomized evaluation of the Sa Aklat Sisikat reading program in the Philippines". *Journal of Human Resources* 49 (3): 611–633.

Agor, Weston H. 1989. "Intuition & Strategic Planning: How Organizations Can Make". *The Futurist* 23 (6): 20.

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations". *Journal of Economic Perspectives* 19 (1): 9–32.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation". *The Quarterly Journal of Economics* 130 (3): 1117. doi:10.1093/qje/qjv015. +.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". *Journal of the American Statistical Association* 103 (484): 1481–1495. doi:10.1198/016214508000000841. eprint: http://dx.doi.org/10.1198/016214508000000841. http://dx.doi.org/10.1198/016214508000000841.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets". *American Economic Review* 107 (6): 1535–63. doi:10.1257/aer.20140774.

Andrabi, Tahir, Natalie Bau, Jishnu Das, and Asim Ijaz Khwaja. 2010. "Are bad public schools public "bads"? Test scores and civic values in public and private schools". Mimeo.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do value-added estimates add value? Accounting for learning dynamics". *American Economic Journal: Applied Economics* 3 (3): 29–54.

Aslam, Monazza, Shenila Rawal, and Sahar Saeed. 2017. *Public-Private Partnerships in Education in Developing Countries: A Rigorous Review of the Evidence*. Tech. rep. Ark Education Partnerships Group.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. 2015. "Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia". National Bureau of Economic Research. doi:10.3386/w20965. http://www.nber.org/papers/w20965.

Baird, Sarah, Craig McIntosh, and Berk Özler. 2011. "Cash or condition? Evidence from a cash transfer experiment". *The Quarterly Journal of Economics* 126 (4): 1709–1753.

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. 2017. "Labor markets and poverty in village economies". *The Quarterly Journal of Economics* 132 (2): 811–870.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015a. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries". *Science* 348 (6236): 1260799.

Banerjee, Abhijit V, Esther Duflo, and Rachel Glennerster. 2008. "Putting a Band-Aid on a corpse: Incentives for nurses in the Indian public health care system". *Journal of the European Economic Association* 6 (2-3): 487–500.

Banerjee, Abhijit V., Rema Hanna, Jordan C Kyle, Benjamin A Olken, and Sudarno Sumarto. 2015b. *Contracting out the Last-Mile of Service Delivery: Subsidized Food Distribution in Indonesia*. Tech. rep. National Bureau of Economic Research.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India". *The Quarterly Journal of Economics* 122 (3): 1235–1264. eprint: http://qje.oxfordjournals.org/content/122/3/1235.full.pdf+html. http://qje.oxfordjournals.org/content/122/3/1235.abstract.

Barrera-Osorio, Felipe, David S Blakeslee, Matthew Hoover, L Linden, Dhushyanth Raju, and SP Rya. 2013. "Leveraging the private sector to improve primary school enrolment: Evidence from a randomized controlled trial in Pakistan". Mimeo.

BBC Africa. 2016. *Liberia – the country that wants to privatise its primary schools*. Visited on 06/01/2017. http://www.bbc.com/news/world-africa-36074964.

Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools". *Journal of Political Economy* 123 (2): 325–364. doi:10.1086/675910. eprint: https://doi.org/10.1086/675910. https://doi.org/10.1086/675910.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014a. "High - Dimensional Methods and Inference on Structural and Treatment Effects". *Journal of Economic Perspectives* 28, no. 2 (): 29–50. doi:10.1257/jep.28.2.29.

— . 2014b. "Inference on Treatment Effects after Selection among High-Dimensional Controls". *The Review of Economic Studies* 81 (2): 608–650. doi:10.1093/restud/rdt044.

Bennedsen, Morten, Kasper Meisner Nielsen, Francisco Pérez-González, and Daniel Wolfenzon. 2007. "Inside the family firm: The role of families in succession decisions and performance". *The Quarterly Journal of Economics* 122 (2): 647–691.

Besley, Timothy, and Maitreesh Ghatak. 2005. "Competition and incentives with motivated agents". *The American economic review* 95 (3): 616–636.

Besley, Timothy, and Torsten Persson. 2010. "State capacity, conflict, and development". *Econometrica* 78 (1): 1–34.

