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

Mauricio Romero* Justin Sandefur† Wayne Aaron Sandholtz*

May 1, 2018‡
(Click here for latest version with appendices)

Abstract

Can outsourcing improve public service delivery in fragile states? To answer this question, we present results from a field experiment in Liberia, where the government delegated management of 93 public schools — staffed by government teachers and run free of charge to students — to private providers. We randomly assigned treatment status at the school level and sampled students from pre-treatment enrollment records to identify the effectiveness of the treatment without confounding the effect of endogenous sorting of pupils into schools. After one academic year, students in outsourced schools scored $.18\sigma$ higher in English and mathematics than those in control schools. Private providers improved scores on an index of managerial practices and significantly reduced teacher absenteeism ("better management"), but also spent significantly more per student and employed more teachers than control schools ("extra resources"). Non-experimental mediation analysis suggests better management and extra resources played roughly equal roles in the observed learning gains. In line with program rules, we find no evidence that providers engaged in selective admissions. Our design allows us to study heterogeneity across providers: While the highest-performing providers increased learning by 0.26σ, the lowest-performing had no impact. Providers also differed in behavior that may generate negative spillovers for the broader school system, including removing pupils to keep class sizes small, and reassigning underperforming teachers to other public schools. These results suggest that leveraging the private sector to improve service delivery in fragile states is promising, but also highlight the importance of procurement rules and contracting details in aligning public and private interests.

Keywords: Public-Private Partnership; Randomized Controlled Trial; School Management
JEL Codes: I25, I28, C93, L32, L33

---

*University of California, San Diego.

†Center for Global Development.

‡We are grateful to the Minister of Education, George K. Werner, Deputy Minister Romelle Horton, Binta Massaquoi, Nisha Makan, and the Partnership Schools for Liberia (PSL) team, as well as Susannah Hares, Robin Horn, and Joe Collins from Ark EPG for their commitment throughout this project to ensuring a rigorous and transparent evaluation of the PSL program. Thanks to Arja Dayal, Dackermue Dolo, and their team at Innovations for Poverty Action who led the data collection. Avi Ahuja, Dev Patel, and Benjamin Tan provided excellent research assistance. We're grateful to Michael Kremer, Karthik Muralidharan, and Pauline Rose who provided detailed comments on the government report of the independent evaluation of the PSL program. The design and analysis benefited from comments and suggestions from Maria Atuesta, Prashant Bharadwaj, Jeffrey Clemens, Joe Collins, Mitch Downey, Susannah Hares, Robin Horn, Isaac Mbiti, Gordon McCord, Craig McIntosh, Karthik Muralidharan, Owen Ozier, Olga Romero, Santiago Saavedra, Diego Vera-Cossio, and seminar participants at the Center for Global Development and UC San Diego. A randomized controlled trials registry entry is available at: https://www.socialscienceregistry.org/trials/1501 as well as the pre-analysis plan. IRB approval was received from IPA (protocol #14227) and the University of Liberia (protocol #17-04-39) prior to any data collection. UCSD IRB approval (protocol #161605S) was received after the first round of data collection but before any other activities were undertaken. The evaluation was supported by the UBS Optimus Foundation and Aestus Trust. Romero and Sandefur acknowledge financial support from the Central Bank of Colombia through the Lauchlin Currie scholarship and the Research on Improving Systems of Education (RISE) program, respectively. The views expressed here are ours, and not those of the Ministry of Education of Liberia or our funders. All errors are our own.

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 & Woolcock, 2004). Influential studies in the 1990s concluded that development aid was least effective in poorly governed states, and advocated directing aid elsewhere (Burnside & Dollar, 2000; Collier & 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 & Risse, 2014; Collier, 2016). 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, & 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, & Schargrodsky, 2005) and food distribution in Indonesia (Banerjee, Hanna, Kyle, Olken, & Sumarto, 2015)), it has failed to do so in others (e.g., prisons in the U.S. (Useem & Goldstone, 2002) and in Brazil (Cabral, Lazzarini, & de 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, & Guáqueta, 2009). On the efficiency side, evidence suggest that private firms (Bloom & Van Reenen, 2010; Bloom, Sadun, & Van Reenen, 2015) and schools (Bloom, Lemos, Sadun, & Van Reenen, 2015; Muralidharan & Sundararaman, 2015) tend to be better managed than their public counterparts. However, fee-charging private schools may increase inequality and induce sorting (Hsieh & Urquiola, 2006; Lucas & 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 (World Bank, 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

[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%).

$\sim$\$50 in the control group) as part of the program, and some providers independently raised and spent far more.[2] In addition, 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 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

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

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 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 & 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, & 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 et al. (1997) argue that the bigger the adverse consequences of non-contractible

[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.

[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.

quality shading, the stronger the case for governments to provide services directly.[7] Some 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); 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, & 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 & Persson, 2009; Besley & Persson, 2010; Muralidharan, Niehaus, & 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, 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.

[7] Empirically, in cases where quality is easy to measure and to enforce, such as water services (Galiani et al., 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 & Harding, 2005). In contrast, for services for which quality is difficult to measure, such as prisons (Useem & Goldstone, 2002; Cabral et al., 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).

[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.

2 Experimental design

2.1 The program

2.1.1 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 (World Bank, 2014). Low *net* enrollment is partially explained by an extraordinary backlog of over-age children (see Figure 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 (Liberia Institute of Statistics and Geo-Information Services, 2016). Learning levels are low: Only 25% of adult women who finish elementary school can read a complete sentence (Liberia Institute of Statistics and Geo-Information Services, 2014) (there is no information for men).

[Figure 1 about here.]

2.1.2 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 government 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' 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

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

[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.

[12] Some agency problems related to contracting out the provision of a public good are alleviated by "mission-matching" (Besley & Ghatak, 2005; Akerlof & 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 3.4 for statistics on these fees.

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 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. 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.

[Table 1 about here.]

2.1.3 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.

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,

[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.

what activities the provider had engaged in. We summarize teachers' responses in Figure 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.

[Figure 2 about here.]

2.1.4 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 (World Bank, 2015c, 2015b).[17]

In the first year, providers spent far more than this amount.[18] *Ex ante* per-pupil 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 3a). *Ex post* per-pupil expenditure submited to the evaluation team at the end of 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 3b). These differences in costs are large relative to differences in treatment effects on learning, implying that cost-effectiveness may be driven

[15] As shown in Section 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 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.

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 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 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]

[Figure 3 about here.]

2.2 Experimental design

2.2.1 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

[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.

[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 & 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.

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 4a shows all public schools in Liberia and those within our sample. Table **??** in Appendix **??** has details on the differences between schools in the experiment and other public schools.

[Figure 4 about here.]

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 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 **??** 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 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

[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.

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

schools and Rising Academies opted not to operate one of their assigned schools. In short, there are 7 non-compliant treatment schools.[27] Figure 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.

