Beyond Short-term Learning Gains: The Impact of Outsourcing Schools in Liberia after Three Years

Mauricio Romero and Justin Sandefur

Abstract

After one year, outsourcing the management of ninety-three randomly-selected government primary schools in Liberia to eight private operators led to modest learning gains (Romero, Sandefur, & Sandholtz, in press). In this paper, we revisit the program two years later. After the first year, treatment effects on learning gains plateaued (e.g., the intention-to-treat effect on English was .18σ after one year, and .16σ after three years, equivalent to 4 words per minute additional reading fluency for the cohort that started in first grade). Looking beyond learning gains, the program reduced corporal punishment (by 4.6 percentage points from a base of 51%), but increased dropout (by 3.3 percentage points from a base of 15%) and failed to reduce sexual abuse. Behind these average effects, the identity of the contractor mattered. Despite facing similar contracts and settings, some providers produced uniformly positive results, while others present stark trade-offs between learning gains, access to education, child safety, and financial sustainability.

Keywords: Public-private partnership; randomized controlled trial; school management

JEL: I25, I28, C93, L32, L33

Center for Global Development
www.cgdev.org

Working Paper 521
December 2019

Beyond Short-term Learning Gains:
The Impact of Outsourcing Schools in Liberia after Three Years

Mauricio Romero
ITAM
mtromero@itam.mx

Justin Sandefur
Center for Global Development

We are grateful to the Minister George K. Werner and his team, Minister Prof. Ansu. D Sonii, Sr and his team, the Partnership Schools for Liberia (PSL) team, Susannah Hares, Robin Horn, and Joe Collins from Ark EPG, and the team at Social Finance for their commitment throughout this project to ensuring a rigorous and transparent evaluation of the PSL/LEAP program. We are especially grateful to Wayne A. Sandholtz for his collaboration in the early stages of this project and subsequent discussions. Thanks to Arja Dayal, Dackermue Dolo, and their team at Innovations for Poverty Action who led the data collection. Avi Ahuja, Miguel Ángel Jiménez-Gallardo, Dev Patel, Rony Rodríguez-Ramírez, and Benjamin Tan provided excellent research assistance. We are grateful to Laura Johnson who provided guidance on the sexual violence survey module and protocol. 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, Aestus Trust, and the UK's Economic and Social Research Council (grant number ES/P00604). Romero acknowledges financial support from the Asociación Mexicana de Cultura. The views expressed here are ours, and not those of the Ministry of Education of Liberia or our funders. All errors are our own.

Mauricio Romero and Justin Sandefur, 2019. "Beyond Short-term Learning Gains: The Impact of Outsourcing Schools in Liberia after Three Years." CGD Working Paper 521. Washington, DC: Center for Global Development. www.cgdev.org/publication/beyond-short-term-learning-gains-impact-outsourcing-schools-liberia-after-three-years

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

202.416.4000
(f) 202.416.4050

www.cgdev.org

1 Introduction

Public-private partnerships in education are common around the world (Patrinos, Barrera-Osorio, & Guáqueta, 2009; Aslam, Rawal, & Saeed, 2017). While canonical results from contract theory might suggest education is well suited for outsourcing (Hart, Shleifer, & Vishny, 1997), the required assumptions may not always hold in practice. Governments may struggle to observe key outputs (e.g., student selection, student attitudes, and non-cognitive skills), and in remote rural areas where alternatives are limited, beneficiaries may be unable to opt out of outsourced services. Furthermore, a narrow focus on raising test scores as the primary goal of public-private partnerships in education may result in a multi-tasking problem (Holmstrom & Milgrom, 1991), leading to quality shading in other important aspects (e.g., access and child safety). While dynamic incentives in repeated contracting may overcome some of these pitfalls (Banerjee & Duflo, 2000; Corts & Singh, 2004), weakly-governed states may be unwilling or unable to sanction bad performance by private contractors. In this paper, we test these hypotheses in a large-scale field experiment of a program that outsourced management of existing public schools to private providers, and report on results after one and three years.

In 2016, the Liberian government outsourced the management of 93 randomly-selected public schools, comprising 8.6% of public school students, to eight different private operators, including a mix of both for-profit companies and non-profit charities, as well as local and international organizations. The program bundled private management with, in theory, a doubling of education expenditure per child. After one academic year, Romero et al. (in press) found modest improvements in learning outcomes and several important caveats: students in outsourced schools scored 0.18σ higher in English and mathematics, but costs far surpassed original projections, and some providers engaged in unforeseen and harmful behavior, including mass removal of students and efforts to conceal serious sexual abuse allegations, complicating any assessment of long-term welfare gains.

Returning after three academic years, there are several reasons to expect that the program's performance would have improved over time. Both learning-by-doing and selective contract renewal — in which successful operators were rewarded with more schools — point toward larger impacts on the program's primary goal of raising learning outcomes. On non-learning outcomes, the earlier evaluation results summarized in Romero et al. (in press) and media coverage of various program failings in the first year provided opportunities to revise contracts and mitigate unintended consequences in later years. Finally, several private operators disputed any cost-benefit analysis in the first year on the grounds that

start-up investments had inflated their costs, and that unit costs would fall rapidly in subsequent years as the program expanded. On the other hand, the enormous media scrutiny of this program in its first year — including on-the-ground reporting by *The New York Times* (Rosenberg, 2016; Tyre, 2017; Kristof, 2017), *Financial Times* (Pilling, 2017), The Economist (2017), and other international outlets — dissipated in the second and third years, which might have reduced the pressure for good performance.

In this paper, we study whether the short-term learning effects endure or increase after three years. However, since the program had a narrow goal of improving test scores, this may create moral hazard and contractual incompleteness in relation to less-emphasized outcomes. Thus, we also measure the program's impact on drop-out rates. In light of widely-reported incidents that have led to criminal proceedings involving staff of two of the providers, we measure whether the program had any effect on rates of either corporal punishment or sexual abuse in schools.

Two features of the experimental design merit special emphasis before we turn to the results: block randomization and intention-to-treat analysis. First, because we randomized treatment within matched pairs, our design amounts to eight internally-valid experiments, one per provider. Hence, we are able to study heterogeneity across providers, which is central to any theoretical interpretation of our results. Second, to avoid confounding the treatment effect of the program with sorting of students across schools, we sampled students from pre-treatment enrollment records and follow them for three years (we were able to interview over 96% of the original sample three years later). Thus, by assigning each student to their "original" school, regardless of what school (if any) they attend in later years, we are able to provide unbiased estimates of the program's intention-to-treat (ITT) effect on various outcomes.

We find that treatment effects on test scores remain statistically significant after three years, but plateau after the first year. The intention-to-treat (ITT) treatment effect of outsourcing after one academic year was $.18\sigma$ on English and $.18\sigma$ on math. After three years, the ITT treatment effects on English and math were $.16\sigma$ (p-value < 0.001) and $.21\sigma$ (p-value < 0.001) — equivalent to 4 words per minute additional reading fluency for the cohort that started in first grade. The ITT effect combines the treatment effect of the program on students who remained in partnership schools for three years, different degrees of student level non-compliance (students who were exposed for zero, one, or two years because either they change schools or dropped out of school), persistence of the treatment effect for students who graduated from primary school, and school level non-compliance (i.e., providers not taking control of some of the schools assigned to them). The treatment-on-the-treated (ToT) effect is 0.26σ for English (p-value < 0.001) and 0.35σ for math (p-value < 0.001).

A possible explanation for why treatment effects on test scores plateau is that the effects on teacher behavior also dissipate after the first year. While teachers were more likely to be in school and more likely to be in a classroom during a spot check after the first year — 20 percentage points (p-value < 0.001) and 15 percentage points (p-value < 0.001) more likely, respectively — this is no longer true after three years when the treatment effects are smaller and insignificant — 6.8 percentage points (p-value .12) and 7.3 percentage points (p-value .1) more likely, respectively. This does not seem to be driven by control schools improving teacher attendance since this remains relatively constant across time. We can only speculate about the reasons for the change in teacher behavior in treatment schools from year one to year three, but there are at least four possible explanations: a) teachers' enthusiasm for the program waned over time; b) teachers realized that providers lacked "sticks" to use against them since they remained unionized and on the government payroll; c) providers failed to provide credible "carrots" (e.g., promises to get teachers on payroll and guarantee regular payments that were not fulfilled in many cases); and d) providers exerted less effort after the initial surge in media and donor scrutiny subsided.

Turning to non-learning outcomes, outsourcing increased dropout rates among the students originally in partnership schools. Negative effects on pupil-level enrollment emerged in the first year due to mass expulsions by one private operator, Bridge International Academies. In schools where enrollment was already high and class sizes were large at baseline, the program led to a significant decline in enrollment Romero et al. (in press). However, most of these students were absorbed by nearby traditional public schools. After three years, the picture is somewhat different: the negative effects are significant in the whole sample, not just a sub-sample of schools, and pupils are not absorbed by other schools. Students enrolled in partnership schools at baseline in 2015/2016 are 3.3 percentage points (p-value .0042) less likely to be enrolled in any school after three years (from a base of 85%). This negative effect is again explained by Bridge International Academies, but is not driven by students who were removed from their schools in the first year due to large class sizes. Instead, the effect seems to be driven by older students (who are less likely to enroll in secondary school after they finish primary school in treatment schools) and by girls (who are more likely to report dropping out due to pregnancy in treatment schools).

The outsourcing program did not reduce (or increase) rates of sexual violence perpetrated by school staff, as reported in our data. Prior studies have found that sexual violence is widespread in Liberian schools (Postmus et al., 2015; Steiner, Johnson, Postmus, & Davis, 2018). However, the issue was not highlighted in the design of the outsourcing program. In 2018, a journalistic investigation revealed that one of the private operators in the program, More Than Me Academies, had concealed the full extent of

a sexual abuse scandal involving the rape of as many as dozens of students by its co-founder prior to the launch of the program (F. Young, 2018). A subsequent sexual abuse case reported by news media involving another provider working in a non-program school heightened concerns about child protection under the outsourcing program (Baysah, 2016). Both of these incidents occurred prior to the launch of the program, but were revealed in full only after the program launched. In response to these concerns, a survey module was added in 2019 to measure self-reported experiences of sexual abuse by pupils. Across treatment and control schools, 5.0% of pupils reported sexual intercourse with a teacher since coming to their current school, and 7.4% reported inappropriate physical contact of any sort.[1] On average, the program had no impact on either measure of abuse, and small but positive (i.e., higher rates) and statistically significant impacts on reported rates of forced sex by fellow students and family members.

Average effects conceal very different results across private operators, both in learning gains and other dimensions. The positive treatment effect on learning reflects an underlying null effect for three of eight operators (BRAC, Omega Schools, and Stella Maris), and ITT effects of roughly 0.4σ for the remaining five. On access, the treatment effect on the probability of still being in school is negative for six of eight operators, but only statistically significant in the case of Bridge International Academies, which posted a negative effect of 6.5 percentage points. In the case of sex with teachers, most effects are close to zero and statistically insignificant, though one operator, BRAC, showed a significant reduction in abuse of 3.7% percentage points. Impacts across dimensions are not perfectly correlated across operators. While some operators produce uniformly positive results, even on non-contracted outcomes, others present stark trade-offs between learning gains and other goals.

Consistent with the idea that high costs in the first year were driven by start-up investments and fixed costs that would decline in importance as the program grew, unit costs have fallen since the first year, but remain above the original budget projections. The ministry expects providers to operate for USD 50 per pupil or less, which it deemed a realistic medium-term increase in the budget for the education sector. In the first year, the average expenditure was roughly USD 300 per pupil, with some providers spending the target amount (50 USD per pupil) and other spending over USD 600 per pupil. After three years, the average (self-reported) expenditure has fallen to USD 119 per pupil. However, Bridge International Academies and More Than Me continue to spend at least three times as much as the government target.

Our results are closely related to the literature on outsourcing education, most of which comes from

[1] Because these figures cover any cases since the pupil arrived at the school, some incidents will pre-date the partnership schools program. However, randomization should ensure balance on pre-treatment levels of sexual abuse, and thus any differences between treatment and control schools will be attributable to the program.

U.S. charter schools (see Betts and Tang (2014) and Chabrier, Cohodes, and Oreopoulos (2016) for reviews). We join the growing literature studying outsourcing in other settings (e.g., Barrera-Osorio et al. (2017) in Pakistan, Barrera-Osorio (2007) and Bonilla (2010) in Colombia, and Eyles and Machin (2019) in the U.K.).

Two important features of Liberia's outsourcing initiative, which speak to its external validity and policy relevance, are that it provides evidence from a setting with low state capacity and where teachers in outsourced schools remain civil servants. Reliance on civil service teachers rather than non-unionized contract workers as in many similar programs, addresses a core element of the political economy underlying public-private partnerships in education. Low state capacity – both in delivering public services, and in managing private contractors – is a feature of many developing countries that may seek a way to improve education via outsourcing.

2 Research design

Below we summarize the most important features of the program and the experimental design. Further details are provided in Romero et al. (in press).

2.1 The program

2.1.1 Context

The government's primary, stated motivation, for the outsourcing program was the low level of learning outcomes in public schools. According to Demographic and Health Survey data, only 25% of adult Liberian women who had completed primary school were literate (Liberia Institute of Statistics and Geo-Information Services, 2014), one of the lowest rates for any country in the world. Our baseline data shows that roughly 25% of pupils enrolled in fifth grade could read zero words.