Betts, Julian R, and Y Emily Tang. 2014. *A Meta-Analysis of the Literature on the Effect of Charter Schools on Student Achievement*. Tech. rep. Society for Research on Educational Effectiveness.

Blimpo, Moussa Pouguinimpo, Moussa Blimpo, David Evans, and Nathalie Lahire. 2015. "Parental human capital and effective school management: evidence from The Gambia". Policy Research Working Paper;No. 7238. World Bank, Washington, DC.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2015. "Do Private Equity Owned Firms Have Better Management Practices?" *American Economic Review* 105, no. 5 (): 442–46. doi:10.1257/aer.p20151000.

Bloom, Nicholas, and John Van Reenen. 2010. "Why do management practices differ across firms and countries?" *The Journal of Economic Perspectives* 24 (1): 203–224.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. 2013. "Does management matter? Evidence from India". *The Quarterly Journal of Economics* 128 (1): 1–51.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen. 2015. "Does Management Matter in schools?" *The Economic Journal* 125 (584): 647–674. ISSN: 1468-0297. doi:10.1111/ecoj.12267.

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying. 2014. "Does working from home work? Evidence from a Chinese experiment". *The Quarterly Journal of Economics* 130 (1): 165–218.

Bold, Tessa, Mwangi S. Kimenyi, and Justin Sandefur. 2013. “Public and Private Provision of Education in Kenya”. *Journal of African Economies* 22 (suppl 2): ii39–ii56. eprint: http://jae.oxfordjournals.org/content/22/suppl_2/ii39.full.pdf+html. http://jae.oxfordjournals.org/content/22/suppl_2/ii39.abstract.

Brault, MW. 2011. *School-aged children with disabilities in U.S. metropolitan statistical areas: 2010. American community survey briefs*. Tech. rep. ACSBR/10-12. US Census Bureau.

Bridge International Academies. 2017. *Bridge International Academies' written evidence to the International Development Committee Inquiry on DFID's work on education: Leaving no one behind?* Tech. rep. House of Commons, International Development Committee.

Bruhn, Miriam, and David McKenzie. 2009. “In Pursuit of Balance: Randomization in Practice in Development Field Experiments”. *American Economic Journal: Applied Economics* 1, no. 4 (): 200–232. doi:10.1257/app.1.4.200.

Bruns, Barbara, and Javier Luque. 2014. *Great teachers: How to raise student learning in Latin America and the Caribbean*. World Bank Publications.

Bullock, John G, and Shang E Ha. 2011. “Mediation Analysis Is Harder than It Looks”. Chap. 35, ed. by James N Druckman, Donald P Green, James H Kuklinski, and Arthur Lupia, 959. Cambridge University Press.

Burde, Dana, and Leigh L Linden. 2013. “Bringing education to Afghan girls: A randomized controlled trial of village-based schools”. *American Economic Journal: Applied Economics* 5 (3): 27–40.

Burnside, Craig, and David Dollar. 2000. “Aid, Policies, and Growth”. *The American Economic Review* 90 (4): 847–868. ISSN: 00028282. http://www.jstor.org/stable/117311.

Cabral, Sandro, Sergio G. Lazzarini, and Paulo Furquim de Azevedo. 2013. “Private Entrepreneurs in Public Services: A Longitudinal Examination of Outsourcing and Statization of Prisons”. *Strategic Entrepreneurship Journal* 7 (1): 6–25. ISSN: 1932-443X. doi:10.1002/sej.1149.

Calefati, Jessica. 2016. *Dozens of California districts with worst test scores excluded from extra state help*. Visited on 05/05/2018. https://calmatters.org/articles/dozens-california-districts-worst-test-scores-excluded-extra-state-help/.

Cameron, Lisa, and Manisha Shah. 2017. “Scaling Up Sanitation: Evidence from an RCT in Indonesia”. Mimeo.

Chabrier, Julia, Sarah Cohodes, and Philip Oreopoulos. 2016. "What Can We Learn from Charter School Lotteries?" *The Journal of Economic Perspectives* 30 (3): 57–84.

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 (1): 91–116. doi:10.1257/089533006776526058. http://www.aeaweb.org/articles?id=10.1257/089533006776526058.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood". *American Economic Review* 104, no. 9 (): 2633–79. doi:10.1257/aer.104.9.2633.