2.2.2 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 **??** in Appendix **??** 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

[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. For details on school allocation, see Appendix **??** which contains a complete list of the schools related to the PSL program. Table **??** summarizes the overlap between schools in our main sample and the set of schools actually managed by PSL providers.

[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 **??** in Appendix **??**. 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 2). Second, the effects on English and math test scores appear to materialize in the later weeks of the fieldwork, as shown in Figure **??**, 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 **??** in Appendix **??**.

treatment was truly randomly assigned (see Section 2.2.5).

2.2.3 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 **??** has details on the construction of the test.

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.

2.2.4 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, 2015a). Furthermore, we conducted school-level surveys to collect information about school facilities, the teacher roster, input availability (e.g., textbooks) and expenditures.

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

[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, Das, Khwaja, and Zajonc (2011); Andrabi, Das, and Khwaja (2017); Singh (2015b, 2016); Muralidharan, Singh, and Ganimian (2016); Mbiti et al. (2017). Das and Zajonc (2010) provide a nice introduction to IRT models, while van der Linden (2017) provides a full treatment of IRT models.

Enumerators collected information on some school practices. Specifically, enumerators recorded whether the school has an enrollment log and what information it 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, Osborn, Uhl-Bien, & Hunt, 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.

2.2.5 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 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 **??** in Appendix **??**).

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 **??** has more details on the tracking and attrition that took place in each round of data collection.

[Table 2 about here.]

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

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]

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} is an error term.

$$Y_{isg} = \alpha_g + \beta_1 treat_s + \varepsilon_{isg} \quad (1)$$

$$Y_{isg} = \alpha_g + \beta_2 treat_s + \gamma_2 X_i + \delta_2 Z_s + \varepsilon_{isg} \quad (2)$$

$$Y_{isg} = \alpha_g + \beta_3 treat_s + \gamma_3 X_i + \delta_3 Z_s + \zeta_3 Y_{isg,-1} + \varepsilon_{isg} \quad (3)$$

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 (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 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

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

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

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

students tested during the first and the second half of the first round of data collection (see Figure **??** in Appendix **??**), and show that the treatment effects fade in during the course of field work.

In our preferred specification (Column 6) the treatment effect of PSL after one academic year is $.18\sigma$ for English (p-value < 0.001) and $.18\sigma$ 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 **??** in Appendix **??** which shows both the ITT and the ToT. Our results are robust to different measures of student ability (see Table **??** in Appendix **??** for details).

[Table 3 about here.]

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 ($.19\sigma$, p-value < 0.001), and statistically indistinguishable from the treatment effect over all the items ($.19\sigma$, 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 ($.12\sigma$, 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 **??** in Appendix **??**).

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 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 $.31\sigma$ in English and $.28\sigma$ in math. Thus, the treatment effect is roughly 0.56 EYOS for English and 0.65 EYOS for math. See Appendix **??** for a detailed explanation of the methodology to estimate EYOS, and a comparison of EYOS and standard deviation across countries. Additionally, Appendix **??** 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.

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 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 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]

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 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 4).[39] Among our sample of students (i.e., students sampled from the 2015/2016 enrollment log), students are equally likely across treatment and

[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 **??** uses EMIS data, while Table 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.

[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).

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 4). Finally, selection analysis using student-level data on wealth, gender, and age finds no evidence of systematic exclusions (see Table **??** in Appendix **??**).

[Table 4 about here.]

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 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 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 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, & 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 **??** in Appendix **??**).

[Table 5 about here.]

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.

3.3.1 Inputs and resources

Teachers, one of the most important inputs of education, change in several ways (see Panels A/B in Table 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 (de Ree, Muralidharan, Pradhan, & Rogers, 2015).

Our enumerators conducted a "materials" check during classroom observations (See Panels C - Table 6). Since we could not conduct classroom observations in schools that were out of session during our visit, Table **??** in Appendix **??** 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

[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, Korda, Nordstrum, and Edwards (2015) for details on this program.

cannot rule out that there is no overall effect as zero is between the Lee (2009) bounds.

[Table 6 about here.]

3.3.2 School management

Two important management changes are shown in Table 7: PSL schools are 8.7 percentage points more likely to be in session (i.e., the school is open, students 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.

[Table 7 about here.]

3.3.3 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 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 8). Also, during these spot checks we used

[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 **??** has details on every component of the good practices index.

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

the Stallings classroom observation instrument to study teacher time use and classroom management (see Panel C in Table 8).

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 et al. (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 & Luque, 2014).

[Table 8 about here.]

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 **??** 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 **??** in Appendix **??** shows 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, Liang, Roberts, & Ying, 2014; Bloom, Eifert, Mahajan, McKenzie, & Roberts, 2013; Bennedsen, Nielsen, Pérez-González, & Wolfenzon, 2007).

[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.

3.4 Other outcomes

Student data (Table 9, Panel C) and household data (Table 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 **??** in Appendix **??** 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 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 2.1.3). Indeed, household expenditures on fees, textbooks, and uniforms drop (see Table **??** 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 **??** for the effect on each component).

To complement the effect of the program on cognitive skills, we study student attitudes and opinions (see Table 9, Panel C). 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, Bau, Das, and Khwaja (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 9 about here.]

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

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

$$
\begin{aligned}
M_{isg} &= \alpha_g + \beta_4 treat_s + \gamma_4 X_i + \delta_4 Z_s + u_{isg} && (4)\\
Y_{isg} &= \alpha_g + \beta_5 treat_s + \gamma_5 X_i + \delta_5 Z_s + \theta_5 M_{isg} + \varepsilon_{isg}, && (5)
\end{aligned}
$$

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 4 is used to estimate the effect of treatment on the mediator (β_4), while equation 5 is used to estimate the effect of the mediator on learning (θ_5).

The mediation effect is $\beta_4 \times \theta_5$, i.e., the effect of the mediator on learning gains (θ_5) combined with changes in the mediator caused by treatment (β_4). β_5 captures the treatment effect that is not mediated by M_{isg}. β_5 is often refereed to as the "direct effect", but it can be a treatment effect mediated by unmeasured mediators. The mediation effect ($\beta_4 \times \theta_5$) and the direct effect (β_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 θ_5 (and therefore of β_5). Imai, Keele, and Yamamoto (2010) show that the OLS estimators for β_5 and θ_5 are consistent under the following assumption:

Assumption 1 (Sequential ignorability)

$$Y_i(t,m), M_i(t) \perp\!\!\!\perp T_i | X_i = x \tag{6}$$

$$Y_i(t,m) \perp\!\!\!\perp M_i(t) | X_i = x, T_i = t \tag{7}$$

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 5 shows the difference between a randomization model without mediation (5a), a mediation model with all the possible causal relationships (5b), and a mediation model under assumption 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).

[Figure 5 about here.]