In addition to low learning levels, access remains an unresolved issue. The last nationally representative household survey prior to the experiment found net primary enrollment of 38%. This low rate is partially explained by high levels of over-age enrollment, with 60% of fifteen-year-olds still in primary school and many young children not enrolled (World Bank, 2014).

The experimental sample analyzed below is not intended to be representative of the country as a whole. Circa 2016, Liberia had 2,619 public primary schools across fifteen counties. To take part in the pilot, schools were required to meet minimum infrastructure standards. While 13 counties were included in the pilot, only 299 schools satisfied all the criteria. Finally, private providers were allowed to filter the

list of potential pilot schools before random assignment based on proximity to roads and availability of 3G service, leaving a final sample of 185 eligible schools.

Public primary school is nominally free in Liberia, though informal fees are common. In contrast, fees are officially permitted for pre-primary classes. At baseline, government spending on public primary schools was roughly $50 (USD) per pupil, almost entirely devoted to teacher salaries. Yet, in our sample of public schools, 36% of teachers were not on the government payroll at baseline. Anecdotally, many received little or no cash payment whatsoever, and relied on in-kind contributions from the community. Separately, existing private schools — which were excluded from the outsourcing program — educated 29.6% of primary-age children as of 2016 (Ministry of Education - Republic of Liberia, 2015-2016).

2.1.2 Intervention

The Liberian Education Advancement Program (LEAP) — formerly known as the Partnership Schools for Liberia (PSL) program — is a public-private partnership (PPP) for school *management*. Under the program, the government delegated the management of 93 public schools, representing 8.6% of all public school students, to eight different private organizations. Providers were paid on a per-pupil basis (with some exceptions noted below) and forbidden from charging fees or screening students based on ability.

Of the eight private organizations, three are for-profit companies: Bridge International Academies (which was allocated 23 schools in the pilot), Omega Schools (allocated 19 schools), and Rising Academies (allocated 5 schools). The other five non-profit providers include BRAC (allocated 20 schools), Street Child (allocated 12 schools), More than Me (allocated 6 schools), Youth Movement for Collective Action (allocated 4 schools), and Stella Maris (allocated 4 schools). Youth Movement for Collective Action and Stella Maris are Liberian organizations, the other six are international. While Stella Maris never actually took control of their assigned schools, the government still considers them part of the program (e.g., they were allocated more schools in an expansion of the program not studied in this paper (Ministry of Education - Republic of Liberia, 2017b)).

In contrast to some other public-private partnerships in education (e.g., U.S. charter schools), the teachers in the Liberian public schools which were outsourced to private providers were to remain civil servants and were still paid directly by the Liberian government. While similar initiatives elsewhere — from charter schools in the U.S. to contract schools in Punjab, Pakistan — often allow private actors to hire teachers with more flexible contracts or lower wages, the Liberian providers were supposed to improve teacher performance with a fixed corps of teachers, and without additional contractual or monetary incentives.

Below we discuss efforts by private providers to enroll onto the government payroll some teachers who were, at baseline, employed in public schools but not paid by the government, which we interpret as a treatment effect of the program.

There are three noteworthy features of the evolution of the intervention since it started in 2016. In 2017, the program expanded to an additional 98 schools. These schools were mostly located in the southeast (the most disadvantaged part of the country) and were not experimentally assigned nor embedded into the randomized evaluation (see Figure A.1). Thus, the results in this report do not speak to the treatment effect in these schools.[2] Second, the program changed some of its operating rules. All providers were given uniform contracts (unlike the first year, when Bridge International Academies had a different contract) and the Ministry of Education did not allow capping class sizes; however, in the experimental sample this had little effect as student expulsions in the first year meant few classes remained above the size cap. Finally, the country had a presidential election in late 2017.[3] The new administration which took office in early 2018 says that it has stopped prioritizing partnership schools in the assignment of teachers or in the process of bringing existing teachers onto the payroll.

2.1.3 What do providers do?

Providers must teach the Liberian national curriculum, but beyond that they have considerable flexibility in defining the intervention. They may choose to use school resources in different ways (e.g., providing remedial programs, prioritizing subjects, having longer school days, or other non-academic activities). They can also provide more inputs such as extra teachers, books, or uniforms, as long as they pay for them. To get some insights on what *actually* happened in treatment schools we administered a survey module to teachers, asking if they had heard of the provider, and if so, what activities the provider had engaged in. We summarize teachers' responses in Figure 1, which shows considerable variation in the specific activities and the total activity level of providers.

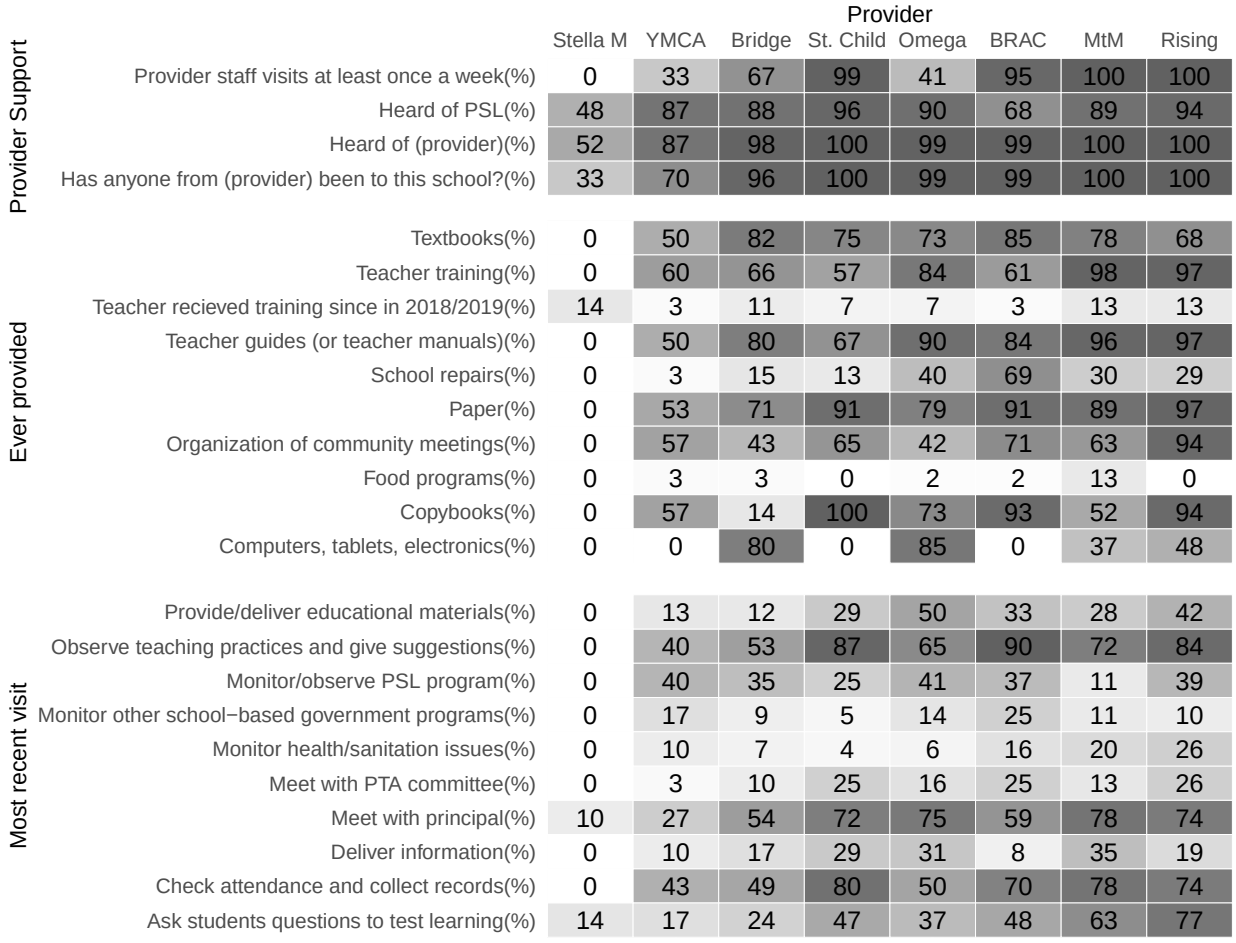
Many providers visit their schools on a weekly basis (Street Child, BRAC, More Than Me, and Rising Academies). Some organizations rely heavily on teacher training and teacher guides (More Than Me and Rising). Others emphasize observing teachers in the classroom and providing feedback (Street Child, BRAC, and Rising). The data confirms that Stella Maris never actually took control of their schools. The

[2] BRAC was assigned an additional 13 schools, Bridge International Academies was assigned 43 more schools, the Liberian Youth Network (previously known as the Youth Movement for Collective Action) was assigned 2 more schools, More Than Me was assigned 12 more schools, Omega Academies was assigned 2 more schools, Street Child was assigned 11 more schools, Stella Maris was assigned 4 more schools, and Rising Academies was assigned 11 more schools.

[3] Sandholtz (2019) examines the political economy question of whether the policy created incentives for politicians to adopt it.

data also confirms that Omega and Bridge provided computers to schools, which fits with the stated approach of scripted lessons through tablets of these two firms. Rising Academies is the organization that engages the local community the most.

Figure 1: What did providers do?



The figure reports simple proportions (not treatment effects) of teachers surveyed in parthernship schools who reported whether the provider responsible for their school had engaged in each of the activities listed. The sample size, n, of teachers interviewed for each provider is: Stella Maris, 21; Omega, 125; YMCA, 30; BRAC, 128; Bridge, 123; Street Child, 75; Rising Academy, 31; More than Me, 46. This sample only includes compliant treatment schools.

2.1.4 Cost data and assumptions

On paper, the Ministry of Education's financial obligation to partnership schools is the same as to any other government-run school: It provides teachers and maintenance, valued at about USD 50 per student on average nationwide. In addition, providers receive *extra* funding (of USD 50 per student), coordinated

by the Ministry of Education but paid by third-party philanthropies. Providers have complete autonomy over the use of these funds (e.g., they can be used for teacher training, school inputs, or management personnel). On top of that, providers may raise more funds on their own (see Section 4 for more details on providers' costs).

2.2 Experimental design

2.2.1 Sampling and random assignment

Two key features of the sampling and randomization process are that (a) providers agreed to a list of schools they would be willing to serve before random assignment took place, and (b) pupils were sampled from lists made before the program began and tracked regardless of where they went.

As noted above, private providers and the government agreed on a list of 185 eligible program schools (out of 2,619 public primary schools).[4] Based on providers' preferences and requirements schools were non-randomly partitioned across providers. The eligible schools allocated to each provider were then paired based on their infrastructure. Finally, within each pair schools were randomly assigned to treatment or control. Private providers did not manage all the schools originally assigned to treatment and we treat these schools as non-compliant, presenting results in an intention-to-treat framework (Table A.1 provides more details on compliance).

Treatment assignment may change the student composition across schools. To prevent differences in the composition of students from driving differences in outcomes, we sampled 20 students per school (from K1 to grade 5) from enrollment logs from the 2015/2016 school 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 assignment at the school level with measuring outcomes of a fixed and comparable pool of students allows us to provide unbiasedestimates of the program's intention-to-treat (ITT) effect within the student population originally attending study schools, uncontaminated by selection.[5]

[4] Schools in the RCT generally have better facilities and infrastructure than most schools in the country, which limits the external validity of the results. Romero et al. (in press) provides more details on the differences between schools in the experiment and other public schools.

[5] As a consequence of this design, we are unable to study the effect on students who were previously out-of-school and who may have decided to enroll in school due to the program.

2.2.2 Timeline of research and intervention activities

We collected data in schools three times: At the beginning of the school year in September/October 2016, at the end of the school year in May/June 2017, and in March/April of 2019. The focus of this paper is on the results from the third round of data collection, but often we will place the results from the second round of data collection alongside the results from the third round to study the evolution of treatment and control schools. Figure A.2 provides a detailed description of the research activities.

2.2.3 Test design

The test design during the third round of data collection is similar to the one in the previous two rounds, described in Romero et al. (in press). Specifically, we conduct one-on-one tests in which an enumerator sits with the student, asks questions, and records the answers since literacy cannot be assumed at any grade level. We use a single adaptive test for all students, regardless of the grade. The test has stop rules that skip higher-order skills if the student is not able to answer questions related to more basic skills. We estimate an item response theory (IRT) model for each round of data collection. 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. For a randomly selected class within each school, we conducted a classroom observation using the Stallings Classroom Observation Tool (World Bank, 2015). Furthermore, we conducted school-level surveys to collect information about school facilities, the teacher roster, input availability (e.g., textbooks), and expenditures.

During the third round of data collection, we added two additional modules to the survey. In schools, we ask Development World Management Survey (DWMS) style-questions Lemos and Scur (2016).[6] While the DWMS is performed with open-ended questions that are then scored against a rubric, we opted to ask multiple-choice questions based on the DWMS scoring rubric. Thus, our management index based on the DWMS is not directly comparable to measures from other countries.

Finally, given the concerns about child safety in partnership schools raised by the sexual abuse scandals involving two of the private providers, we added a sexual violence module to the student survey. Sexual abuse is inherently difficult to measure, and is rarely reported through official channels in Liberian schools.

[6]The DWMS webpage is available at the following link: https://developingmanagement.org/

We collected data via an anonymous survey.[7] In short, enumerators asked the student questions regarding sexual abuse at school (by teachers and peers) and at home, the student filled in an anonymous answer sheet (pre-filled with the school id and the gender of the child) and placed it in a closed ballot box. The survey was adapted to make it appropriate for children in the Liberian context. Section B provides more details on the survey protocol and survey instrument.