Cohen, Peter A, James A. Kulik, and Chen-Lin C. Kulik. 1982. "Educational Outcomes of Tutoring: A Meta-analysis of Findings". *American Educational Research Journal* 19 (2): 237–248. doi:10.3102/00028312019002237. eprint: http://aer.sagepub.com/content/19/2/237.full.pdf+html. http://aer.sagepub.com/content/19/2/237.abstract.

Collier, Kiah. 2016a. *Lawmakers Look at Tying School Funding to Performance*. Visited on 05/05/2018. https://www.texastribune.org/2016/08/03/senators-examining-performance-based-funding-schoo/.

Collier, Paul. 2016b. *Fragile States and International Support*. Working Papers P175. FERDI. https://ideas.repec.org/p/fdi/wpaper/3375.html.

Collier, Paul, and David Dollar. 2002. "Aid allocation and poverty reduction". *European economic review* 46 (8): 1475–1500.

Contreras, Dante, and Tomás Rau. 2012. "Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile". *Economic Development and Cultural Change* 61 (1): 219–246. doi:10.1086/666955. eprint: https://doi.org/10.1086/666955. https://doi.org/10.1086/666955.

Crawfurd, Lee. 2017. "School Management and Public-Private Partnerships in Uganda". *Journal of African Economies* 26 (5): 539–560.

Cremata, Edward, Devora Davis, Kathleen Dickey, Kristina Lawyer, Yohannes Negassi, Margaret Raymond, and James L. Woodworth. 2013. *National charter school study*. Tech. rep. Center for Research on Education Outcomes, Stanford University.

Dal Bó, Ernesto, Frederico Finan, and Martín A. Rossi. 2013. "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service". *The Quarterly Journal of Economics* 128 (3): 1169–1218. doi:10.1093/qje/qjt008. eprint:

/oup/backfile/content_public/journal/qje/128/3/10.1093_qje_qjt008/4/qjt008.pdf. http://dx.doi.org/10.1093/qje/qjt008.

Das, Jishnu, and Tristan Zajonc. 2010. "India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement". *Journal of Development Economics* 92 (2): 175–187. ISSN: 0304-3878. doi:10.1016/j.jdeveco.2009.03.004. http://www.sciencedirect.com/science/article/pii/S0304387809000273.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. 2013. "School Inputs, Household Substitution, and Test Scores". *American Economic Journal: Applied Economics* 5 (2): 29–57.

Deci, E., and R.M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. Perspectives in Social Psychology. Springer US. ISBN: 9780306420221. https://books.google.com/books?id=p96Wmn-ER4QC.

Dhaliwal, Iqbal, and Rema Hanna. 2014. *Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India*. Tech. rep. National Bureau of Economic Research.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. 2013. "Comparative cost-effectiveness analysis to inform policy in developing countries: a general framework with applications for education". *Education Policy in Developing Countries*: 285–338.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya". *American Economic Review* 101 (5): 1739–74. doi:10.1257/aer.101.5.1739.

— . 2015. "School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools". *Journal of Public Economics* 123:92–110.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School". *American Economic Review* 102 (4): 1241–78. doi:10.1257/aer.102.4.1241.

Duggan, Mark. 2004. "Does contracting out increase the efficiency of government programs? Evidence from Medicaid HMOs". *Journal of Public Economics* 88 (12): 2549–2572. ISSN: 0047-2727. doi:10.1016/j.jpubeco.2003.08.003. http://www.sciencedirect.com/science/article/pii/S0047272703001415.

Evans, David, and Anna Popova. 2016. "What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews". *The World Bank Research Observer* 31 (2): 242–270.

Evans, David, and Fei Yuan. 2017. “The Economic Returns to Interventions that Increase Learning”. Mimeo.

Fehr, Ernst, and Armin Falk. 2002. “Psychological foundations of incentives”. *European economic review* 46 (4): 687–724.

Filmer, Deon, Amer Hasan, and Lant Pritchett. 2006. “A millennium learning goal: Measuring real progress in education”. Center for Global Development Working Paper 97.

Foreign Policy. 2016. *Liberia's Education Fire Sale*. Visited on 07/20/2017. http://foreignpolicy.com/2016/06/30/liberias-education-fire-sale/.