While randomization implies that equation 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 7 in Assumption 1 (Green, Ha, & Bullock, 2010; Bullock & 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 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, & Hansen, 2014b, 2014a; Urminsky, Hansen, & 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 5 with these mediators are shown in Columns 2 and 3 of Table 10. 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 5 with these mediators are shown in Columns 4 and 5 of Table 10. For reference, we include a regression with no mediators (Column 1) which replicates the results from Table 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

[46]Lasso is similar to OLS but penalizes according to the number of controls used. See James, Witten, Hastie, and Tibshirani (2014) for a recent discussion.

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.

[Table 10 about here.]

In Section 3 we estimated equation (4) for several mediators. Combining those results with the results from Table 10, we show in Figure 6 the mediation effect ($\beta_4 \times \theta_5$) for the intermediate outcomes selected by "Double Lasso", as well as the direct effect (β_5). The left panel uses only raw inputs as mediators, while the right panel also includes changes in the use of inputs. Figure **??** in Appendix **??** 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 3.3.3 suggest that providers are able to increase teacher attendance even if the pool of 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 θ_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 **??** in Appendix **??**). 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.

[Figure 6 about here.]

5 Provider comparisons

The main results in Section 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.

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 **??** 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 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, Carlin, Stern, and Rubin (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

[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.

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

differences and estimating the treatment effect of each provider on the average school in our sample.[49]

Formally, let

$$Y_{isgc} = \alpha_g + \beta_c treat_s + \varepsilon_{isgc} \tag{8}$$

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 8 and equation 1 is that the treatment effect (β_c) is provider specific.

Asymptotically, the estimator of the treatment effect for each provider is normally distributed (assuming the standard error is known):[50]

$$\hat{\beta}_c \sim N(\beta_c, \sigma_c^2) \tag{9}$$

The bayesian hirerichal model further assumes that

$$\beta_c \sim N(\mu, \tau^2) \tag{10}$$

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 **??** 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.

5.2 Baseline differences

As discussed in Section 2.2.1 and shown in Table **??**, PSL schools are not a representative sample of public schools. Furthermore, there is heterogeneity in school characteristics across providers. This is unsurprising

[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).

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

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 11 (Table **??** in Appendix **??** shows simple summary statistics for the schools of each provider). We reject 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 11 about here.]

5.3 Learning outcomes

The raw treatment effects on test scores for each individual provider shown in Figure 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 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 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 12.

[51] Figure **??** in Appendix **??** 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).

[Figure 7 about here.]

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 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]

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 et al., 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 & Ghatak, 2005; Akerlof & 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

[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.

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 5.2. To measure teacher selection, we study the number of teachers dismissed and the number of new teachers recruited (Table 12 - 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 12 and Figure **??** in Appendix **??** 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 12 about here.]

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, Duflo, Glennerster, and Tulloch (2013) for a review). Section 2.1.4 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

[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.

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 .19σ, 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 & Popova, 2016), a review by Kremer et al. (2013) of other interventions subject to experimental evaluation in developing countries highlights various interventions that yield higher per-dollar gains than PSL (see Figure 8).

[Figure 8 about here.]

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 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, Mwabu, Ng'ang'a, and Sandefur (2013); Dhaliwal and Hanna (2014); Kerwin and Thornton (2015); 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.

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

[56] For example, Table **??** 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 interventions.

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 non-existent, especially in developing countries (Aslam et al., 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.

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 et al. (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.

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

References

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

Agor, W. H. (1989). Intuition & strategic planning: How organizations can make. *The Futurist*, *23*(6), 20.

Akerlof, G. A., & Kranton, R. E. (2005). Identity and the economics of organizations. *Journal of Economic Perspectives*, *19*(1), 9-32. doi: 10.1257/0895330053147930

Andrabi, T., Bau, N., Das, J., & Khwaja, A. I. (2010). *Are bad public schools public "bads"? test scores and civic values in public and private schools.* (Mimeo)

Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review*, *107*(6), 1535-63. Retrieved from `http://www.aeaweb.org/articles?id=10.1257/aer.20140774` doi: 10.1257/aer.20140774

Andrabi, T., Das, J., Khwaja, A. I., & Zajonc, T. (2011). Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal: Applied Economics*, *3*(3), 29–54.

Aslam, M., Rawal, S., & Saeed, S. (2017). *Public-private partnerships in education in developing countries: A rigorous review of the evidence* (Tech. Rep.). Ark Education Partnerships Group.

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

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*(3), 1235–1264. Retrieved from `http://qje.oxfordjournals.org/content/122/3/1235.abstract` doi: 10.1162/qjec.122.3.1235

Banerjee, A. V., Duflo, E., & Glennerster, R. (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, A. V., Hanna, R., Kyle, J. C., Olken, B. A., & Sumarto, S. (2015). *Contracting out the last-mile of service delivery: Subsidized food distribution in Indonesia* (Tech. Rep.). National Bureau of Economic Research.

Barrera-Osorio, F., Blakeslee, D. S., Hoover, M., Linden, L., Raju, D., & Rya, S. (2013). *Leveraging the private sector to improve primary school enrolment: Evidence from a randomized controlled trial in Pakistan.* (Mimeo)

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a, May). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29-50. Retrieved from http://www.aeaweb.org/articles?id=10.1257/jep.28.2.29 doi: 10.1257/jep.28.2.29

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608-650. Retrieved from http://dx.doi.org/10.1093/restud/rdt044 doi: 10.1093/restud/rdt044

Bennedsen, M., Nielsen, K. M., Pérez-González, F., & Wolfenzon, D. (2007). Inside the family firm: The role of families in succession decisions and performance. *The Quarterly Journal of Economics*, *122*(2), 647–691.

Besley, T., & Ghatak, M. (2005). Competition and incentives with motivated agents. *The American economic review*, *95*(3), 616–636.

Besley, T., & Persson, T. (2010). State capacity, conflict, and development. *Econometrica*, *78*(1), 1–34.

Betts, J. R., & Tang, Y. E. (2014). *A meta-analysis of the literature on the effect of charter schools on student achievement* (Tech. Rep.). Society for Research on Educational Effectiveness.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., & Roberts, J. (2013). Does management matter? evidence from India. *The Quarterly Journal of Economics*, *128*(1), 1–51.

Bloom, N., Lemos, R., Sadun, R., & Van Reenen, J. (2015). Does management matter in schools? *The Economic Journal*, *125*(584), 647–674. doi: 10.1111/ecoj.12267

Bloom, N., Liang, J., Roberts, J., & Ying, Z. J. (2014). Does working from home work? evidence from a Chinese experiment. *The Quarterly Journal of Economics*, *130*(1), 165–218.

Bloom, N., Sadun, R., & Van Reenen, J. (2015, May). Do private equity owned firms have better management practices? *American Economic Review*, *105*(5), 442-46. Retrieved from http://www.aeaweb.org/articles?id=10.1257/aer.p20151000 doi: 10.1257/aer.p20151000

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

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). *Scaling up what works: experimental evidence on external validity in Kenyan education.* (Mimeo)

Brault, M. (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.