2.2.5 Balance and attrition

Romero et al. (in press) shows that time-invariant characteristics of both schools and students, as well as pre-treatment administrative measures of school characteristics are balanced across treatment and control schools.

Given our study design, we put considerable effort and resources into minimizing attrition (extra training, generous tracking time, and specialized tracking teams). 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%. Attrition in the third wave of data collection is balanced between treatment and control and is below 3% (see Table 1).[8]

Table 1: % Interviewed

	Year 1		Year 3	
	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)
% interviewed	95.74	-0.35	97.19	-0.47
	(20.20)	(0.44)	(16.53)	(0.41)
Observations	1,761	3,511	1,780	3,622

Notes: This table presents the attrition rate (proportion of students interviewed in Year 1 and Year 3). Columns 1 and 3 show the mean and standard deviation (in parentheses) for the control in Year 1 and Year 3. The differences between treatment and control for Year 1 and Year 3 are presented in Column 2 and Column 4, with standard errors (in parenthesis). The differences take into into account the randomization design (i.e., including "pair" fixed effects). Standard errors are clustered at the school level.

[7] We ask students about sexual abuse via three different methods: An anonymous survey with the enumerator reading the questions directly to the student, list randomization, and an anonymous survey where the enumerator reads the questions to the full class. In a companion paper, Johnson, Romero, Sandefur, and Sandholtz (2019) compare the relative effectiveness of different methods to measure sexual abuse in primary schools in Liberia. Overall, students do not seem to understand list randomization. The two anonymous surveys portray similar levels of sexual abuse. In this paper, we focus on the first method since it was performed in all schools (the third method was not).

[8] We sampled more students during the second wave of data collection from 2015/2016 enrollment logs. Thus, to calculate attrition for the second wave of data collection we only consider students sampled during the first wave. To calculate attrition for the third wave of data collection we consider students sampled during the first and the second wave.

3 Overall policy impact

In this section, we estimate the overall impact of the Liberian government's outsourcing program — aggregating results across eight operators — before turning in the following section to the impact of specific providers. In both cases, we focus on four margins: 1) *access*, defined as impacts on enrollment and grade attainment for a fixed sample of pupils; 2) *learning*, as measured by test scores; 3) *sustainability*, which hinges, in part, on whether the program effects come from increases in material inputs or staffing versus improvements in school management, as well as on the degree of negative externalities imposed on the broader system; and 4) *child safety*, as measured by corporal punishment and sexual abuse.

3.1 Access

The program reduced enrollment and increased dropout for the sample of students originally enrolled in partnership schools. At the same time, the program had a positive treatment effect on total enrollment in partnership schools, which implies they pulled in new students from outside our baseline sample.[9]

First, we focus on school-level enrollment. After three years, the program had a positive treatment effect on enrollment — a net increase of 36 students per school (p-value .052). This treatment effect comes from enrollment shrinking less in treatment schools since overall enrollment fell in both treatment and control schools during this period.

Provider compensation is based on the number of students enrolled rather than the number of students actively attending school. Yet, student attendance (measured during a spot-check by our enumerators) is higher in partnership schools by 11 percentage points (p-value .017, see Panel A, Table 2).

While the treatment effect on enrollment increased between 2016 and 2019, there was a decrease in the treatment effect on attendance. This can be explained by control schools increasing student attendance from 33% to 42% in this period.

A possible explanation for the positive impact on school-level enrollment is that providers are not allowed to charge fees and program schools should be free at all levels, including early-childhood education (ECE). In contrast, control schools are officialy permitted to charge fees to ECE students and charge informal fees to primary students. Indeed, the likelihood that principals report charging fees in primary decreases in program schools by 21 percentage points (p-value .0025) after three years (see Table A.12,

[9] Romero et al. (in press) had shown that providers do not seem to engage in cream-skimming. This is still the case. There is no evidence that any group of students is systematically excluded from treatment schools after three years (Table A.4 provides more details).

Panel A) . The effect on ECE fees is similar. However, the reduction in the average yearly fee paid by parents is close to zero.

Turning to student-level enrollment, however, students enrolled in partnership schools in 2015/2016 are less likely to be enrolled in any school after three years (see Panel B, Table 2). This is not driven by students who were removed from their schools in the first year due to large class sizes. Instead, the effect seems to be driven by older students and by girls, who are more likely to drop out in treatment schools (see Table A.5).

We classify the reasons why students are no longer attending school into four broad categories: 1) Left school to work, 2) pregnancy, 3) could not afford school fees, and 4) others. Pregnancy seems to be driving this reduction in enrollment. Students originally enrolled in partnership schools are 2.32 percentage points (p-value < 0.001) more likely to drop out of schools because of pregnancy (from a base of 3.1%), consistent with the effect being driven by girls — Table A.9 provides more details. In addition, the effect is driven by students who were originally enrolled in Grade 4 and 5 in 2015/2016 (see Table A.8) and thus students are less likely to be enrolled in some form of secondary school (see Table A.11).

There are at least two explanations for the increase in dropout which we cannot rule out. The first, is that three treatment schools — all assigned to the same provider, Bridge International Academies, at the company's request — had a secondary school on the same premises. After Bridge took control of these schools it reassigned the classrooms and teachers assigned to secondary grades to primary, effectively shutting down the secondary school. One possibility is that students are less likely to progress to secondary school simply because the nearest secondary school was shut down. This is consistent with the overall negative effect on school attainment being driven by Bridge (see Section 4). However, due to small sample sizes any estimate of the impact of shutting down these grades is very noisy. The second possibility is that providers strengthened enforcement of the national policy requiring pregnant girls to drop out of school until after childbirth (Martinez & Odhiambo, 2018). Thus, an increase in the number of students dropping out due to pregnancy may reflect a stricter policy enforcement and not more pregnancies in treatment schools.

Table 2: ITT treatment effects on enrollment, attendance, and selection

	Year 1		Year 3	
	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)
Panel A: School level data				
Enrollment change	-6.06	24.60*	-63.13	35.93*
	(82.25)	(14.35)	(96.41)	(18.21)
Attendance % (spot check)	32.83	15.57***	41.99	10.71**
	(26.55)	(3.13)	(31.66)	(4.39)
% of students with disabilities	0.39	0.21	0.61	0.19
	(0.67)	(0.15)	(1.07)	(0.31)
Observations	87	175	90	181
Panel B: Student level data				
% enrolled in the same school	83.16	0.71	41.04	2.10
	(37.43)	(2.06)	(49.21)	(1.81)
% enrolled in school	93.99	1.23	84.50	-3.34***
	(23.77)	(0.87)	(36.20)	(1.15)
Days missed, previous week	0.85	-0.06	0.64	0.06
	(1.40)	(0.07)	(1.21)	(0.05)
Observations	1,786	3,639	1,780	3,624

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1 in Year 1 and Column 3 in Year 3), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2 in Year 1 and Column 4 in Year 3. Panel A presents school level data including enrollment (taken from enrollment logs) and student attendance measure by our enumerators during a 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. 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). Panel B presents student-level data including whether the student is still enrolled in the same schools, whether he is enrolled in school at all, and whether it missed school in the previous week (conditional on being enrolled in school). Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels is indicated by ***, **, and *.

Providers were authorized to cap class sizes in the first year, which could lead to students being excluded from their previous school (and either transferred to another school or to no school at all). Indeed, both after one year and after three years, enrollment decreased in classes originally above the cap, while enrollment increased in classes below the cap (see Column 1 in Table 3). However, this does not explain why students who were enrolled in partnership schools in 2015/2016 are less likely to be enrolled in any school after three years: The likelihood that students are enrolled in any school (Column 3) is the same for classes originally below and above the cap.

An alternative explanation for why the program caused students to drop out of school is lower motivation, but we find little corroboration for this hypothesis (see Table A.12, Panel B). The program had a

positive and significant treatment effect on student satisfaction after one year, although this is no longer the case after three years: students originally enrolled in partnership schools are 1.4 percentage points (p-value .56) more likely to think going to school is fun after three years. Likewise, after one year the program had a positive impact on students' likelihood to consider that what they learned in school was useful, but no impact after three years.

Table 3: ITT treatment effects by whether class size caps are binding

	Δ enrollment	% same school	% in school	Test scores
	(1)	(2)	(3)	(4)
Panel A: Year 1				
Constrained=0 × Treatment	5.30***	3.90***	1.65**	0.15***
	(1.11)	(1.40)	(0.73)	(0.034)
Constrained=1 × Treatment	-11.7*	-12.5	0.085	0.35***
	(6.47)	(7.72)	(4.12)	(0.11)
Observations	1,635	3,637	3,485	3,490
Mean control (Unconstrained)	-0.75	81.89	93.38	0.13
Mean control (Constrained)	-7.73	83.85	94.81	-0.08
α_0 = Constrained - Unconstrained	-17.05	-16.36	-1.56	0.20
p-value ($H_0 : \alpha_0 = 0$)	0.01	0.04	0.71	0.07
Panel B: Year 3				
Constrained=0 × Treatment	5.93***	3.27*	-3.33***	0.17***
	(1.22)	(1.75)	(1.27)	(0.035)
Constrained=1 × Treatment	-8.44	-2.03	-3.14	0.29**
	(7.07)	(5.92)	(4.93)	(0.13)
Observations	1,596	3,622	3,508	3,508
Mean control (Unconstrained)	-5.14	36.39	82.46	0.10
Mean control (Constrained)	-21.97	42.86	86.58	-0.07
α_0 = Constrained - Unconstrained	-14.37	-5.29	0.19	0.12
p-value ($H_0 : \alpha_0 = 0$)	0.05	0.40	0.97	0.38

Notes: Column 1 uses school-grade level data and the outcome is the change in enrollment at the grade level. Columns 2-4 use student-level data. The outcomes are whether the student is in the same school (Column 2), whether the student is still enrolled in any school (Column 3), and the composite test score (Column 4). Panel A presents the estimations for year 1 and Panel B for year 3. Standard errors are clustered at the school level. There were 194 constrained classes before treatment (holding 30% of students), and 1,468 unconstrained classes before treatment (holding 70% of students). Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *.

To reiterate, the program had a positive treatment effect ontotal enrollment — by slowing the rate of enrollment decline relative to control schools — but reduced enrollment and increased dropout for students originally enrolled in those schools. While the effect on access to education is clearly negative for the children in our sample, the net effect on access to education overall depends on an unknown parameter: the share of new students in partnerships schools who were previously unenrolled versus enrolled in other schools. It is unlikely that these students were previously unenrolled given the secular

decline in enrollment observed across all schools, suggesting a negative impact of the program on access to education in Liberia.

3.2 Learning

Following our pre-analysis plan, we report intention-to-treat (ITT) estimates from two specifications. The first specification amounts to a simple comparison of post-treatment outcomes for treatment and control individuals in which we only control for matched-pair fixed effect (i.e., stratification-level dummies). The second specification controls for time-invariant characteristics measured at the individual level and school level (Table A.2 provides a list of controls). We estimate both specifications via ordinary least squares, clustering the standard errors at the school level.

The intention-to-treat treatment effect of the program after three academic years is $.16\sigma$ for English (p-value < 0.001) and $.21\sigma$ for math (p-value < 0.001), as shown in Column 8 of Table 4. Treatment effects plateau after one year (when the treatment effects on English and math were $.18\sigma$ and $.18\sigma$, respectively, as shown in Column 5). Inclusion of student- and school-level controls has little effect on these results, as can be seen by comparing columns 4 and 5 as well as 7 and 8.

While we focus on the ITT effect, we also report treatment-on-the-treated (ToT) estimates (i.e., the treatment effect for students that actually attended a treated school). We focus on the ITT for several reasons. First, it requires fewer assumptions and represents the effect of offering the program, since it would be unethical to force students and parents to enroll in partnership schools. Second, the compliers (over which the ToT is estimated) may experience larger benefits from the program, and not represent the average student in the country. Third, the ITT incorporates any long-term gains from students who have already graduated from primary school. If the treatment effect disappears after students leave treatment schools, the ToT would not reflect this.

To estimate the ToT we use the randomized treatment assignment as an instrument for whether the student is in fact enrolled in a treated school. For the first two waves of data collection (1-2 months and 9-10 months) we instrument actual enrollment in a treated school with the treatment assignment. For the third wave of data collection, since students could have been exposed anywhere from zero to three years to the treatment, we calculate how many years students were enrolled in a treated school, and instrument this with the treatment assignment. To make the ITT and ToT effects comparable, we present the treatment-on-the-treated effect of being enrolled in a treated school for three years.[10]

[10] After one year, the percentage of students originally assigned to treatment schools who are actually in treatment schools at the

The ToT effect is 0.26σ for English (p-value < 0.001) and 0.35σ for math (p-value < 0.001), as shown in Column 9 of Table 4. Our results are robust to different measures of student ability (see Table A.3 for details).