Fryer, R.G. 2017. “Chapter 2 - The Production of Human Capital in Developed Countries: Evidence From 196 Randomized Field Experimentsa”. In *Handbook of Economic Field Experiments*, ed. by Abhijit Vinayak Banerjee and Esther Duflo, 2:95–322. Handbook of Economic Field Experiments, Supplement C. North-Holland. doi:10.1016/bs.hefe.2016.08.006. http://www.sciencedirect.com/science/article/pii/S2214658X16300083.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. “Water for life: The impact of the privatization of water services on child mortality”. *Journal of political economy* 113 (1): 83–120.

Gelman, Andrew. 2006. “Prior distributions for variance parameters in hierarchical models”. *Bayesian Analysis* 1 (3): 515–533.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. 2014. *Bayesian data analysis*. Chapman & Hall/CRC Boca Raton, FL, USA.

Glewwe, P., and K. Muralidharan. 2016. “Chapter 10 - Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications”. In *Handbook of the Economics of Education*, ed. by Stephen Machin Eric A. Hanushek and Ludger Woessmann, 5:653–743. Elsevier. doi:10.1016/B978-0-444-63459-7.00010-5. http://www.sciencedirect.com/science/article/pii/B9780444634597000105.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. “Teacher incentives”. *American Economic Journal: Applied Economics* 2 (3): 205–227.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. “Many Children Left Behind? Textbooks and Test Scores in Kenya”. *American Economic Journal: Applied Economics* 1 (1): 112–35. doi:10.1257/app.1.1.112.

Green, Donald P., Shang E. Ha, and John G. Bullock. 2010. “Enough Already about “Black Box” Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose”. *The ANNALS of the American Academy of Political and Social Science* 628 (1): 200–208. doi:10.1177/0002716209351526. eprint: http://dx.doi.org/10.1177/0002716209351526. http://dx.doi.org/10.1177/0002716209351526.

Gurkan, Asli, Kai Kaiser, and Doris Voorbraak. 2009. “Implementing public expenditure tracking surveys for results: lessons from a decade of global experience”. PREM Notes; No. 145. World Bank, Washington, DC.

Hanushek, Eric A, John F Kain, and Steven G Rivkin. 2004. “Disruption versus Tiebout improvement: The costs and benefits of switching schools”. *Journal of public Economics* 88 (9): 1721–1746.

Hanushek, Eric A., and Dennis D. Kimko. 2000. “Schooling, Labor-Force Quality, and the Growth of Nations”. *American Economic Review* 90 (5): 1184–1208. doi:10.1257/aer.90.5.1184.

Hanushek, Eric A, and Ludger Wößmann. 2007. “The role of education quality for economic growth”. World Bank Policy Research Working Paper.

Hart, Oliver, Andrei Shleifer, and Robert W Vishny. 1997. “The proper scope of government: theory and an application to prisons”. *The Quarterly Journal of Economics* 112 (4): 1127–1161.

Heckman, James, and Rodrigo Pinto. 2015. “Econometric Mediation Analyses: Identifying the Sources of Treatment Effects from Experimentally Estimated Production Technologies with Unmeasured and Mismeasured Inputs”. *Econometric Reviews* 34 (1-2): 6–31. http://dx.doi.org/10.1080/07474938.2014.944466.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. “Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes”. *American Economic Review* 103 (6): 2052–86. doi:10.1257/aer.103.6.2052.

Heckman, James J, and Jeffrey A Smith. 1995. “Assessing the case for social experiments”. *The Journal of Economic Perspectives* 9 (2): 85–110.

Hirshleifer, Sarojini. 2017. “Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance”. Mimeo.

Holmstrom, Bengt, and Paul Milgrom. 1991. “Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design”. *Journal of Law, Economics, & Organization* 7:24–52.