Bruns, B., & Luque, J. (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. World Bank Publications.

Bullock, J. G., & Ha, S. E. (2011). Mediation analysis is harder than it looks. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), (p. 959). Cambridge University Press.

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

Burnside, C., & Dollar, D. (2000). Aid, policies, and growth. *The American Economic Review*, *90*(4), 847-868. Retrieved from http://www.jstor.org/stable/117311

Cabral, S., Lazzarini, S. G., & de Azevedo, P. F. (2013). Private entrepreneurs in public services: A longitudinal examination of outsourcing and statization of prisons. *Strategic Entrepreneurship Journal*, *7*(1), 6–25. Retrieved from http://dx.doi.org/10.1002/sej.1149 doi: 10.1002/sej.1149

Cameron, L., & Shah, M. (2017). *Scaling up sanitation: Evidence from an RCT in indonesia.* (mimeo)

Chabrier, J., Cohodes, S., & Oreopoulos, P. (2016). What can we learn from charter school lotteries? *The Journal of Economic Perspectives*, *30*(3), 57–84.

Collier, P. (2016, November). *Fragile States and International Support* (Working Papers No. P175). FERDI. Retrieved from https://ideas.repec.org/p/fdi/wpaper/3375.html

Collier, P., & Dollar, D. (2002). Aid allocation and poverty reduction. *European economic review*, *46*(8), 1475–1500.

Crawfurd, L. (in press). School management in Uganda.
(Journal of African Economies)

Cremata, E., Davis, D., Dickey, K., Lawyer, K., Negassi, Y., Raymond, M., & Woodworth, J. L. (2013). *National charter school study* (Tech. Rep.). Center for Research on Education Outcomes, Stanford University.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2013). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics*, *5*(2), 29–57.

Das, J., & Zajonc, T. (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. Retrieved from http://www.sciencedirect.com/science/article/pii/S0304387809000273 doi: http://dx.doi.org/10.1016/j.jdeveco.2009.03.004

de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. (2015). *Double for nothing? experimental evidence on* *the impact of an unconditional teacher salary increase on student performance in Indonesia* (Working Paper No. 21806). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w21806` doi: 10.3386/w21806

Dhaliwal, I., Duflo, E., Glennerster, R., & Tulloch, C. (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.

Dhaliwal, I., & Hanna, R. (2014). *Deal with the devil: The successes and limitations of bureaucratic reform in India* (Tech. Rep.). National Bureau of Economic Research.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739–74. doi: 10.1257/aer.101.5.1739

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*, *123*, 92–110.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, *102*(4), 1241–78. doi: 10.1257/aer.102.4.1241

Duggan, M. (2004). Does contracting out increase the efficiency of government programs? evidence from Medicaid HMOs. *Journal of Public Economics*, *88*(12), 2549 - 2572. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0047272703001415` doi: https://doi.org/10.1016/j.jpubeco.2003.08.003

Evans, D., & Popova, A. (2016). What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews. *The World Bank Research Observer*, *31*(2), 242–270.

Evans, D., & Yuan, F. (2017). *The economic returns to interventions that increase learning.* (mimeo)

Galiani, S., Gertler, P., & Schargrodsky, E. (2005). Water for life: The impact of the privatization of water services on child mortality. *Journal of political economy*, *113*(1), 83–120.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis*. Chapman & Hall/CRC Boca Raton, FL, USA.

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2*(3), 205–227.

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? textbooks and test scores in Kenya. *American Economic Journal: Applied Economics*, *1*(1), 112-35. Retrieved from `http://www`

.aeaweb.org/articles?id=10.1257/app.1.1.112 doi: 10.1257/app.1.1.112

Green, D. P., Ha, S. E., & Bullock, J. G. (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. Retrieved from http://dx.doi.org/10.1177/0002716209351526 doi: 10.1177/0002716209351526

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Disruption versus tiebout improvement: The costs and benefits of switching schools. *Journal of public Economics*, *88*(9), 1721–1746.

Hart, O., Shleifer, A., & Vishny, R. W. (1997). The proper scope of government: theory and an application to prisons. *The Quarterly Journal of Economics*, *112*(4), 1127–1161.

Heckman, J., & Pinto, R. (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. Retrieved from http://dx.doi.org/10.1080/07474938.2014.944466 doi: 10.1080/07474938.2014.944466

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, *103*(6), 2052-86. Retrieved from http://www.aeaweb.org/articles?id=10.1257/aer.103.6.2052 doi: 10.1257/aer.103.6.2052

Hsieh, C.-T., & Urquiola, M. (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, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological methods*, *15*(4), 309.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, *25*(1), 51–71.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning with applications in R.* (Springer)

Kerwin, J. T., & Thornton, R. (2015). *Making the grade: Understanding what works for teaching literacy in rural Uganda.* (mimeo)

Kiessel, J., & Duflo, A. (2014). *Cost-effectiveness report: The teacher community assistant initiative (TCAI).* Retrieved 6/08/2017, from http://www.poverty-action.org/sites/default/files/publications/TCAI_Cost-Effectiveness_2014.3.26.pdf

King, S., Korda, M., Nordstrum, L., & Edwards, S. (2015). *Liberia teacher training program: Endline assessment* *of the impact of early grade reading and mathematics interventions* (Tech. Rep.). RTI International.

Krasner, S. D., & Risse, T. (2014). External actors, state-building, and service provision in areas of limited statehood: Introduction. *Governance*, *27*(4), 545–567.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*(6130), 297–300. Retrieved from http://science.sciencemag.org/content/340/6130/297 doi: 10.1126/science.1235350

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, *91*(3), 437–456.

Kwauk, C., & Robinson, J. P. (2016). *Bridge International Academies: Delivering quality education at a low cost in Kenya, Nigeria, and Uganda* (Tech. Rep.). The Brookings Institution. Retrieved 09/08/2017, from http://www.bridgeinternationalacademies.com/wp-content/uploads/2016/09/Brookings-Millions-Learning-case-study.pdf

Ladner, P., & Persson, T. (2009). The origins of state capacity: Property rights, taxation, and politics. *The American Economic Review*, *99*(4), 1218–1244.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*(3), 1071–1102.

Lemos, R., & Scur, D. (2016). *Developing management: An expanded evaluation tool for developing countries.* (mimeo)

Liberia Institute of Statistics and Geo-Information Services. (2014). *Liberia demographic and health survey 2013*. Liberia Institute of Statistics and Geo-Information Services.

Liberia Institute of Statistics and Geo-Information Services. (2016). *Liberia - household income and expenditure survey 2014-2015*. Liberia Institute of Statistics and Geo-Information Services.

Loevinsohn, B., & Harding, A. (2005). Buying results? contracting for health service delivery in developing countries. *The Lancet*, *366*(9486), 676–681.