Given the low baseline levels and variance in test scores, an important concern when interpreting these results expressed in standard deviations is how much learning they represent. We use correct words per minute and correct numbers per minute as a benchmark. Students enrolled in Grade 1 in 2015/2016 in control schools are able to read a little over 11 words per minute on average in 2019. Their counterparts in treatment schools can read about 15 words per minute. For students enrolled in Grade 5 in 2015/2016, the difference between treatment (27 words per minute) and control (25 words per minute) in 2019 is less than 2 words per minute. As a benchmark, to understand a simple passage students should read 45-60 words per minute. To meet this standard, students should read around 30 words correct per minute by the end of grade 1 (Abadzi, 2011). The gap between Liberian children (even in treated schools) and children in more developed countries is massive. For example, in Minnesota, third and five grade students can read 114 and 143 words per minute respectively (Silberglitt, Burns, Madyun, & Lail, 2006). See Figures A.4 and A.5 for more details.

end of the 2016/2017 school year is 81%. The percentage of students assigned to control schools who are in treatment schools at the end of the 2016/2017 school year is 0%. After two years, 56% of students originally assigned to treatment schools are actually in treatment schools. After three years, 33% of students originally assigned to treatment schools are actually in treatment schools.

Table 4: ITT treatment effects on learning

	First wave (1-2 months after treatment)			Second wave (9-10 months after treatment)			Third wave (33-34 months after treatment)		
	ITT		ToT	ITT		ToT	ITT		ToT
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
English	0.09*	0.07**	0.08**	0.17***	0.18***	0.21***	0.15***	0.16***	0.26***
	(0.05)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)
Math	0.07*	0.05*	0.06*	0.19***	0.18***	0.22***	0.20***	0.21***	0.35***
	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.06)
Abstract	0.05	0.03	0.04	0.05	0.05	0.06	0.02	0.03	0.05
	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.03)	(0.03)	(0.06)
Composite	0.08*	0.06*	0.07*	0.19***	0.18***	0.22***	0.19***	0.20***	0.33***
	(0.05)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.06)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Observations	3,508	3,508	3,508	3,492	3,492	3,492	3,510	3,510	3,510

Notes: Columns 1-3 are based on the first wave of data and show the difference between treatment and control schools taking into account the randomization design — i.e., including "pair" fixed effects — (Column 1), the difference taking into account other student and school controls (Column 2), and the treatment-on-the-treated (ToT) estimates (Column 3). Columns 4-6 are based on the second wave of data and show the difference between treatment and control taking into account the randomization design — i.e., including "pair" fixed effects — (Column 4), the difference taking into account other student and school controls (Column 5), and the treatment-on-the-treated (ToT) estimates (Column 6). Columns 7-9 are based on the third wave of data and show the difference between treatment and control taking into account the randomization design — i.e., including "pair" fixed effects — (Column 7), the difference taking into account other student and school controls (Column 8), and the treatment-on-the-treated (ToT) estimates (Column 9). The treatment-on-the-treated effects are estimated using the assigned treatment as an instrument for whether the student is in fact enrolled in a partnership school at the time of data collection. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *.

3.3 Sustainability

The outsourcing program changed the management of treated schools, while also increasing the total resources available to them (this was true even before it surpassed the original budget targets). The sustainability of the program depends in part on the relative importance of these two channels. Furthermore, some of these changes may have imposed negative externalities on the broader school system, by shifting students (see Section 3.1) and under-performing teachers to non-program schools. While we do not attempt any formal mediation analysis to quantify the role of competing mechanisms, this section estimates the effect of the program on school resources and management, and explores some of the potential negative externalities.

3.3.1 Inputs and resources

First, we focus on a key input in the education production function: teachers. In the first year, the Ministry of Education agreed to release some underperforming teachers from program schools, replace those teachers, and provide additional new teachers (Romero et al., in press). The net result was that

program schools had 2.6 more teachers on average (p-value < 0.001) after one year. After three years, operators still have 2.2 more teachers on average (p-value < 0.001). As expected, since the gap between treatment and control schools has remained the same across time, there is no effect on additional hiring or release of teachers in treatment schools after the first year. However, since treatment schools also have more students, the impact on the pupil-teacher ratio is small and statistically insignificant after three years.

After three years the composition of teachers is different in program schools compared to control schools. This is a consequence of the re-shuffling of teachers in the first year, which has long-lasting effects since the replacement and extra teachers in the first year were recent graduates from Rural Teacher Training Institutes (see King, Korda, Nordstrum, and Edwards (2015) for details on this program). See Panel B in Table 5 for more details. In short, the average teacher in a program school is younger, less-experienced, more likely to have worked in private schools in the past, and has higher test scores (we conducted simple memory, math, word association, and abstract thinking tests).

Teachers in program schools report higher wages. Yet, this was not mandated by the program rules. In fact, the program's contracts made no provisions to pay teachers differently in treatment and control schools. A potential explanation is that providers asked the government to put teachers who were previously paid by the community (known as 'volunteer' teachers) onto the government payroll, which would likely have resulted in higher wages for these teachers.

Finally, when comparing the materials available to students during classroom observations (see Panels C - Table 5), the positive treatment effect of the program on textbooks and chalk on the first year dissipate after three years. While the increase in the likelihood that schools in the control group have chalk (from 79% to 94%) could explain the latter, it cannot explain the former where the mean in the control group decreased (from 18% to 9.5%). Although the number of seats in each classroom increases, there is no treatment effect on the likelihood that students are sitting on the floor.[11]

[11] Since we could not conduct classroom observations in schools that were out of session during our visit, Table A.6 provides Lee (2009) bounds on these treatment effects (control schools are more likely to be out of session). However, we cannot rule out that there is no overall effect as zero is between the Lee (2009) bounds.

Table 5: ITT treatment effects on inputs and resources

	Year 1		Year 3	
	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)
Panel A: School-level outcomes				
Number of teachers	7.02	2.61***	6.95	2.22***
	(3.12)	(0.37)	(3.08)	(0.38)
Pupil-teacher ratio (PTR)	39.95	-7.82***	30.31	1.16
	(18.27)	(2.12)	(14.11)	(2.14)
New teachers	1.77	3.01***	2.42	0.20
	(2.03)	(0.35)	(1.72)	(0.26)
Teachers dismissed	2.12	1.13**	1.60	-0.23
	(2.62)	(0.47)	(1.46)	(0.22)
Observations	92	185	92	185
Panel B: Teacher-level outcomes				
Age in years	46.37	-7.10***	44.38	-5.79***
	(11.67)	(0.68)	(12.17)	(0.64)
Experience in years	15.79	-5.26***	13.90	-4.38***
	(10.77)	(0.51)	(10.90)	(0.51)
% has worked at a private school	37.50	10.20***	36.60	11.34***
	(48.46)	(2.42)	(48.22)	(2.26)
Test score in standard deviations	-0.01	0.14**	-0.03	0.21***
	(0.99)	(0.06)	(0.99)	(0.06)
% certified (or tertiary education)	58.05	4.20	53.05	17.46***
	(49.39)	(2.99)	(49.95)	(3.08)
Salary (USD/month)–Conditional on salary> 0	104.54	13.90***	107.48	24.09***
	(60.15)	(4.53)	(76.86)	(5.72)
Observations	489	1,167	478	1,142
Panel C: Classroom observation				
Number of seats	20.58	0.58	18.62	4.07*
	(13.57)	(1.90)	(11.74)	(2.18)
% with students sitting on the floor	4.23	-1.51	3.03	0.00
	(20.26)	(2.61)	(17.27)	(2.59)
% with chalk	78.87	16.58***	93.94	3.64
	(41.11)	(5.50)	(24.04)	(3.64)
% of students with textbooks	17.60	22.60***	9.55	8.29
	(35.25)	(6.32)	(21.59)	(5.43)
% of students with pens/pencils	79.67	8.16*	86.92	4.19
	(30.13)	(4.10)	(24.47)	(4.40)
Observations	71	143	66	114

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1 in Year 1 and Column 3 in Year 3), as well as the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2 in Year 1 and Column 4 in Year 3. Panel A has school level outcomes. Panel B presents teacher-level outcomes including their score in tests conducted by our survey teams. Panel C presents data on inputs measured during classroom observations. Since we could not conduct classroom observations in schools that were out of session during our visit, Table A.6 provides Lee (2009) bounds on these treatment effects (control schools are more likely to be out of session). Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *.

3.3.2 School management

The treatment effects on school management are similar after one and three years of the program (see Table 6). Program schools are 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: 8.7 percentage points (p-value .058) after one year, and 12 percentage points (p-value .022) after three years. Schools days are also longer: 3.2 more hours per week of instructional time (p-value .0011) after one year and 3.6 more hours per week (p-value < 0.001) after three years. Principals report spending 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 program schools did not have to double as teachers.

Finally, management practices (as measured by a "good practices" Principal Component Analysis index normalized to a mean of zero and standard deviation of one in the control group) are $.4\sigma$ (p-value .0011) higher in program schools after one year and $.55\sigma$ (p-value < 0.001) higher after three years. Some of these good practices include maintaining an enrollment log (treatment schools are 17 percentage points more likely, p-value .0086, to maintain one, from a base of 68%), having an official schedule posted somewhere on the school premises (treatment schools are 13 percentage points more likely, p-value .0039, to have one, from a base of 84%) or the principal having the contact information for the PTA head (principals in treatment schools are 13 percentage points more likely, p-value .063, to have the contact information, from a base of 35%). In 2019 we also measure management practices using a a DWMS style survey (see Section 2.2.4 for more details). According to this index, management practices improve by $.68\sigma$ (p-value < 0.001) after three years.

Table 6: ITT treatment effects on school management

	Year 1		Year 3	
	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)
% school in session at spot check	83.70	8.66*	75.82	12.41**
	(37.14)	(4.52)	(43.05)	(5.33)
Instruction time (hrs/week)	14.69	3.17***	18.91	3.63***
	(4.04)	(0.65)	(5.74)	(0.86)
Principal's working time (hrs/week)	20.60	0.84	24.33	0.25
	(14.45)	(1.88)	(11.94)	(1.75)
% of principle's time spent on management	53.64	20.09***	44.53	24.94***
	(27.74)	(3.75)	(23.25)	(3.64)
Index of good practices (PCA)	-0.00	0.40***	-0.00	0.55***
	(1.00)	(0.12)	(1.00)	(0.14)
Management index (DWMS-style)			-0.00	0.68***
			(1.00)	(0.14)
Observations	92	185	91	183

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1 in Year 1 and Column 3 in Year 3), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2 in Year 1 and Column 4 in Year 3. The index of good practices is the first component of a principal component analysis of the variables in Table A.7. The index is normalized to have mean zero and standard deviation of one in the control group. The management index is based on Development World Management Survey (DWMS) style-questions. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *.

3.3.3 Teacher behavior

A possible explanation for why treatment effects on test scores plateau after the first year, is that the treatment effects on teacher behavior dissipate after the first year. 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 7). Also, during these spot checks we used the Stallings classroom observation instrument to study teacher time use and classroom management (see Panel C in Table 7).

While teachers were more likely to be in schools and more likely to be in a classroom during a spot check after the first year — 20 percentage points (p-value < 0.001) more likely and 15 percentage points (p-value < 0.001) more likely, respectively — this is no longer true after three years when the treatment effects are smaller and insignificant — 6.8 percentage points (p-value .12) more likely and 7.3 percentage points (p-value .1) more likely, respectively. This does not seem to be driven by control schools improving teacher attendance since this remains relatively constant across time (goes from 40% in 2016 to 46% in 2019) and the average attendance in treatment schools goes down across time (from 60% in 2016 to 54% in 2019).

Classroom observations also show a reduction in the treatment effects on teacher behavior and pedagogical practices after three years. After one year teachers were 25 percentage points (p-value < 0.001) less likely to be off-task during class time. After three years teachers were 14 percentage points (p-value .013) less likely to be off-task. While the treatment effect is still significant, it is over 10 percentage points smaller. Likewise, the treatment effect on instruction time and classroom management is still significant after three years, but smaller. The reduction in treatment effects could be explained by control schools improving over time.[12] In control schools the time off-task goes from 56% in 2016 to 40% in 2019.

To put these numbers in perspective, the Stallings, Knight, and Markham (2014) good practice benchmark is that 85% of total class time should be used for instruction. After three years control schools are using less than 50% of time for instruction, while program schools are using close to 60%.

[12]We cannot rule out that control schools improve due to spillovers as the experiment was not designed to estimate any spillover effects. However, given the decline in teacher attendance in treatment schools noted above, this is unlikely.

Table 7: ITT treatment effects on teacher behavior

	Year 1		Year 3	
	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)
Panel A: Spot checks				
% on schools campus	40.38	19.79***	46.42	6.75
	(25.20)	(3.48)	(28.64)	(4.29)
% in classroom	31.42	15.37***	36.94	7.32
	(25.04)	(3.62)	(30.93)	(4.46)
Observations	92	185	92	185
Panel B: Student reports				
Teacher missed school previous week (%)	25.12	-7.53***	23.65	-2.68*
	(14.93)	(1.95)	(11.91)	(1.59)
Teacher never hits students (%)	48.20	6.59**	48.69	4.59**
	(17.07)	(2.53)	(16.88)	(2.23)
Teacher helps outside the classroom (%)	46.59	3.56	37.19	1.91
	(18.01)	(2.28)	(14.82)	(2.13)
Observations	92	185	92	185
Panel C: Classroom observations				
Instruction (active + passive) (% of class time)	35.00	14.51***	49.57	9.28*
	(37.08)	(4.70)	(39.25)	(5.08)
Classroom management (% class time)	8.70	10.25***	10.00	5.09**
	(14.00)	(2.73)	(14.75)	(2.38)
Teacher off-task (% class time)	56.30	-24.77***	40.43	-14.37**
	(42.55)	(5.48)	(44.15)	(5.70)
Student off-task (% class time)	47.14	2.94	28.00	1.09
	(38.43)	(4.59)	(31.97)	(4.15)
Observations	92	185	92	185

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1 in Year 1 and Column 3 in Year 3), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2 in Year 1 and Column 4 in Year 3. Panel A presents data from spot checks conducted by our survey teams in the middle of a school day. Panel B presents data from our panel of students where we asked them about their teachers' behavior. Panel C presents data from classroom observations. If the school was not in session during a regular school day we mark all teachers not on campus as absent and teachers and students as off-task in the classroom observation. Table A.6 has the results without imputing values for schools not in session. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels is indicated by ***, **, and *, respectively.