Hsieh, Chang-Tai, and Miguel Urquiola. 2006. “The effects of generalized school choice on achievement and stratification: Evidence from Chile’s voucher program”. *Journal of public Economics* 90 (8): 1477–1503.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. “A general approach to causal mediation analysis”. *Psychological methods* 15 (4): 309.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. “Identification, inference and sensitivity analysis for causal mediation effects”. *Statistical science* 25 (1): 51–71.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. 2016. “The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms”. *The Quarterly Journal of Economics* 131 (1): 157–218. doi:10.1093/qje/qjv036. eprint: /oup/backfile/content_public/journal/qje/131/1/10.1093_qje_qjv036/1/qjv036.pdf. http://dx.doi.org/10.1093/qje/qjv036.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. “An Introduction to Statistical Learning with Applications in R”. Springer.

Johnston, Bruce F., and John W. Mellor. 1961. “The Role of Agriculture in Economic Development”. *The American Economic Review* 51 (4): 566–593. issn: 00028282. http://www.jstor.org/stable/1812786.

Jones, Sam, Youdi Schipper, Sara Ruto, and Rakesh Rajani. 2014. “Can Your Child Read and Count? Measuring Learning Outcomes in East Africa”. *Journal of African Economies* 23 (5): 643–672. doi:10.1093/jae/eju009. eprint: /oup/backfile/content_public/journal/jae/23/5/10.1093/jae/eju009/2/eju009.pdf.

Kerwin, Jason T, and Rebecca Thornton. 2015. “Making the Grade: Understanding What Works for Teaching Literacy in Rural Uganda”. Mimeo.

Kerwin, Jason Theodore, and Rebecca L Thornton. 2017. “Making the Grade: The Trade-off between Efficiency and Effectiveness in Improving Student Learning”.

Kiessel, Jessica, and Annie Duflo. 2014. *Cost-effectiveness report: The teacher community assistant initiative (TCAI)*. Visited on 08/06/2017. http://www.poverty-action.org/sites/default/files/publications/TCAI_Cost-Effectiveness_2014.3.26.pdf.

King, Simon, Medina Korda, Lee Nordstrum, and Susan Edwards. 2015. *Liberia Teacher Training Program: ENDLINE ASSESSMENT OF THE IMPACT OF EARLY GRADE READING AND MATHEMATICS INTERVENTIONS*. Tech. rep. RTI International.

Krasner, Stephen D, and Thomas Risse. 2014. “External actors, state-building, and service provision in areas of limited statehood: Introduction”. *Governance* 27 (4): 545–567.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World". *Science* 340 (6130): 297–300. ISSN: 0036-8075. doi:10.1126/science.1235350. eprint: http://science.sciencemag.org/content/340/6130/297.full.pdf. http://science.sciencemag.org/content/340/6130/297.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to learn". *The Review of Economics and Statistics* 91 (3): 437–456.

Kwauk, Christina, and Jenny Perlman Robinson. 2016. "Bridge International Academies: Delivering Quality Education at a Low Cost in Kenya, Nigeria, and Uganda". Visited on 08/09/2017. http://www.bridgeinternationalacademies.com/wp-content/uploads/2016/09/Brookings-Millions-Learning-case-study.pdf.

Ladner, Peter, and Torsten Persson. 2009. "The origins of state capacity: Property rights, taxation, and politics". *The American Economic Review* 99 (4): 1218–1244.

Lavy, V. 2002. "Evaluating the effect of teachers' group performance incentives on pupil achievement". *Journal of Political Economy* 110 (6): 1286–1317.

Lavy, Victor. 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics". *American Economic Review* 99 (5): 1979–2011. doi:10.1257/aer.99.5.1979.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". *The Review of Economic Studies* 76 (3): 1071–1102. doi:10.1111/j.1467-937X.2009.00536.x. eprint: http://restud.oxfordjournals.org/content/76/3/1071.full.pdf+html. http://restud.oxfordjournals.org/content/76/3/1071.abstract.

Lemos, Renata, and Daniela Scur. 2016. "Developing Management: An expanded evaluation tool for developing countries". Mimeo.

Levitt, Steven D, John A List, Susanne Neckermann, and Sally Sadoff. 2016. "The behavioralist goes to school: Leveraging behavioral economics to improve educational performance". *American Economic Journal: Economic Policy* 8 (4): 183–219.

Liberia Institute of Statistics and Geo-Information Services. 2016. *Liberia - Household Income and Expenditure Survey 2014-2015*. Liberia Institute of Statistics / Geo-Information Services.

— . 2014. *Liberia Demographic and Health Survey 2013*. Liberia Institute of Statistics / Geo-Information Services.