Lucas, A. M., & Mbiti, I. M. (2012). Access, sorting, and achievement: the short-run effects of free primary education in Kenya. *American Economic Journal: Applied Economics*, *4*(4), 226–253.

May, S. (2017). *Oral evidence: DFID's work on education: Leaving no one behind?, HC 639* (Tech. Rep.). House of Commons, International Development Committee.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Rajani, R., & Manda, C. (2017). *Inputs, incentives, and complementarities in primary education: Experimental evidence from Tanzania.* (Mimeo)

Meager, R. (2016). *Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit* *literature* (Tech. Rep.).

Ministry of Education - Republic of Liberia. (2016a). *Liberia education statistics report 2015-2106.*

Ministry of Education - Republic of Liberia. (2016b). *Memorandum of understanding between Ministry of Education, Government of Liberia and Bridge International Academies.* Retrieved 6/08/2017, from www.theperspective.org/2016/ppp_mou.pdf

Ministry of Education - Republic of Liberia. (2017). *Getting to best education sector plan, 2017-2021.*

Muralidharan, K., Niehaus, P., & Sukhtankar, S. (2016). Building state capacity: Evidence from biometric smartcards in India. *The American Economic Review*, *106*(10), 2895–2929.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2016). *Disrupting education? experimental evidence on technology-aided instruction in India* (Tech. Rep.). National Bureau of Economic Research.

Muralidharan, K., & Sundararaman, V. (2015). The aggregate effect of school choice: Evidence from a two-stage experiment in India. *The Quarterly Journal of Economics*, *130*(3), 1011. Retrieved from +http://dx.doi.org/10.1093/qje/qjv013 doi: 10.1093/qje/qjv013

Patrinos, H. A., Osorio, F. B., & Guáqueta, J. (2009). *The role and impact of public-private partnerships in education*. World Bank Publications.

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

Pritchett, L., & Woolcock, M. (2004). Solutions when the solution is the problem: Arraying the disarray in development. *World Development*, *32*(2), 191–212.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of educational and behavioral statistics*, *6*(4), 377–401.

Schermerhorn, J., Osborn, R., Uhl-Bien, M., & Hunt, J. (2011). *Organizational behavior*. Wiley. Retrieved from https://books.google.com/books?id=8eRtuZeIguIC

Singh, A. (2015a). *How standard is a standard deviation? a cautionary note on using sds to compare across impact evaluations in education.* Retrieved 31/07/2017, from http://blogs.worldbank.org/impactevaluations/how-standard-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations

Singh, A. (2015b). Private school effects in urban and rural india: Panel estimates at primary and secondary school ages. *Journal of Development Economics*, *113*, 16–32.

Singh, A. (2016). *Learning more with every year: School year productivity and international learning divergence.* (Mimeo)

Stallings, J. A., Knight, S. L., & Markham, D. (2014). *Using the stallings observation system to investigate time* *on task in four countries* (Tech. Rep.). World Bank.

Tuttle, C. C., Gleason, P., & Clark, M. (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.

Urminsky, O., Hansen, C., & Chernozhukov, V. (2016). *Using double-lasso regression for principled variable selection.* (Mimeo)

USAID. (2017). *Request for proposals - SOL-669-17-000004, Read Liberia.* Retrieved 6/08/2017, from https://www.fbo.gov/index?s=opportunity&mode=form&id=e53cb285301f7014f415ce91b14049a3&tab=core&tabmode=list&=

Useem, B., & Goldstone, J. A. (2002). Forging social order and its breakdown: Riot and reform in U.S. prisons. *American Sociological Review*, *67*(4), 499-525. Retrieved from http://www.jstor.org/stable/3088943

van der Linden, W. J. (2017). *Handbook of item response theory*. CRC Press.

Werner, G. K. (2017). Liberia has to work with international private school companies if we want to protect our children's future. *Quartz Africa*. Retrieved 20/07/2017, from https://qz.com/876708/why-liberia-is-working-with-bridge-international-brac-and-rising-academies-by-education-minister-george-werner/

Woodworth, J. L., Raymond, M., Han, C., Negassi, Y., Richardson, W. P., & Snow, W. (2017). *Charter management organizations* (Tech. Rep.). Center for Research on Education Outcomes, Stanford University.

World Bank. (2014). *Life expectancy.* (data retrieved from World Development Indicators, http://data.worldbank.org/indicator/SE.PRM.NENR?locations=LR)

World Bank. (2015a). *Conducting classroom observations: analyzing classrooms dynamics and instructional time, using the stallings' classroom snapshot'observation system. user guide* (Tech. Rep.). World Bank Group.

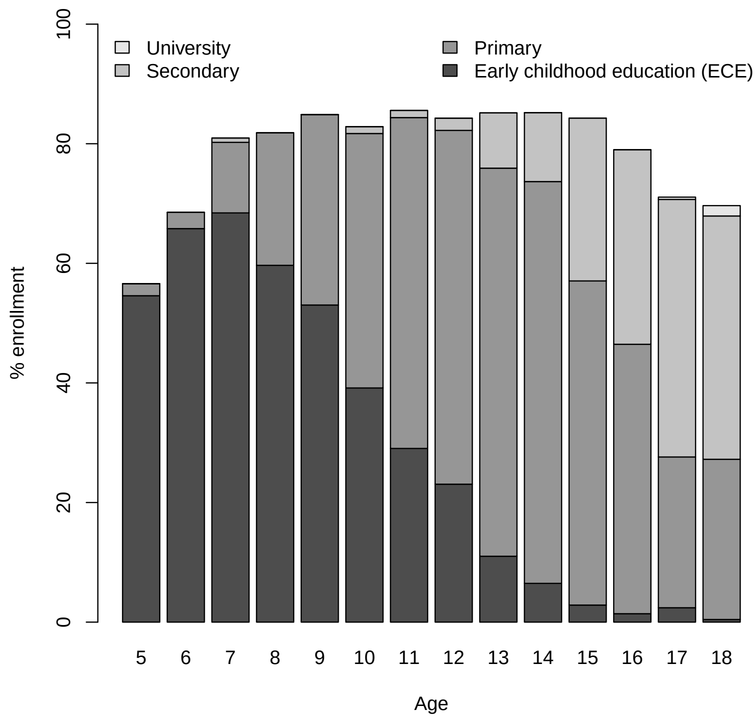
World Bank. (2015b). *GDP per capita (current US£).* (data retrieved from World Development Indicators, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)

World Bank. (2015c). *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)

Zhang, H. (2014). *The mirage of elite schools: evidence from lottery-based school admissions in China.* (Mimeo)

Figures

Figure 1: Enrollment by age



Note: Authors' calculations based on 2014 Household Income and Expenditures Survey.

Figure 2: What did providers do?