Since these estimates combine the effects on individual teacher behavior with changes to teacher composition, we perform additional analyses in Appendix A using administrative data (EMIS) to restrict our sample to teachers who worked at the school the year before the intervention began (2015/2016). In this analysis teachers who no longer worked at the school in the 2018/2019 school year are treated as (non-random) attriters. We then estimate Lee (2009) bounds on the treatment effect for this pool of teachers. However, the number of attriters is large after three years making these bounds uninformative.

3.3.4 Discussion

Three features from the analysis above stand out regarding sustainability. First, some of the advantages given to program schools in terms of staffing in the first year have had long-lasting effects. Program schools have more and better-trained teachers after three years as a consequence (although the pupil-teacher ratio in treatment schools has converged back to the level in control schools). Because the program did not increase the total supply of trained teachers, this raises doubts about the scalability of the program's impacts. Second, observable management practices are better in program schools, and the durability of this result is a positive signal of sustainability. Finally, while teacher behavior and pedagogy are still better in treatment schools, the effect attenuated after the first year. We can only speculate about the reasons for this last result, including the possibility that: a) teachers' enthusiasm for the program waned over time; b) teachers realized that providers lacked "sticks" to use against them since they remained unionized and on the government payroll[13]; c) providers failed to provide credible "carrots" (e.g., promises to get volunteer teachers on the government payroll and guarantee regular payments that were often not fulfilled)[14]; and d) providers exerting less effort after the first year media and international donor community scrutiny dissipated.

3.4 Child safety

We use two measures of child safety within the school: corporal punishment and sexual abuse. Corporal punishment is higher in Sub-Saharan Africa than in other regions of the world (Gershoff, 2017) and prior studies have found that sexual violence is widespread (e.g., rates of sexual coercion of 30% for girls and 22% for boys) in Liberian schools (Postmus et al., 2015; Steiner et al., 2018). In addition, two providers have been involved in sexual abuse scandals, as reported in the media. The most prominent scandal was revealed by a ProPublica investigation that found that thirty or more girls were raped by the co-founder of More than Me prior to the launch of the partnership program (F. Young, 2018). The chair of the board of the Liberian Youth Network (the previous name for the Youth Movement for Collective Action) was also found guilty of raping a teenage boy (Baysah, 2016). Both of these incidents occurred prior to the launch of the program, but were revealed in full only after the program launched.

[13]There are good theoretical reasons to avoid giving providers full control of staffing decisions. Giving providers full staffing control could induce sorting of teachers across schools, increasing inequality in the system without raising the quality or the productivity of the average teacher. In addition, providers might have monopsonistic power in local labor markets in rural areas with few schools.

[14]This failure may be due, inpart, to the withdrawal during years 2 and 3 of some of the special treatment given to providers in the very early stages of the programme, when the teachers in their schools were given priority to be added to the payroll

While sexual abuse was not highlighted in the design of the outsourcing program, a survey module on sexual abuse was added in 2019 in response to these concerns. The survey was administered to students twelve years old and above after the learning assessment, and asked about sexual abuse at school (by teachers and peers) and at home. Students could opt out — the response rate is close to 90% and balanced between treatment and control schools, see Table A.13 — and their responses were anonymous and not linked to any student identifier.

Corporal punishment is widespread across schools: 51% of students in control schools report being hit by their teachers at least occasionally. The program reduced this rate by 4.6 percentage points (p-value .043).

Sexual abuse rates in our data are lower than those reported in previous studies (Postmus et al., 2015; Steiner et al., 2018): 3.6% of students in control schools report having sex with a teacher (statutory rape).[15] The program had a small (-.41 percentage points) and statistically insignificant (p-value .46) impact on this measure. However, the program increased reported forced sexual intercourse by peers by 1.7 percentage points (p-value .0042) from a base of 3.6%. A possible explanation is that the likelihood of *reporting* an incident may have increased in program schools. Indeed, reported cases of forced sexual intercourse at home — where the true rate is unlikely to be affected by the program — increased by 1 percentage points (p-value .071) from a base of 2.9%.

In summary, the program reduced, but did far from eradicated the use of corporal punishment in schools. Despite many credible reports of sexual abuse in schools run by private providers involved in this program, we fail to find any significant change in self-reported sexual abuse as a result of the program. Nevertheless, sexual abuse remains far too widespread, and private providers failed to use an influx of new resources and external oversight to reduce its incidence.

[15] In a companion paper, Johnson et al. (2019) compare the survey protocol used in this paper with a protocol identical to the one used by Postmus et al. (2015) and Steiner et al. (2018) and find similar rates of sexual abuse across protocols.

Table 8: Gender based violence

	All		Girls		Boys	
	Control	Treatment Effect	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher: Sex	3.55	-0.41	3.99	-1.55*	3.24	0.30
	(18.52)	(0.54)	(19.60)	(0.85)	(17.71)	(0.72)
Teacher: Touched	7.47	-0.01	6.67	-0.78	8.05	0.84
	(26.30)	(0.75)	(24.97)	(1.03)	(27.23)	(1.14)
Teacher: Forced sex	2.37	0.06	2.83	0.39	2.04	-0.29
	(15.23)	(0.42)	(16.61)	(0.74)	(14.16)	(0.50)
Student: Touched	16.36	0.08	20.03	-4.50***	13.72	2.23
	(37.01)	(1.09)	(40.06)	(1.70)	(34.42)	(1.47)
Student: Forced sex	3.56	1.71***	3.18	0.95	3.84	1.76*
	(18.54)	(0.59)	(17.55)	(0.88)	(19.22)	(0.90)
Family: Touched	9.49	-0.64	10.15	-1.91	9.01	0.29
	(29.32)	(0.81)	(30.22)	(1.42)	(28.66)	(1.08)
Family: Forced sex	2.93	1.01*	3.84	-0.16	2.28	1.64**
	(16.87)	(0.56)	(19.23)	(1.00)	(14.93)	(0.70)
Observations	1,435	2,869	601	1,239	834	1,630

Notes: This table presents the mean and standard deviation (in parentheses) for the control as well as the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) for all students (Column 1-2), only girls (Column 3-4), and only boys (Columns 5-6). Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels is indicated by ***, **, and *, respectively.

4 Provider level heterogeneity

We now turn to provider-by-provider outcomes. We focus on the same four margins as the policy impact: access, learning, sustainability, and child safety. We document two facts. First, there is heterogeneity in treatment effects across providers in all these dimensions. Second, the group of providers that performs well in various dimensions is different, posing trade-offs for policymakers who must decide on the weights to attach to each outcome.

The experimental design amounts to having eight experiments, one for each provider. However, there is heterogeneity in school characteristics across the experiments for each provider. Thus, while the raw treatment effects for each individual provider are internally valid, they are not strictly comparable with each other without further assumptions. Romero et al. (in press) estimates provider-specific results with and without controls for baseline school characteristics, and finds baseline heterogeneity does little to explain heterogeneity in treatment effects.

In addition, we cannot cluster at the school level since the number of schools per operator is not enough

to guarantee convergence. Instead, we use randomization inference which provides exact tests of sharp hypotheses no matter the sample size.[16]

4.1 Heterogeneity across providers

4.1.1 Access

On access, the overall negative treatment effect on the baseline sample of students still attending any school is driven by Bridge International Academies (p-value 0.088), Omega (p-value 0.163) and More Than Me (p-value 0.342). As expected given the overall results, the effect is not driven by students who were originally enrolled in constrained classes, where students were dismissed en-mass during the first year. Instead, the effect is driven by an increase in the likelihood that students dropout of school due to pregnancy in Bridge International Academies (see Table A.10). As mentioned in Section 3.1, another explanation may be that Bridge effectively shut down the nearest secondary school for some students. To recap, some of the schools in the experiment had a secondary school on the same campus. The Ministry attempted to exclude schools such schools from the outsourcing program, but Bridge successfully appealed for the inclusion of a handful of them in the eligible list. All these schools were assigned to Bridge International Academies at their request, prior to randomization. After Bridge took control of the subset selected for treatment it reassigned the classrooms and teachers assigned to secondary grades to primary, effectively shutting down the secondary school.

The treatment effect on school size — i.e. attracting new students not in our baseline sample – is positive across all providers, except Stella Maris and Street Child (Table 9 - Panel D). However, only BRAC, More Than Me, and the Youth Movement for Collective Action have a statistically significant treatment effect. BRAC, More Than Me, Rising, Street Child, and the Youth Movement for Collective Action have a positive and statistically significant treatment effect on student attendance.

4.1.2 Learning

While the program as a whole raised composite learning outcomes by $.2\sigma$, three of the eight operators produced negligible and statistically insignificant learning gains, while the other five generated similar ITT effects of roughly 0.4σ (Table 9 - Panel A and Figure A.3a). Of this latter group, four of five (Rising Academies, More than Me, Bridge International Academies, and Street Child) have learning effects that

[16] For a detailed discussion of the (theoretical and practical) effects of using conventional econometric tests that rely on asymptotic results see A. Young (2018).

are statistically significant using randomization inference. The Youth Movement for Collective Action has an ITT effect of a similar magnitude but it is statistically insignificant, perhaps due in part to a smaller sample size. The treatment-on-the-treated effects (Table 9 - Panel B and Figure A.3b) shows a higher variance across operators, with effects as high as 0.9σ for More Than Me, and as low as -0.12σ for Omega Academies.

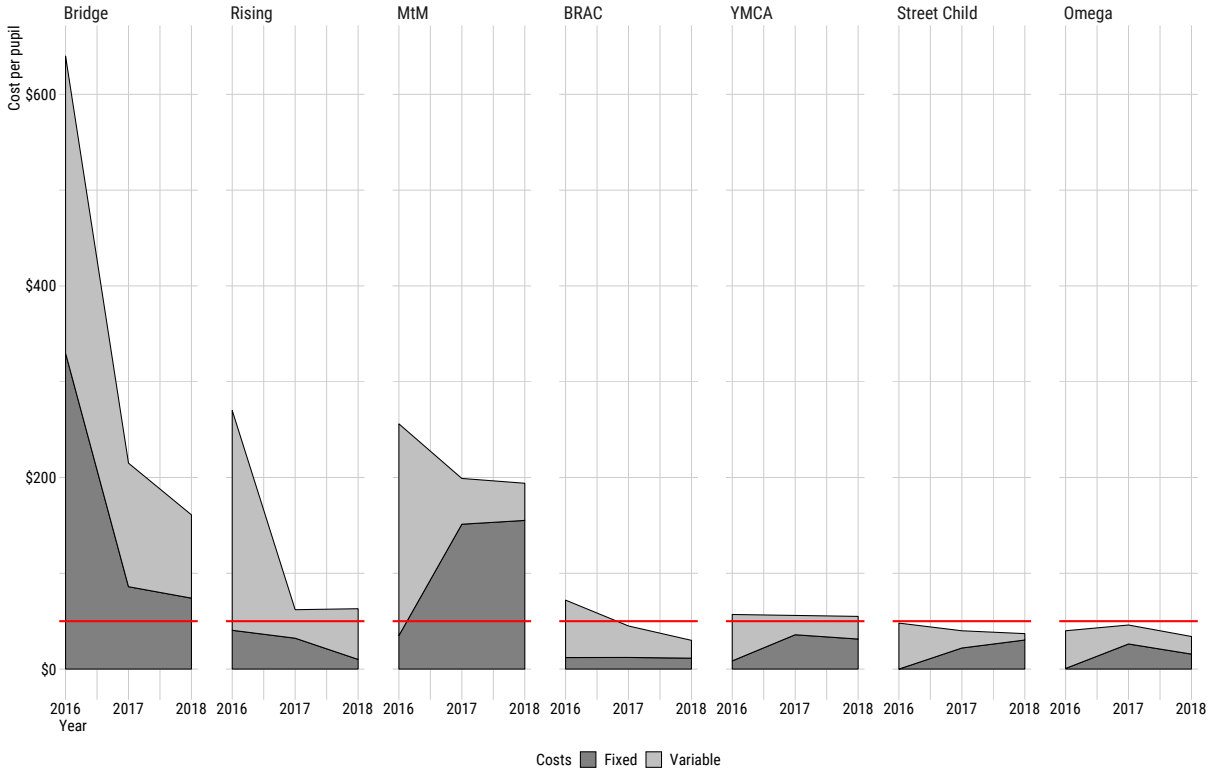
4.1.3 Sustainability

Here we focus on the main cost of the program paid by the government (teachers) and on the extra cost incurred by providers beyond the government transfer per pupil.

On the former point, we study the number of teachers dismissed and the number of new teachers recruited (Table 9 - Panel C). As noted in Section 3.3.1, the program led to the assignment of 2.2 extra teachers per school, even after taking into account that the program led to .26 additional teachers exiting per school. Large-scale dismissal of teachers is driven by Bridge International Academies (p-value 0.000) and More Than Me (p-value 0.001). However, most providers were able to get several new teachers, except for Omega Academies, Stella Maris, and the Youth Movement for Collective Action. Notably, the treatment effect on the percentage of new teachers is 90 percentage points for More Than Me. While weeding out bad teachers is important, a reshuffling of teachers is unlikely to raise average performance in the system as a whole. We are unable to verify whether the teachers dismissed from program schools were reassigned to other public schools.