Linden, Wim J van der. 2017. *Handbook of Item Response Theory*. CRC Press.

Loevinsohn, Benjamin, and April Harding. 2005. "Buying results? Contracting for health service delivery in developing countries". *The Lancet* 366 (9486): 676–681.

Lucas, Adrienne M, and Isaac M Mbiti. 2012. "Access, sorting, and achievement: the short-run effects of free primary education in Kenya". *American Economic Journal: Applied Economics* 4 (4): 226–253.

Mail & Guardian Africa. 2016a. *An Africa first! Liberia outsources entire education system to a private American firm. Why all should pay attention*. Visited on 07/20/2017. http://mgafrica.com/article/2016-03-31-liberia-plans-to-outsource-its-entire-education-system-to-a-private-company-why-this-is-a-very-big-deal-and-africa-should-pay-attention.

— . 2016b. *An update on Bridge Academies in Liberia, and why people need dreams - and yes, sweet lies - too*. Visited on 07/20/2017. http://mgafrica.com/article/2016-05-07-an-update-on-bridge-academies-in-liberia-and-why-people-need-dreams-and-yes-sweet-lies-too.

May, Shannon. 2017. *Oral evidence: DFID's work on education: Leaving no one behind?, HC 639*. Tech. rep. House of Commons, International Development Committee.

Mbiti, Isaac. 2016. "The Need for Accountability in Education in Developing Countries". *Journal of Economic Perspectives* 30 (3): 109–32. doi:10.1257/jep.30.3.109.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Rakesh Rajani, and Constantine Manda. 2017. "Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania". Mimeo.

Meager, Rachael. 2016. "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature". Mimeo.

Mesecar, Doug, and Don Soifer. 2016. *How performance-based funding can improve education funding*. Visited on 05/05/2018. https://www.brookings.edu/blog/brown-center-chalkboard/2016/02/24/how-performance-based-funding-can-improve-education-funding/.

Ministry of Education - Republic of Liberia. 2017. *Getting to Best Education Sector Plan, 2017-2021*.

— . 2016a. *Liberia Education Statistics Report 2015-2106*.

— . 2016b. *Memorandum of Understanding BETWEEN Ministry of Education, Government of Liberia AND Bridge International Academies*. Visited on 08/06/2017. www.theperspective.org/2016/ppp_mou.pdf.

Mullainathan, Sendhil. 2005. "Development economics through the lens of psychology". In *Annual World Bank Conference on Development Economics 2005: Lessons of Experience*.

Munk, N. 2013. *The Idealist: Jeffrey Sachs and the Quest to End Poverty*. Knopf Doubleday Publishing Group. ISBN: 9780385537742. https://books.google.com/books?id=lF8vezXqSawC.

Munley, Vincent G., Eoghan Garvey, and Michael J. McConnell. 2010. "The Effectiveness of Peer Tutoring on Student Achievement at the University Level". *The American Economic Review* 100 (2): 277–282. ISSN: 00028282. http://www.jstor.org/stable/27805004.

Muralidharan, K. 2017. "Chapter 3 - Field Experiments in Education in Developing Countries". In *Handbook of Economic Field Experiments*, ed. by Abhijit Vinayak Banerjee and Esther Duflo, 2:323–385. Handbook of Economic Field Experiments, Supplement C. North-Holland. doi:10.1016/bs.hefe.2016.09.004. http://www.sciencedirect.com/science/article/pii/S2214658X16300125.

Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at Scale". *Journal of Economic Perspectives* 31 (4): 103–24.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2016. "Building state capacity: Evidence from biometric smartcards in India". *The American Economic Review* 106 (10): 2895–2929.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wuthrich. 2018. "Improving Inference in Experiments with Factorial Designs". Mimeo.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian. 2016. *Disrupting education? Experimental evidence on technology-aided instruction in India*. Tech. rep. National Bureau of Economic Research.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. *Contract teachers: Experimental evidence from India*. Tech. rep. National Bureau of Economic Research.

— . 2011a. "Teacher opinions on performance pay: Evidence from India". *Economics of Education Review* 30 (3): 394–403.

— . 2011b. "Teacher Performance Pay: Experimental Evidence from India". *Journal of Political Economy* 119 (1): 39–77. ISSN: 00223808. doi:10.1086/659655.