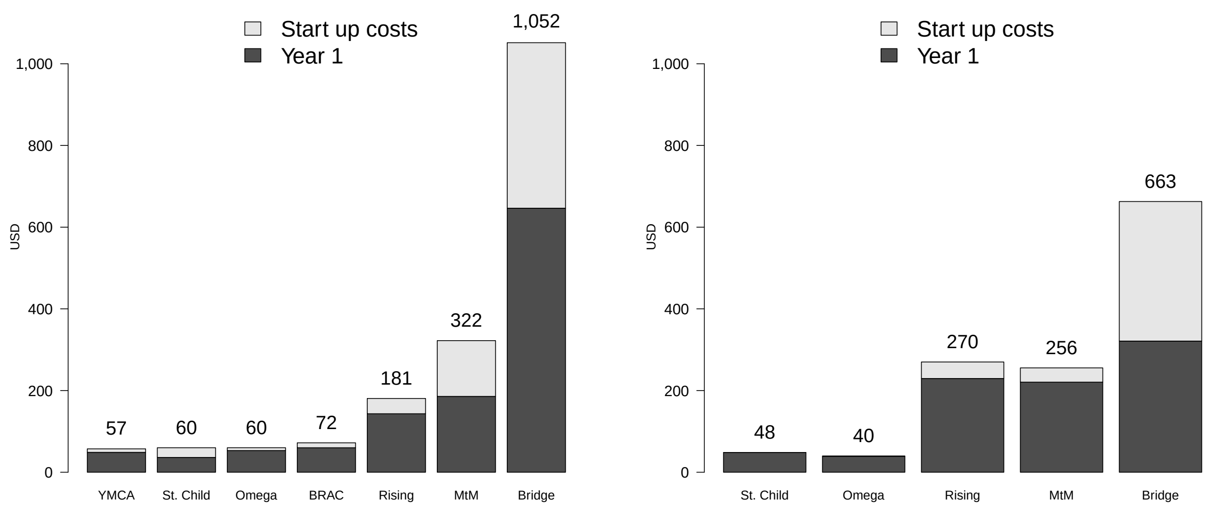
		Provider							
		Stella M	YMCA	Omega	BRAC	Bridge	Rising	St. Child	MtM
Provider Support	Provider staff visits at least once a week(%)	0	54	13	93	76	94	91	96
	Heard of PSL(%)	42	85	61	42	87	90	68	85
	Heard of provider(%)	46	96	100	95	100	100	100	100
	Has anyone from (provider) been to this school?(%)	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

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.

Figure 3: Budget and costs as reported by providers

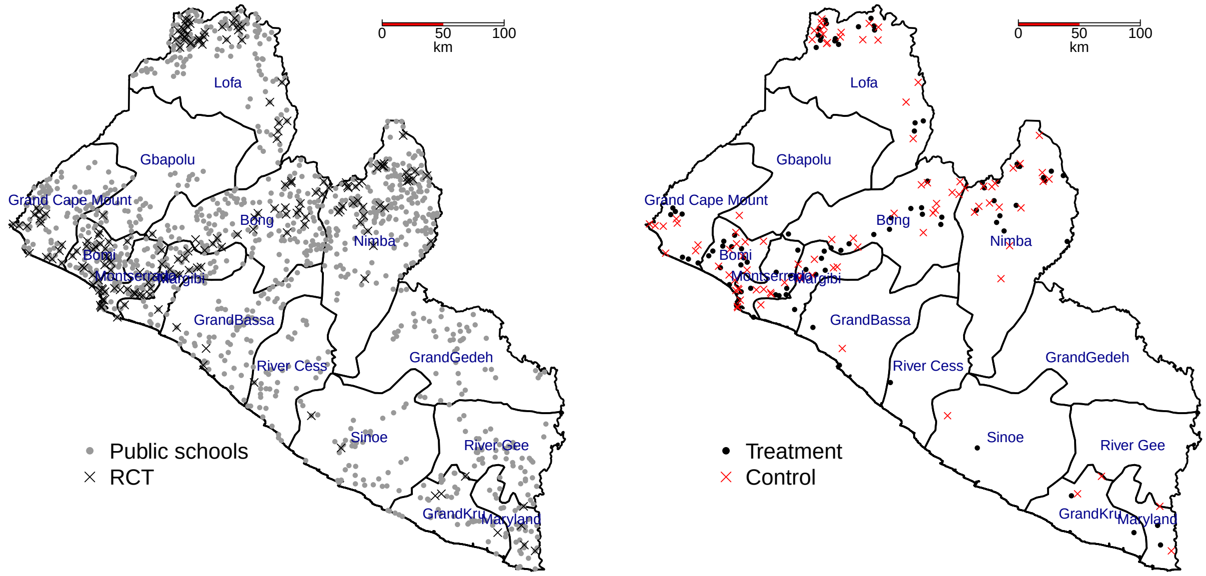
(a) Ex ante budget per pupil

(b) Ex post cost per pupil



Note: Numbers in 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 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.

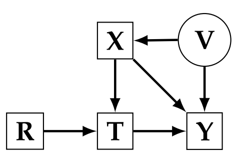
Figure 4: Public primary schools in Liberia



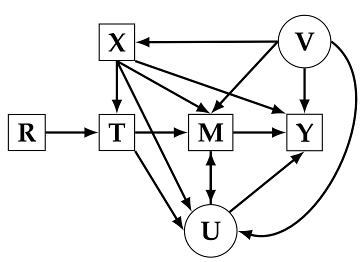
(a) Geographical distribution of all public schools in Liberia and those within the RCT.

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

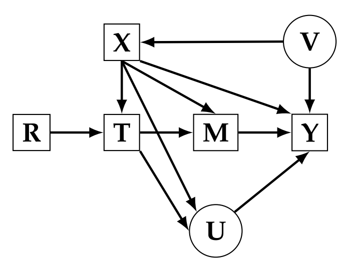
Figure 5: Causal relationships under different models