In terms of additional direct costs of the program, the ministry expects providers to operate for USD 50 or less in the long term. Self-reported unit costs have fallen for most providers over time. Yet, for Bridge International Academies and More Than Me, total unit costs remain at least three times as much as the government target. While we report fixed and variable costs in Figure 2, it appears providers fail to distinguish between these categories. For instance, Street Child and Youth Movement for Collective action claim nearly zero variable cost, implying the potential for costless expansion of the program, which is unrealistic.

Figure 2: Per pupil cost



Note: Numbers are based on budgets submitted to Social Finance, who managed the pool of funds that paid providers the per pupil subsidy. Stella Maris did not provide budget data. Numbers do not include the cost of teaching staff borne by the Ministry of Education. The red line represents USD 50 per pupil.

4.1.4 Child Safety

First, we focus on corporal punishment. The treatment effect on the likelihood that students report never being hit by teachers is positive and statistically significant only for Street Child (p-value 0.031). The treatment effect is positive and insignificant for four other operators, and negative and insignificant for the remaining three.

In terms of students reporting sex with teachers, there is only a negative treatment effect (i.e., lower rates) for BRAC (p-value 0.053), with negative and insignificant point estimates for four other providers and positive and insignificant estimates for three others.

4.2 Tradeoffs between outcomes

We provide treatment effects across a list of providers, carefully vetted by the government. Despite facing similar contracts and settings, the identity of the provider matters (for comparison, see Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur (2018)). Thus, the success of a public-private does not depending only on the context and the details of the contract, but also on which partners are selected through the procurement process.

Some providers perform well in almost every dimension. Street Child produced positive and statistically significant learning gains without a negative effect on access. They were also able to reduce corporal punishment. While they received a considerable number of new teachers, their external cost is low. Rising Academies also produced a consistently positive pattern of results across learning, access, and safety dimensions, though these are not always statistically significant and Rising's costs were higher than the government target.

At the other extreme, some providers underperformed across the board. Omega Academies and Stella Maris both show no measurable, positive effect on any dimension, including point estimates on learning outcomes that are actually negative.

Each of the other providers presents policymakers with a tradeoff between outcome dimensions.[17] Bridge produced high learning gains at a high cost and reduced access to education for some children. Children in BRAC schools were less likely to report sexual abuse, but saw almost no learning gains. Yet, school-level enrollment increased in BRAC schools. More Than Me produced the highest learning gains (by ToT estimates) and increased school-level enrollment, but the gains come at a high cost to the government and external donors. Moreover, the organization has failed to protect children from sexual abuse. YMCA poses the same trade-off: its results are quite strong across the board, but must be considered in light of the sexual abuse case involving its board chair. While both cases of sexual abuses involving providers staff occurred before the start of the program, an important difference between the More Than Me and YMCA scandals is that the board of YMCA was fired as soon as the scandal came to light. F. Young (2018) investigation shows how the More than Me leadership failed to accept responsibility, and successfully conceal the case from public scrutiny for many years.

[17] Note that there is no clear pattern linking results to whether the provider is a for-profit business or non-profit charity.

Table 9: Treatment effects by provider (randomization inference)

	BRAC (1)	Bridge (2)	MtM (3)	Omega (4)	Rising (5)	St. Child (6)	Stella M (7)	YMCA (8)
Panel A: Student test scores (ITT)								
English	0.088	0.223	0.294	-0.057	0.317	0.245	-0.201	0.625
	[0.283]	[0.090]	[0.014]	[0.593]	[0.195]	[0.050]	[0.510]	[0.123]
Math	0.056	0.393	0.448	-0.099	0.433	0.286	0.076	0.285
	[0.519]	[0.001]	[0.006]	[0.311]	[0.064]	[0.116]	[0.786]	[0.127]
Composite	0.058	0.348	0.424	-0.083	0.417	0.285	-0.042	0.419
	[0.473]	[0.006]	[0.006]	[0.430]	[0.063]	[0.071]	[0.974]	[0.127]
Panel B: Student test scores (ToT)								
English	0.118	0.398	0.624	-0.080	0.514	0.463	0.000	0.806
	[0.321]	[0.011]	[0.076]	[0.573]	[0.059]	[0.002]	[.]	[0.001]
Math	0.074	0.696	1.008	-0.154	0.721	0.520	0.000	0.324
	[0.523]	[0.000]	[0.001]	[0.317]	[0.008]	[0.001]	[.]	[0.181]
Composite	0.076	0.617	0.941	-0.125	0.689	0.526	0.000	0.508
	[0.518]	[0.000]	[0.002]	[0.424]	[0.014]	[0.001]	[.]	[0.036]
Panel C: Changes to the pool of teachers								
% teachers dismissed	-16.327	43.931	26.191	-3.719	-0.056	-13.815	-4.180	-13.720
	[0.000]	[0.000]	[0.001]	[0.663]	[0.936]	[0.128]	[0.744]	[0.370]
% new teachers	58.271	61.647	90.441	17.034	52.500	40.439	-46.740	22.024
	[0.000]	[0.001]	[0.027]	[0.344]	[0.002]	[0.021]	[0.000]	[0.125]
Age in years (teachers)	-3.956	-11.474	-8.646	-5.677	-10.326	-1.997	-8.826	0.830
	[0.000]	[0.000]	[0.000]	[0.000]	[0.074]	[0.364]	[0.001]	[0.751]
Test score (teachers)	0.238	0.239	-0.049	0.268	0.293	0.262	-0.216	0.282
	[0.048]	[0.159]	[0.748]	[0.172]	[0.385]	[0.278]	[0.388]	[0.014]
Panel D: Enrollment and access								
Δ enrollment	57.275	25.045	69.900	26.789	75.000	-5.182	-1.625	92.625
	[0.082]	[0.683]	[0.026]	[0.384]	[0.215]	[0.899]	[0.917]	[0.051]
Δ enrollment constrained	0.000	-10.638	0.000	-5.182	0.000	-38.500	0.000	0.000
	[.]	[0.229]	[.]	[0.808]	[.]	[0.151]	[.]	[.]
Student attendance	20.320	4.921	44.265	18.158	29.513	20.853	6.259	13.573
	[0.001]	[0.421]	[0.000]	[0.229]	[0.000]	[0.030]	[0.490]	[0.000]
% still attending any school	-1.101	-6.470	-3.807	-4.844	0.935	-0.440	3.177	-0.077
	[0.647]	[0.088]	[0.342]	[0.163]	[0.890]	[0.933]	[0.663]	[0.917]
% still attending same school	3.495	-4.514	4.002	5.028	11.779	7.341	8.968	-0.392
	[0.481]	[0.356]	[0.741]	[0.471]	[0.517]	[0.218]	[0.378]	[0.877]
Panel E: Child Safety								
Teacher never hits students (%)	6.35	-2.01	19.39	-0.66	9.57	14.55	9.59	-5.23
	[0.157]	[0.692]	[0.142]	[0.876]	[0.304]	[0.031]	[0.619]	[0.381]
Teacher (unforced sex)	-3.68	0.52	-0.12	1.20	-1.51	-0.54	7.73	-2.72
	[0.053]	[0.727]	[0.777]	[0.492]	[0.871]	[0.589]	[0.470]	[0.317]
Number of schools	40	45	12	38	10	24	8	8

Notes: This table presents the raw treatment effect for each provider on different outcomes using randomization inference. The estimates for each provider are not comparable to each other without further assumptions, and thus we do not include a test of equality. The randomization inference uses 5,000 iterations to calculate p-values based upon the distribution of squared t-statistics following A. Young (2018). Panel A presents data on intention-to-treat estimates on students' test scores. Panel B presents data on treatment-on-the-treated estimates on students' test scores. Panel C presents data related to the pool of teachers in each school. Panel D presents data related to school enrollment. Panle E presents data related to child safety (corporal punishment and sexual abuse). p-values from randomization inference are presented in brackets. Number of schools refers to the number of schools in both treatment and control groups.

5 Conclusions

Romero et al. (in press) show that Liberia's initiative to outsource management of ninety-three randomly-selected government primary schools to eight private operators led to short-term learning gains, partially offset by high costs and early signs of negative side-effects on the broader school system. In this paper we summarize impacts (a) over a longer time horizon, (b) on a range of outcomes beyond test scores, and (c) distinguishing the average impact of the outsourcing policy, which was modest overall, from the larger effects of some specific private operators.

Beyond the first year, treatment effects on learning gains plateau. After one year the treatment effects on English and math were .18σ and .18σ, respectively. After three academic years the treatment effect is .16σ for English (p-value < 0.001) and .21σ for math (p-value < 0.001). This corresponds to an increase in reading fluency of 4 and 2 words per minute for students enrolled in first and fifth grade in 2015/2016, respectively. As a benchmark, to understand a simple passage students should read 45-60 words per minute (Abadzi, 2011) and children in more developed countries can read over 140 words per minute by fifth grade (Silberglitt et al., 2006).

Beyond learning gains, the program reduced enrollment and increased dropout for the sample of students originally enrolled in partnership schools. While the program reduced the use of corporal punishment in schools, abuse remains widespread. Despite an influx of new resources and external oversight, sexual abuse did not decline in partnership schools. In addition, some of the advantages given to program schools in terms of staffing in the first year have had long-lasting effects. Finally, Bridge International Academies and More Than Me still spend at least three times as much as the government target.

Beyond these average effects, we document substantial heterogeneity in impacts across the eight private providers which managed schools as part of the overall program. Heterogeneity exists not only in learning, where three of eight providers had fairly precise null effects while others produced significant gains, but also in impacts on access to education and child safety, as well as in the sustainability of providers' models. Complicating the policy analysis, impacts on these various dimensions are not perfectly correlated across providers. While some providers show almost uniformly positive or null (and even negative) effects, several providers present trade-offs for policymakers who must decide on the weights to attach to positive gains on one dimension and losses in another.

A few limitations of our experiment bear repeating. First, since the evaluation randomized treatment status among a very selected school sample, we are unable to assess the effect of the program on the

average school in Liberia. Out of 2,619 public primary schools in Liberia, only 299 schools satisfied all the criteria to participate in the program. Since schools that met the criteria to take part in the program had above-average resources and infrastructure, this also means the program concentrated an influx of new funding on a narrow set of schools that were already advantaged. A reasonable policymaker might consider not only whether a policy, such as outsourcing, improves efficiency, but also how it affects equity.

Second, while our experimental design, which tracks pupils originally enrolled in treatment and control schools over three years, allows us to estimate unbiased intention-to-treat effects, it also prevents us from estimating welfare impacts on a broader population of children — including possible negative externalities from the expulsion of children and release of under-performing teachers, and possible positive impacts due to attracting new children into treatment schools.

Third, the unique features of this program should shape any attempt to extrapolate our results to other public-private partnerships in education globally. Compared to other programs in this literature (Cremata et al., 2013; Woodworth et al., 2017; Barrera-Osorio et al., 2017; Aslam et al., 2017), Liberia's outsourcing initiative provides limited top-down control at two levels. At the school level, private providers have limited control over teachers in their schools — some of whom remain unionized civil servants, while others remain off-payroll and are rarely paid. At the program level, the Liberian government also has limited capacity to incentivize or sanction private providers, who are reimbursed directly by third-party philanthropies. The attenuation of treatment effects on learning gains after the first year could be explained by a lack of accountability at these two levels.

Sanctioning under-performing providers is an obvious policy recommendation of our results. The dynamic here is reminiscent of the commitment problems studied in the contract theory literature on aid (Svensson, 2000), but with a different source of time inconsistency. Some of the philanthropies operating in Liberia conditioned their aid on the participation of specific private providers, undermining the government's credibility when linking provider contracts to results. Even providers who did nothing in the first year, or who were implicated in serious sexual abuse cases, were rewarded with more schools when the program expanded (e.g., Stella Maris and More than Me, respectively). Similarly, Bridge International Academies, which expelled students en masse at the beginning of the program, was the provider rewarded with the most schools when the program expanded (Ministry of Education - Republic of Liberia, 2017b, 2017a).

Finally, our results suggest the identity of private contractors matters for the performance of public-private partnerships, even when holding constant the setting and contractual terms. Some operators

performed well even on outcomes that were not specified in the original contracts. When state capacity to monitor performance and enforce contracts is weak — as in the case of Liberia — selecting private providers who are aligned with the public interest, and disinclined to exploit contractual incompleteness, may be important for the success of outsourcing schemes (à la Besley and Ghatak (2005); Akerlof and Kranton (2005)).

References

Abadzi, H. (2011). *Reading fluency measurements in efa fti partner countries: Outcomes and improvement prospects.* World Bank.

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

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

Banerjee, A. V., & Duflo, E. (2000). Reputation effects and the limits of contracting: A study of the Indian software industry. *The Quarterly Journal of Economics*, *115*(3), 989–1017.

Barrera-Osorio, F. (2007). *The impact of private provision of public education: empirical evidence from Bogota's concession schools.* (Mimeo)

Barrera-Osorio, F., Blakeslee, D. S., Hoover, M., Linden, L., Raju, D., & Ryan, S. P. (2017, September). *Delivering education to the underserved through a public-private partnership program in Pakistan* (Working Paper No. 23870). National Bureau of Economic Research.