— . 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India". *The Quarterly Journal of Economics* 130 (3): 1011. doi:10.

1093/qje/qjv013. eprint: /oup/backfile/content_public/journal/qje/130/3/10.1093_qje_qjv013/4/qjv013.pdf. +.

Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal. 2017. "The fiscal cost of weak governance: Evidence from teacher absence in India". *Journal of Public Economics* 145:116–135.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability", *Review of Economics and Statistics* 92, no. 2 (): 263–283. ISSN: 0034-6535.

OHCHR. 2016. *UN rights expert urges Liberia not to hand public education over to a private company*. Visited on 06/01/2017. http://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=18506.

Patrinos, Harry Anthony, Felipe Barrera Osorio, and Juliana Guáqueta. 2009. *The role and impact of public-private partnerships in education*. World Bank Publications.

Piper, Benjamin, and Medina Korda. 2011. "EGRA Plus: Liberia. Program Evaluation Report". RTI International.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha. 2014. "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia". *American Economic Journal: Applied Economics* 6, no. 2 (): 105–26. doi:10.1257/app.6.2.105.

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

Pritchett, Lant, and Michael Woolcock. 2004. "Solutions when the Solution is the Problem: Arraying the Disarray in Development". *World Development* 32 (2): 191–212.

Quintilianus, M.F., and K. Halm. 1869. *Institutio oratoria*. Bibliotheca scriptorum Graecorum et Romanorum Teubneriana v. 2. Teubner. https://books.google.com/books?id=jaI9AAAAcAAJ.

Ray, D. 1998. *Development Economics*. Princeton University Press. ISBN: 9781400835898. https://books.google.com.co/books?id=GKr5RxWT4uAC.

Ree, Joppe de, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2015. *Double for Nothing? Experimental Evidence on the Impact of an Unconditional Teacher Salary Increase on Student Performance in Indonesia*. Working Paper, Working Paper Series 21806. National Bureau of Economic Research. doi:10.3386/w21806. http://www.nber.org/papers/w21806.

Reinikka, Ritva, and Nathanael Smith. 2004. *Public expenditure tracking surveys in education*. UNESCO, International Institute for Educational Planning.

Rubin, Donald B. 1981. "Estimation in parallel randomized experiments". *Journal of educational and behavioral statistics* 6 (4): 377–401.

Sabarwal, Shwetlena, David K. Evans, and Anastasia Marshak. 2014. *The permanent input hypothesis : the case of textbooks and (no) student learning in Sierra Leone*. Policy Research Working Paper Series 7021. The World Bank.

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates". *The Quarterly Journal of Economics* 116 (2): 681–704. eprint: http://qje.oxfordjournals.org/content/116/2/681.full.pdf+html. http://qje.oxfordjournals.org/content/116/2/681.abstract.

Sachs, J. 2006. *The End of Poverty: Economic Possibilities for Our Time*. Penguin Books. isbn: 9780143036586. https://books.google.com/books?id=PNI9tqKVicIC.

Schermerhorn, J.R., R.N. Osborn, M. Uhl-Bien, and J.G. Hunt. 2011. *Organizational Behavior*. Wiley. isbn: 9780470878200. https://books.google.com/books?id=8eRtuZeIguIC.

Shenderovich, Yulia, Allen Thurston, and Sarah Miller. 2016. "Cross-age tutoring in kindergarten and elementary school settings: A systematic review and meta-analysis". *International Journal of Educational Research* 76:190–210.

Singh, Abhijeet. 2015a. *How standard is a standard deviation? A cautionary note on using SDs to compare across impact evaluations in education*. Visited on 07/31/2017. http://blogs.worldbank.org/impactevaluations/how-standard-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations.

— . 2016. "Learning more with every year: School year productivity and international learning divergence". Mimeo.

— . 2015b. "Private school effects in urban and rural India: Panel estimates at primary and secondary school ages". *Journal of Development Economics* 113:16–32.

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt, and Emmanuel Jimenez. 2016. "The impact of education programmes on learning and school participation in low-and middle-income countries". International Initiative for Impact Education.