(a) Randomization



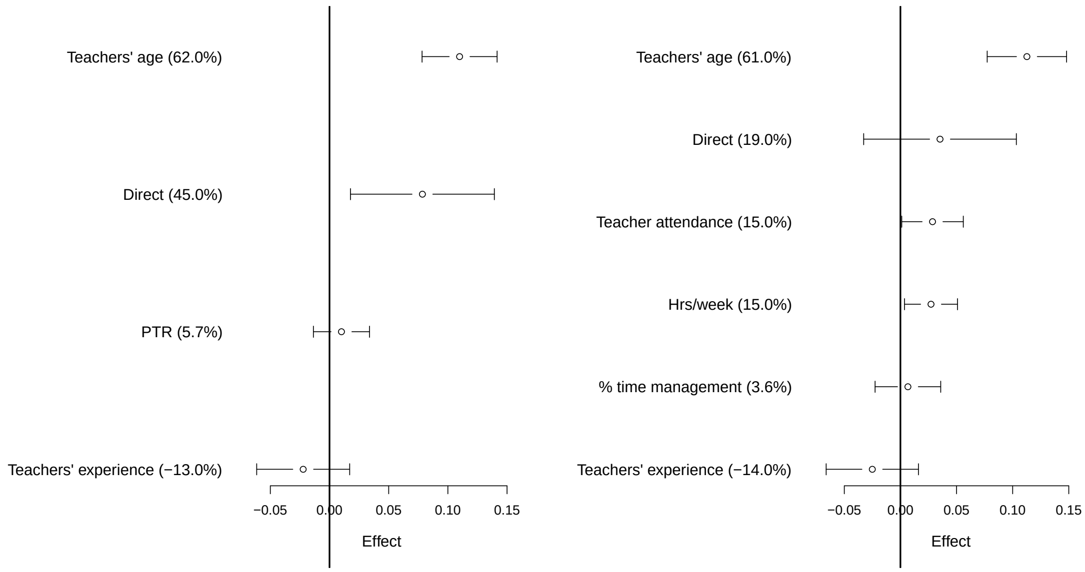
(b) Mediation



(c) Mediation under assumption 1

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.

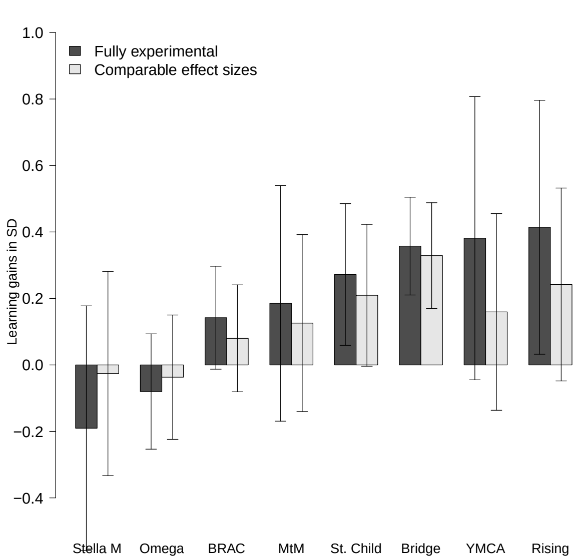
Figure 6: Direct and mediation effects



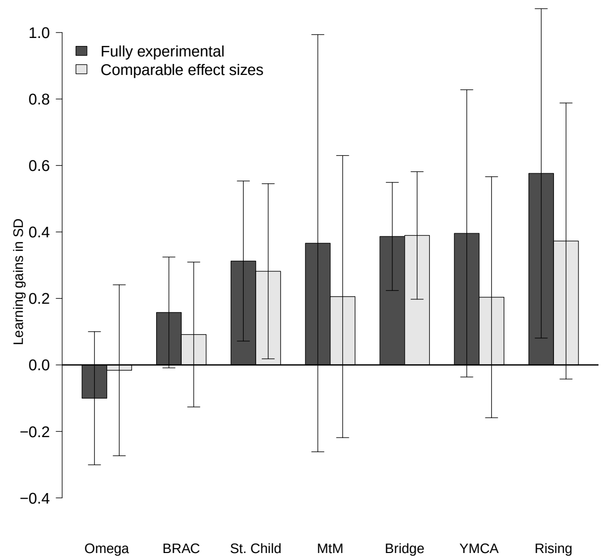
(a) Inputs (b) Inputs & Management

Note: Direct (β_5) and mediation effects ($\beta_4 \times \theta_5$) 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 6a shows treatment effects allowing only change in inputs as mediators. Panel 6b shows treatment effects allowing change in inputs and in the use of inputs as mediators.

Figure 7: Treatment effects by provider



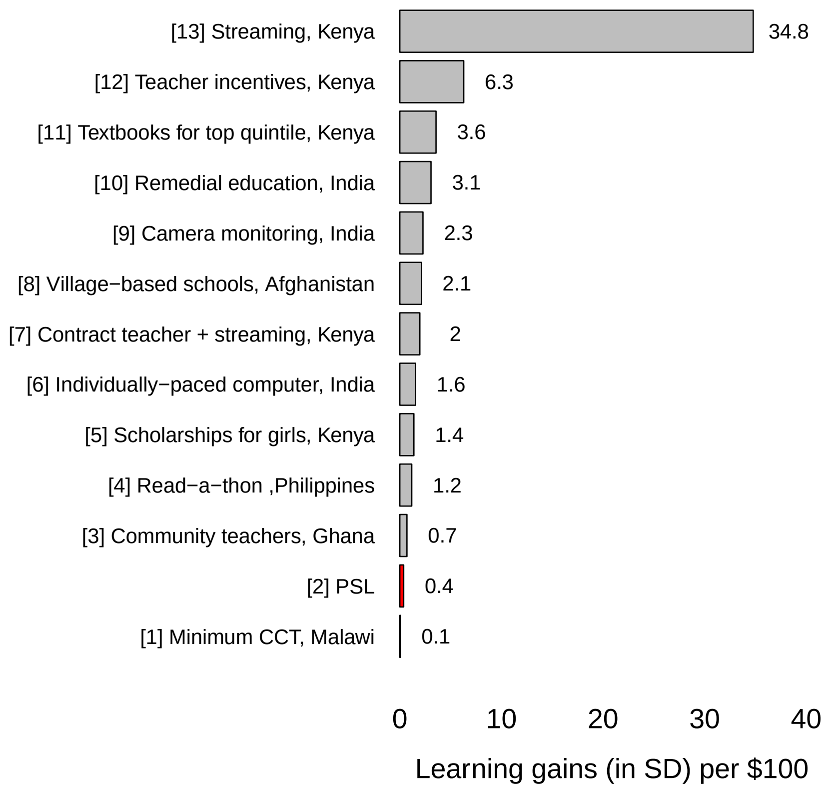
(a) Intention-to-treat (ITT) effect



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

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 7a shows the intention-to-treat (ITT) effect, while Figure 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.

Figure 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, Cole, Duflo, and Linden (2007); [8] Burde and Linden (2013); [9] Duflo, Hanna, and Ryan (2012); [11] Glewwe, Kremer, and Moulin (2009); [12] Glewwe, Ilias, and Kremer (2010).

Tables

Table 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 fund-raising
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.

Table 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)

The first wave of data was collected 2 to 8 weeks after the beginning of treatment; hence, the focus here is on time-invariant characteristics (some of these characteristics may vary in response to the program in the long run, but are time-invariant given the duration of our study). This table presents the mean and standard error of the mean (in parentheses) 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. Panel A has two measures of school infrastructure: The first is a school infrastructure index 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 second is whether the school ever needs to hold classes outside due to lack of classrooms. There are two measures of school rurality: First, a binary variable and second, the time it takes to travel by motorcycle to the nearest bank. Panel B has student characteristics. The wealth index is the first component of a principal component analysis of indicator variables for whether the student's household has a television, radio, electricity, a refrigerator, a mattress, a motorbike, a fan, and a phone. Panel C shows the attrition rate (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. The sample is the original treatment and control allocation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: ITT treatment effects on learning

	First wave (1-2 months after treatment)			Second wave (9-10 months after treatment)		
	Difference	Difference (F.E.)	Difference (F.E.+Controls)	Difference	Difference (F.E.)	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 modules				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).