Baysah, A. M., Jr. (2016, Nov). *Liberia: Police charge youth activist for sodomy.* Retrieved from https://web.archive.org/web/20161103182507/https://allafrica.com/stories/201611020824.html

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

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.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics*, *168*, 1–20.

Bonilla, J. D. (2010). *Contracting out public schools for academic achievement: Evidence from Colombia.* (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.

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

Corts, K. S., & Singh, J. (2004). The effect of repeated interaction on contract choice: Evidence from offshore drilling. *Journal of Law, Economics, and Organization*, *20*(1), 230–260.

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.

Eyles, A., & Machin, S. (2019). The introduction of academy schools to england's education. *Journal of the European Economic Association*, *17*(4), 1107–1146.

Gershoff, E. T. (2017). School corporal punishment in global perspective: prevalence, outcomes, and efforts at intervention. *Psychology, health & medicine*, *22*(sup1), 224–239.

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.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7*, 24–52.

Johnson, L., Romero, M., Sandefur, J., & Sandholtz, W. (2019). *Comparing three measures of sexual violence in Liberian schools.* (Mimeo)

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.

Kristof, N. (2017). A solution when a nation's schools fail. *The New York Times*. Retrieved 15/07/2017, from https://www.nytimes.com/2017/07/15/opinion/sunday/bridge-schools-liberia.html

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.

Martinez, E., & Odhiambo, A. (2018). *Leave no girl behind in africa: Discrimination in education against pregnant girls and adolescent mothers.* (Tech. Rep.). Human Rights Watch. Retrieved from https://www.hrw.org/sites/default/files/report_pdf/au0618_insert_webspreads.pdf

Ministry of Education - Republic of Liberia. (2015-2016). *Education Management Information System (EMIS).* http://moe-liberia.org/emis-data/.

Ministry of Education - Republic of Liberia. (2017a). *Partnership schools for liberia school allocation 2017.*

Ministry of Education - Republic of Liberia. (2017b). *PSL school allocation: Decision points.* Retrieved 28/07/2017, from http://moe.gov.lr/wp-content/uploads/2017/06/Allocation-final.pdf

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

Pilling, D. (2017). Liberia is outsourcing education. Can it work? *The Financial Times*. Retrieved 13/09/2017, from https://www.ft.com/content/291b7fca-2487-11e7-a34a-538b4cb30025

Postmus, J. L., Hoge, G. L., Davis, R., Johnson, L., Koechlein, E., & Winter, S. (2015). Examining gender based violence and abuse among liberian school students in four counties: An exploratory study. *Child abuse & neglect*, *44*, 76–86.

Romero, M., Sandefur, J., & Sandholtz, W. A. (in press). Outsourcing education: Experimental evidence from liberia. *American Economic Review*.

Rosenberg, T. (2016). Liberia, desperate to educate, turns to charter schools. *The New York Times*. Retrieved 20/07/2016, from http://www.nytimes.com/2016/06/14/opinion/liberia-desperate-to-educate-turns-to-charter-schools.html

Sandholtz, W. A. (2019). *Do voters reward service delivery? Experimental evidence from Liberia.* (Mimeo)

Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools*, *43*(5), 527-535. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/pits.20175 doi: 10.1002/pits.20175

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.

Steiner, J. J., Johnson, L., Postmus, J. L., & Davis, R. (2018). Sexual violence of liberian school age students: An investigation of perpetration, gender, and forms of abuse. *Journal of child sexual abuse*, 1–20.

Svensson, J. (2000). When is foreign aid policy credible? aid dependence and conditionality. *Journal of development economics*, *61*(1), 61–84.

The Economist. (2017). *Liberia's bold experiment in school reform.* Retrieved 13/09/2017, from https://www.economist.com/news/middle-east-and-africa/21717379-war-scorched-state-where-almost-nothing-works-tries-charter-schools-liberias

Tyre, P. (2017). Can a tech start-up successfully educate children in the developing world? *The New York Times*. Retrieved 27/06/2017, from https://www.nytimes.com/2017/06/27/magazine/can-a-tech-start-up-successfully-educate-children-in-the-developing-world.html

UNICEF. (2013). *The state of the world's children: Children with disabilities* (Tech. Rep.). United Nations.

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

Young, A. (2018, 11). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *The Quarterly Journal of Economics*, *134*(2), 557-598. doi: 10.1093/qje/qjy029

Young, F. (2018, Oct). *Unprotected.* Retrieved from https://features.propublica.org/liberia/unprotected-more-than-me-katie-meyler-liberia-sexual-exploitation/

A Additional tables and figures

Table A.1: Number of schools by provider

	Randomly assigned	Noncompliant	Replacement	Outside sample	Managed (1)-(2)+(3)+(4)	% compliant in sample [(1)-(2)]/(1)
	(1)	(2)	(3)	(4)	(5)	(6)
BRAC	20	0	0	0	20	100%
Bridge	23	0	0	2	25	100%
YMCA	4	0	0	0	4	100%
MtM	6	2	2	0	6	67%
Omega	19	2	0	0	17	89%
Rising	5	1	0	1	5	80%
Stella	4	4	0	0	0	0%
St. Child	12	2	2	0	12	83%

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

Table A.2: Control variables

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

Notes: This table shows the control variables at the students and the school level.

Table A.3: Treatment effects across various measures of difference in student ability

	Year 1		Year 3	
	ITT	ToT	ITT	ToT
	(1)	(2)	(3)	(4)
Panel A: Base IRT model				
English	0.18***	0.21***	0.16***	0.26***
	(0.03)	(0.04)	(0.03)	(0.05)
Math	0.18***	0.22***	0.21***	0.36***
	(0.03)	(0.04)	(0.04)	(0.06)
Panel B: Base IRT model standardized by grade				
English	0.23***	0.28***	0.18***	0.31***
	(0.04)	(0.05)	(0.04)	(0.06)
Math	0.23***	0.27***	0.24***	0.39***
	(0.04)	(0.05)	(0.04)	(0.07)
Panel C: PCA				
English	0.16***	0.19***	0.13***	0.21***
	(0.03)	(0.04)	(0.03)	(0.05)
Math	0.24***	0.28***	0.23***	0.38***
	(0.04)	(0.05)	(0.04)	(0.06)
Panel D: PCA standardized by grade				
English	0.19***	0.23***	0.13***	0.21***
	(0.04)	(0.05)	(0.03)	(0.05)
Math	0.28***	0.33***	0.25***	0.41***
	(0.05)	(0.06)	(0.04)	(0.07)
Panel E: % correct answers				
English	2.96***	3.56***	2.37***	3.96***
	(0.55)	(0.66)	(0.50)	(0.83)
Math	4.24***	5.09***	4.28***	7.14***
	(0.71)	(0.84)	(0.70)	(1.17)
Observations	3,492	3,492	3,510	3,510

Notes: Column 1 and 3 show the intention-to-treat treatment effect estimated with a specification that takes into account the randomization design — i.e., includes "pair" fixed effects — and includes student and school controls for year 1 and year 3 respectively. The treatment-on-the-treated effect (Column 2 and Column 3 for year 1 and year 3 respectively) is estimated using the assigned treatment as an instrument for whether the student is in fact enrolled in a PSL school during the 2016/2017 academic year. Panel A uses our default IRT model and normalizes test scores using the same mean and standard deviation across all grades. Panel B uses the same IRT model as Panel A, but normalizes test scores using a different mean and standard deviation for each grade. Panel C estimates students' ability as the first component from a principal component analysis (PCA), and normalizes test scores using a common mean and standard deviation across all grades. Panel D uses the same model as Panel C but normalizes test scores using a different mean and standard deviation per grade. Panel E calculates the percentage of correct responses. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.4: Student selection

	Same school (1)	Same school (2)	Same school (3)
Panel A: Year 1			
Treatment	0.061	0.012	0.021
	(0.082)	(0.026)	(0.019)
Treatment × Age	-0.0042		
	(0.0064)		
Treatment × Male		-0.011	
		(0.028)	
Treatment × Asset Index (PCA)			-0.0059
			(0.011)
Observations	3,487	3,487	3,428
Panel B: Year 3			
Treatment	0.32***	0.036	0.026
	(0.12)	(0.029)	(0.019)
Treatment × Age	-0.020**		
	(0.0076)		
Treatment × Male		-0.023	
		(0.038)	
Treatment × Asset Index (PCA)			-0.017
			(0.014)
Observations	3,510	3,510	3,421

Notes: Panel A presents estimations for year 1 and Panel B for year 3. The outcome variable is whether the student is enrolled at the end of the 2016/2017 school year in the same schools he or she was enrolled in the 2015/2016 school year. All regressions include "pair" fixed effects. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.5: Student selection still going to school

	In school (1)	In school (2)	In school (3)
Panel A: Year 1			
Treatment	1.76	1.58	1.81**
	(6.32)	(1.49)	(0.86)
Treatment × Age	-0.039		
	(0.50)		
Treatment × Male		-0.57	
		(1.79)	
Treatment × Asset Index (PCA)			0.041
			(0.63)
Observations	3,487	3,487	3,288
Panel B: Year 3			
Treatment	15.2*	-6.97***	-3.67***
	(8.76)	(2.19)	(1.19)
Treatment × Age	-1.23**		
	(0.59)		
Treatment × Male		6.92**	
		(2.92)	
Treatment × Asset Index (PCA)			-0.79
			(0.91)
Observations	3,510	3,510	3,316

Notes: Panel A presents estimations for year 1 and Panel B for year 3. The outcome variable is whether the student is enrolled at the end of the 2016/2017 school year in the same schools he or she was enrolled in the 2015/2016 school year. All regressions include "pair" fixed effects. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.6: Intensive margin effect on teacher attendance and classroom observation with Lee bounds

	Year 1			Year 3		
	Control	Treatment Effect	90% CI (bounds)	Control	Treatment Effect	90% CI (bounds)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Spot check						
% on schools campus	52.40	14.17***	2.67	61.49	0.46	-5.84
	(50.00)	(3.75)	27.96	(48.74)	(4.54)	9.80
% in classroom	41.05	9.96**	-1.21	49.84	4.31	-2.23
	(49.25)	(3.86)	24.26	(50.08)	(4.95)	13.56
Observations	458	929	929	309	654	654
B: Classroom observation						
Active instruction (% class time)	30.13	7.62	-4.75	43.94	2.11	-11.70
	(32.11)	(4.75)	19.92	(27.22)	(4.62)	12.12
Passive instruction (% class time)	12.80	4.72	-4.93	25.15	4.74	-8.22
	(19.83)	(3.23)	9.62	(21.43)	(3.77)	11.85
Classroom management (% class time)	10.67	10.33***	0.77	13.94	1.40	-4.68
	(14.83)	(3.32)	16.99	(15.78)	(3.15)	11.22
Teacher off-task (% class time)	46.40	-22.66***	-40.24	16.97	-8.25*	-17.57
	(41.09)	(6.26)	-10.32	(27.40)	(4.14)	-0.71
Student off-task (% class time)	57.60	-5.19	-16.05	38.64	-1.93	-17.25
	(34.87)	(4.88)	12.63	(32.00)	(4.93)	8.72
Observations	71	143	143	66	114	114
Panel C: Inputs						
Number of seats	20.58	0.58	-7.22	18.62	4.07*	-1.31
	(13.57)	(1.90)	5.36	(11.74)	(2.18)	9.95
% with students sitting on the floor	4.23	-1.51	-7.48	3.03	0.00	-6.06
	(20.26)	(2.61)	2.76	(17.27)	(2.59)	2.17
% with chalk	78.87	16.58***	9.47	93.94	3.64	-0.28
	(41.11)	(5.50)	27.85	(24.04)	(3.64)	10.38
% of students with textbooks	17.60	22.60***	-1.21	9.55	8.29	-10.97
	(35.25)	(6.32)	34.87	(21.59)	(5.43)	21.82
% of students with pens/pencils	79.67	8.16*	1.36	86.92	4.19	-5.31
	(30.13)	(4.10)	20.98	(24.47)	(4.40)	15.80
Observations	71	143	143	66	114	114

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1 in Year 1 and Column 4 in Year 3), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2 in Year 1 and Column 5 in Year 3. Column 3 and 6 show the 90% confidence interval using Lee (2009) bounds for Year 1 and Year 3, respectively. Panel A provides results from the spot check using the EMIS data (2015/2016) on teachers as a baseline, and treating teachers who no longer teach at school as attriters. Panel B and C provide the classroom observation information without imputing values for schools not in session during our visit, and treating the missing information as attrition. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.7: Treatment effect on schools' good practices (%)

	Year 1		Year 3	
	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)
Maintains an enrollment log	80.43	10.11*	68.13	16.79***
	(39.89)	(5.40)	(46.85)	(6.25)
Log contains student name	81.52	7.58	71.43	14.60**
	(39.02)	(5.15)	(45.43)	(6.31)
Log contains student grade	83.70	9.75**	69.23	15.69**
	(37.14)	(4.87)	(46.41)	(6.38)
Log contains student age	64.13	0.00	43.96	21.53***
	(48.22)	(6.87)	(49.91)	(7.51)
Log contains student gender	82.61	6.50	65.93	14.60**
	(38.11)	(5.50)	(47.66)	(6.50)
Log contains student contact information	13.04	12.64**	13.19	17.15***
	(33.86)	(5.58)	(34.02)	(6.32)
Enrollment log is clean and neat	26.09	12.64*	35.16	11.31
	(44.15)	(6.73)	(48.01)	(7.51)
Maintains official schedule	89.13	8.66***	83.52	12.77***
	(31.30)	(2.94)	(37.31)	(4.31)
Official schedule is posted	69.57	14.44**	69.23	16.06***
	(46.27)	(5.92)	(46.41)	(5.77)
Has a PTA	97.83	1.08	96.70	3.28*
	(14.66)	(1.88)	(17.95)	(1.87)
Principal has PTA head's number at hand	26.09	14.80**	35.16	12.77*
	(44.15)	(6.31)	(48.01)	(6.80)
Maintains expenditure records	8.70	5.05	19.78	-1.09
	(28.33)	(4.81)	(40.05)	(5.70)
Maintains a written budget	21.74	4.33	36.67	-2.58
	(41.47)	(5.73)	(48.46)	(7.01)
Observations	92	185	90	181