Stallings, Jane A, Stephanie L Knight, and David Markham. 2014. *Using the stallings observation system to investigate time on task in four countries*. Tech. rep. World Bank.

The New York Times. 2016. *Liberia, Desperate to Educate, Turns to Charter Schools*. Visited on 07/20/2017. http://www.nytimes.com/2016/06/14/opinion/liberia-desperate-to-educate-turns-to-charter-schools.html.

Todd, Petra E, and Kenneth I Wolpin. 2003. "On the specification and estimation of the production function for cognitive achievement". *The Economic Journal* 113 (485): F3–F33.

Tuttle, Christina Clark, Philip Gleason, and Melissa Clark. 2012. "Using lotteries to evaluate schools of choice: Evidence from a national study of charter schools". *Economics of Education Review* 31 (2): 237–253.

UNESCO. 2016. *Global Monitoring Report 2016*. Tech. rep. United Nations.

UNICEF. 2013. *The State of the World's Children: Children with Disabilities*. Tech. rep. United Nations.

United Nations. 2015. *The millennium development goals report 2015*. United Nations Publications.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. 2016. "Using Double-Lasso Regression for Principled Variable Selection". Mimeo.

USAID. 2017. "Request for proposals - SOL-669-17-000004, Read Liberia". Visited on 08/06/2017. https://www.fbo.gov/index?s=opportunity&mode=form&id=e53cb285301f7014f415ce91b14049a3&tab=core&tabmode=list&=.

Useem, Bert, and Jack A. Goldstone. 2002. "Forging Social Order and Its Breakdown: Riot and Reform in U.S. Prisons". *American Sociological Review* 67 (4): 499–525. ISSN: 00031224. http://www.jstor.org/stable/3088943.

Uwezo. 2017. *Are Our Children Learning?* Tech. rep. Accessed on 02-02-2018. Uwezo.

Uwezo. 2013. *Are Our Children Learning? Numeracy and Literacy across East Africa*. Uwezo East-Africa Report. Accessed on 05-12-2014. Nairobi: Uwezo.

Valente, Christine. 2015. "Primary Education Expansion and Quality of Schooling: Evidence from Tanzania". IZA Discussion Paper.

Vox World. 2016. *Liberia is outsourcing primary schools to a startup backed by Mark Zuckerberg*. Visited on 07/20/2017. http://www.vox.com/2016/4/8/11347796/liberia-outsourcing-schools.

Werner, George K. 2017. "Liberia has to work with international private school companies if we want to protect our children's future". *Quartz Africa*. Visited on

07/20/2017. https://qz.com/876708/why-liberia-is-working-with-bridge-international-brac-and-rising-academies-by-education-minister-george-werner/.

Woodworth, James L., Margaret Raymond, Chunping Han, Yohannes Negassi, W. Payton Richardson, and Will Snow. 2017. *Charter Management Organizations*. Tech. rep. Center for Research on Education Outcomes, Stanford University.

World Bank. 2015. *Conducting classroom observations: analyzing classrooms dynamics and instructional time, using the Stallings' classroom snapshot'observation system. User guide*. Tech. rep. World Bank Group.

— . 2015a. *GDP per capita (current US$)*. Data retrieved from World Development Indicators, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD.

— . 2007. *Global Monitoring Report*. doi:10.1596/978-0-8213-6975-3. eprint: http://elibrary.worldbank.org/doi/pdf/10.1596/978-0-8213-6975-3.

— . 2015b. *Government expenditure per student, primary (% of GDP per capita)*. Data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SE.XPD.PRIM.PC.ZS.

— . 2014. *Life expectancy*. Data retrieved from World Development Indicators, http://data.worldbank.org/indicator/SE.PRM.NENR?locations=LR.

— . 2015c. *Net primary enrollment in low-income countries*. Data retrieved from World Development Indicators, http://data.worldbank.org/indicator/SE.PRM.NENR?locations=XM.

— . 2012. *Tanzania Service Delivery Indicators*. Tech. rep. Washington D.C.: World Bank.

Zhang, Hongliang. 2014. "The mirage of elite schools: evidence from lottery-based school admissions in China". Mimeo.

Zimmerman, David J. 2003. "Peer effects in academic outcomes: Evidence from a natural experiment". *Review of Economics and Statistics* 85 (1): 9–23.