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

Table 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 change	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.

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

Table 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).

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

Table 6: ITT treatment effects on inputs and resources

	(1) Treatment	(2) Control	(3) Difference	(4) 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)–Conditional on salary>0	121.36	104.54	16.82**	13.90***
	(44.42)	(60.15)	(6.56)	(4.53)
Panel C: Classroom observation (N = 143)				
Number of seats	20.64	20.58	0.06	0.58
	(13.33)	(13.57)	(2.21)	(1.90)
% with students sitting on the floor	2.41	4.23	-1.82	-1.51
	(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. The sample is the original treatment and control allocation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: ITT treatment effects on school management

	(1) Treatment	(2) Control	(3) Difference	(4) 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)
Principal's 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 **??**. 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 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 (N = 185)				
Teacher missed school previous week (%)	17.69	25.12	-7.43***	-7.55***
	(10.75)	(14.92)	(1.91)	(1.94)
Teacher never hits students (%)	54.71	48.21	6.50**	6.56***
	(18.74)	(17.06)	(2.63)	(2.52)
Teacher 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 (active + passive) (% of class time)	49.68	35.00	14.68***	14.51***
	(32.22)	(37.08)	(5.11)	(4.70)
Classroom 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 **??** has the results without imputing values. Standard errors are clustered at the school level. The sample is the original treatment and control allocation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: ITT treatment effects on household behavior, fees, and student attitudes

	(1) Treatment	(2) Control	(3) Difference	(4) Difference (F.E)
Panel A: Household behavior (N = 1,115)				
% satisfied with school	74.87	67.46	7.42**	7.44**
	(19.25)	(23.95)	(3.20)	(3.23)
% paying any fees	48.11	73.56	-25.45***	-25.69***
	(50.01)	(44.14)	(4.73)	(3.26)
Fees (USD/year)	5.72	8.04	-2.32**	-2.89***
	(10.22)	(9.73)	(0.96)	(0.61)
Expenditure (USD/year)	65.52	73.61	-8.09	-6.74
	(74.78)	(79.53)	(6.96)	(4.13)
Engagement index (PCA)	-0.11	-0.09	-0.02	-0.03
	(0.84)	(0.91)	(0.07)	(0.06)
Panel B: Fees (N = 184)				
% with >0 ECE fees	11.83	30.77	-18.94***	-18.98***
	(32.47)	(46.41)	(5.92)	(5.42)
% with >0 primary fees	12.90	29.67	-16.77***	-16.79***
	(33.71)	(45.93)	(5.95)	(5.71)
ECE Fee (USD/year)	0.57	1.42	-0.85**	-0.87***
	(1.92)	(2.78)	(0.35)	(0.33)
Primary Fee (USD/year)	0.54	1.22	-0.68**	-0.70**
	(1.71)	(2.40)	(0.31)	(0.31)
Panel C: Student attitudes (N = 3,492)				
School is fun	0.58	0.53	0.05**	0.05**
	(0.49)	(0.50)	(0.02)	(0.02)
I use what I'm learning outside of school	0.52	0.49	0.04	0.04***
	(0.50)	(0.50)	(0.02)	(0.02)
If I work hard, I will succeed.	0.60	0.55	0.05*	0.04***
	(0.49)	(0.50)	(0.03)	(0.02)
Elections are the best way to choose a president	0.90	0.88	0.03*	0.03***
	(0.30)	(0.33)	(0.01)	(0.01)
Boys are smarter than girls	0.69	0.69	-0.00	0.01
	(0.46)	(0.46)	(0.02)	(0.01)
Some tribes in Liberia are bad	0.76	0.79	-0.03	-0.03**
	(0.43)	(0.41)	(0.02)	(0.01)

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. The index for parent engagement is the first component from a principal component analysis across several measures of parental engagement; see Table **??** for details.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: 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)
% exp. in private schools			-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)
% of time spent 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)
Index of good practices (PCA)					0.079***
					(0.024)
Student attendance					-0.048
					(0.081)
Instruction (Classroom obs)					-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

The independent variable in all regressions is the composite IRT score across all test items. All dependent variables are standardized to have mean zero and standard deviation of 1 except the treatment dummy. Column 1 replicates the results from Table 3 and columns 2 and 3 include only raw inputs. Columns 4 and 5 include raw inputs and the use of these inputs. Column 2 and column 4 only include mediators selected via "Double Lasso", and columns 3 and 5 include all the mediators. The dependent variable is the composite test score (IRT score using both math and English questions). 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 11: Baseline differences between treatment schools and average public schools, by provider

	(1) BRAC	(2) Bridge	(3) YMCA	(4) MtM	(5) Omega	(6) Rising	(7) St. Child	(8) Stella M	(9) 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Comparable ITT treatment effects by provider

	(1) BRAC	(2) Bridge	(3) YMCA	(4) MtM	(5) Omega	(6) Rising	(7) St. Child	(8) Stella M	(9) 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)	
Test score in standard deviations (teachers)	0.12	0.24*	0.23	0.17	0.17	0.23	0.17	0.05	0.46
	(0.13)	(0.14)	(0.18)	(0.13)	(0.18)	(0.16)	(0.18)	(0.23)	
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)	
Panel D: Satisfaction									
% satisfied with school (parents)	11.64*	10.98*	3.72	1.70	2.44	-0.63	9.97	8.54	0.23
	(6.31)	(6.40)	(8.40)	(6.32)	(9.06)	(8.44)	(9.40)	(9.18)	
% students who think school is fun	4.04	2.68	2.47	3.24	3.48	2.61	-0.02	4.80	0.59
	(3.89)	(3.64)	(5.41)	(4.06)	(5.64)	(4.64)	(6.64)	(6.20)	
Observations	40	45	8	12	38	10	24	8	

This table presents the ITT treatment effect for each provider, after adjusting for differences in baseline school characteristics, based on a Bayesian hierarchical model. Thus, this number should be interpreted as the difference between treatment and control schools, not as the mean in treatment schools. Column 9 shows the p-value for testing $H_0 : \beta_{BRAC} = \beta_{Bridge} = \beta_{YMCA} = \beta_{MtM} = \beta_{Omega} = \beta_{Rising} = \beta_{St.Child} = \beta_{StellaM}$. Some operators had no schools with class sizes above the caps. Table **??** in Appendix **??** has the raw experimental treatment effects by provider. Standard errors are shown in parentheses. Estimation is conducted on collapsed, school-level data. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$