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1 in Year 1 and Column 3 in Year 3), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2 in Year 1 and Column 4 in Year 3. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.8: Probability of still going to school

	Control	Treatment Effect
	(1)	(2)
Panel A: Grade 4th and 5th in 2015		
Goes to the same school	10.14	-0.35
	(30.22)	(2.26)
Still goes to school	80.17	-10.05***
	(39.91)	(2.90)
Observations	447	919
Panel B: Grade $<$ 4th grade in 2015		
Goes to the same school	51.05	2.53
	(50.01)	(2.21)
Still goes to school	85.65	-1.67
	(35.07)	(1.17)
Observations	1,282	2,591

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2. Panel A restricts the sample to students who were in 4th grade or higher in 2015. Panel B restricts the sample for those students who where in 3rd grade or lower in 2015. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.9: Probability of dropping out of school

	Control	Treatment Effect
	(1)	(2)
Work	0.49	0.23
	(6.99)	(0.20)
Pregnancy	3.11	2.32***
	(17.37)	(0.62)
Parents can't afford school	4.93	0.87
	(21.66)	(0.69)
Observations	1,730	3,511

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2. The sample is restricted to students who were in 4th grade or higher in 2015. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.10: Dropout reasons by provider (Randomization Inference)

	BRAC (1)	Bridge (2)	MtM (3)	Omega (4)	Rising (5)	St. Child (6)	Stella M (7)	YMCA (8)
Panle A: Dropout reasons								
Work	0.028	1.885	-0.022	0.943	-0.049	-1.617	-0.069	-0.056
	[0.127]	[0.404]	[0.641]	[0.413]	[0.610]	[0.543]	[0.117]	[0.110]
Pregnancy	1.562	16.029	4.807	2.020	19.033	3.568	-17.287	25.606
	[0.636]	[0.041]	[0.381]	[0.779]	[0.235]	[0.635]	[0.739]	[0.163]
Parents can't afford school	-0.009	-1.796	0.620	0.174	-12.481	-7.390	-27.335	-10.972
	[0.998]	[0.679]	[0.794]	[0.977]	[0.192]	[0.463]	[0.691]	[0.279]
Panel B: Constrained classes								
Goes to the same school	-37.276	-3.278	0.000	15.574	0.000	0.000	-17.357	0.000
	[0.209]	[0.680]	[.]	[0.427]	[.]	[.]	[0.407]	[.]
Still goes to school	-8.226	-4.746	0.000	5.689	0.000	0.000	-28.658	0.000
	[0.115]	[0.464]	[.]	[0.616]	[.]	[.]	[0.322]	[.]
Number of schools	40	45	12	38	10	24	8	8

Notes: This table presents the treatment effect for each provider on different outcomes using randomization inference. The estimates for each provider are not comparable to each other without further assumptions, and thus we do not include a test of equality. The randomization inference uses 5000 iterations to calculate p-values based upon the distribution of squared tstatistics following A. Young (2018). Standard errors (not shown) are clustered at the school level. Number of observations refers to the number of schools in both treatment and control groups. The sample is restricted to students who were in 4th grade or higher in 2015.

Table A.11: Probability of attending secondary school

	Control (1)	Treatment Effect (2)
Panel A: All students		
Attending secondary school	17.57	-2.19**
	(38.07)	(0.91)
Observations	1,780	3,624
Panel B: Grade 4th and 5th in 2015		
Attending secondary school	66.80	-11.37***
	(47.14)	(3.45)
Observations	456	947

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2. In Panel B, the sample is restricted to students who were in 4th grade or higher in 2015. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.12: ITT treatment effects on household behavior, fees, and student attitudes

	Year 1		Year 3	
	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)
Panel A: Fees				
% with > 0 ECE fees	30.77	-18.98***	52.75	-21.03***
	(46.41)	(5.42)	(50.20)	(6.75)
% with > 0 primary fees	29.67	-16.79***	58.24	-22.99***
	(45.93)	(5.71)	(49.59)	(6.23)
ECE Fee (USD/year)	1.42	-0.87***	0.26	-0.15
	(2.78)	(0.33)	(0.77)	(0.10)
Primary Fee (USD/year)	1.22	-0.70**	0.32	-0.21*
	(2.40)	(0.31)	(0.87)	(0.11)
Observations	91	184	90	183
Panel B: Student attitudes				
School is fun	0.53	0.05**	0.53	0.02
	(0.50)	(0.02)	(0.50)	(0.02)
I use what I'm learning outside of school	0.49	0.04***	0.54	-0.00
	(0.50)	(0.02)	(0.50)	(0.02)
Observations	1,713	3,492	1,730	3,510

Notes: This table presents the mean and standard deviation (in parentheses) for the control (Column 1 in Year 1 and Column 3 in Year 3), as well as the the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) in Column 2 in Year 1 and Column 4 in Year 3. Panel A presents data from household surveys. Panel B presents data from school principals on what fees schools charge. Panel C presents data on whether students agree or disagree with several statements. Standard errors are clustered at the school level. Standard errors are clustered at the school level. Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.13: Gender based violence survey - Balance response rate

	All students		Girls		Boys	
	Control	Treatment Effect	Control	Treatment Effect	Control	Treatment Effect
	(1)	(2)	(3)	(4)	(5)	(6)
Responde rate	0.89	-0.02	0.92	-0.02	0.93	-0.03
	(0.09)	(0.01)	(0.12)	(0.02)	(0.11)	(0.02)

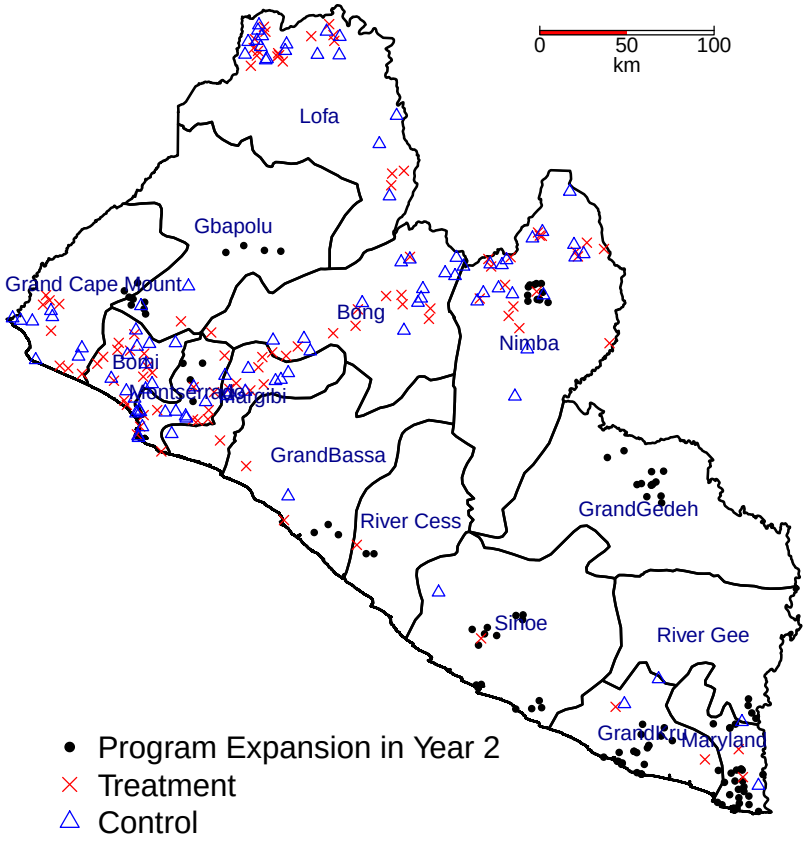
Notes: This table presents the mean and the standard deviation (in parentheses) for the control as well as the difference and the standard error of the difference (in parentheses) taking into account the randomization design (i.e., including "pair" fixed effects) for all students (Column 1-2), only girls (Column 3-4), and only boys (Columns 5-6). Statistical significance at the 1, 5, 10% levels are indicated by ***, **, and *, respectively.

Table A.14: External validity: Differences in characteristics of schools in the PSL program (new schools) and other public schools

	PSL program expansion (1)	Other public schools (2)	Difference (3)
Students: ECE	110.92	108.02	9.71
	(74.33)	(65.35)	(7.51)
Students: Primary	138.56	127.41	24.47*
	(145.03)	(154.90)	(13.29)
Students	221.31	225.58	21.37
	(160.95)	(173.64)	(14.56)
Classrooms per 100 students	1.18	0.81	0.52**
	(2.24)	(1.96)	(0.24)
Teachers per 100 students	3.80	3.95	0.07
	(2.78)	(15.47)	(0.29)
Textbooks per 100 students	112.80	105.68	14.56
	(113.20)	(194.44)	(10.76)
Chairs per 100 students	16.77	14.18	7.77***
	(27.83)	(41.89)	(2.63)
Food from Gov or NGO	0.56	0.31	0.21***
	(0.50)	(0.46)	(0.05)
Solid building	0.28	0.27	0.09**
	(0.45)	(0.44)	(0.04)
Water pump	0.50	0.45	0.10**
	(0.50)	(0.50)	(0.05)
Latrine/toilet	0.81	0.70	0.14***
	(0.39)	(0.45)	(0.04)
Observations	123	1,597	1,720

Notes: This table presents the mean and standard deviation of the mean (in parentheses) for schools in the PSL program (Column 1) and other public schools (Column 2), as well as the difference in means and standard error of the mean across both groups (Column 3). ECE = Early childhood education. Authors' calculations based on 2015/2016 EMIS data. Sample is restricted to those counties where the expansion of the program took place. Regressions include county fixed effects. Standard errors are clustered at the school level.

Figure A.1: Program schools



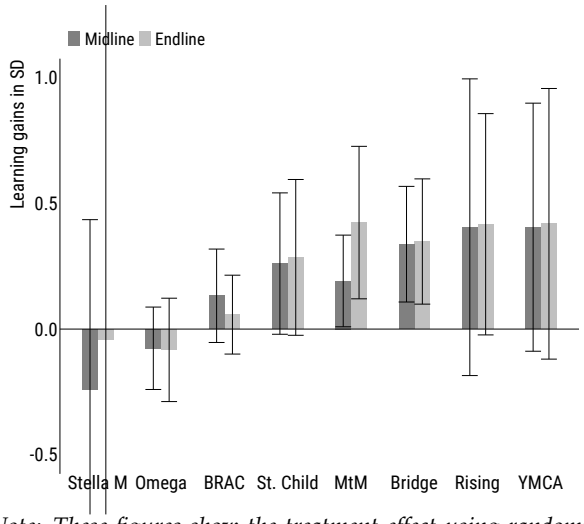
Notes: This shows the schools in the experiment (both treatment and control), as well as the schools that became program schools after the expansion of the program in 2017 to an additional 98 schools. The schools that were assigned to providers in 2017 are not experimentally assigned nor embedded into the randomized evaluation. Thus, the results in this report do not speak to the treatment effect in these schools.

Figure A.2: Timeline

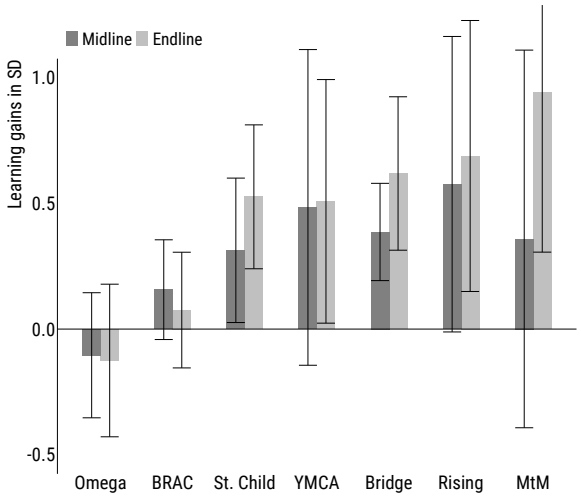
Research Activities	Year	Month	Intervention Activities
	2016	Jun	Operator selection
Randomization		Jul	
		Aug	
First Wave		Sep	School year begins
		Oct	
		Nov	
		Dec	
	2017	Jan	
		Feb	
		Mar	
		Apr	
Second Wave		May	
		Jun	
		Jul	School year ends
		Aug	
		Sep	School year begins
		Oct	
		Nov	
		Dec	
	2018	Jan	
		Feb	
		Mar	
		Apr	
		May	
		Jun	
		Jul	School year ends
		Aug	
		Sep	School year begins
		Oct	
		Nov	
		Dec	
	2019	Jan	
		Feb	
Third Wave		Mar	
		Apr	
		May	

Figure A.3: Treatment effects by provider

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

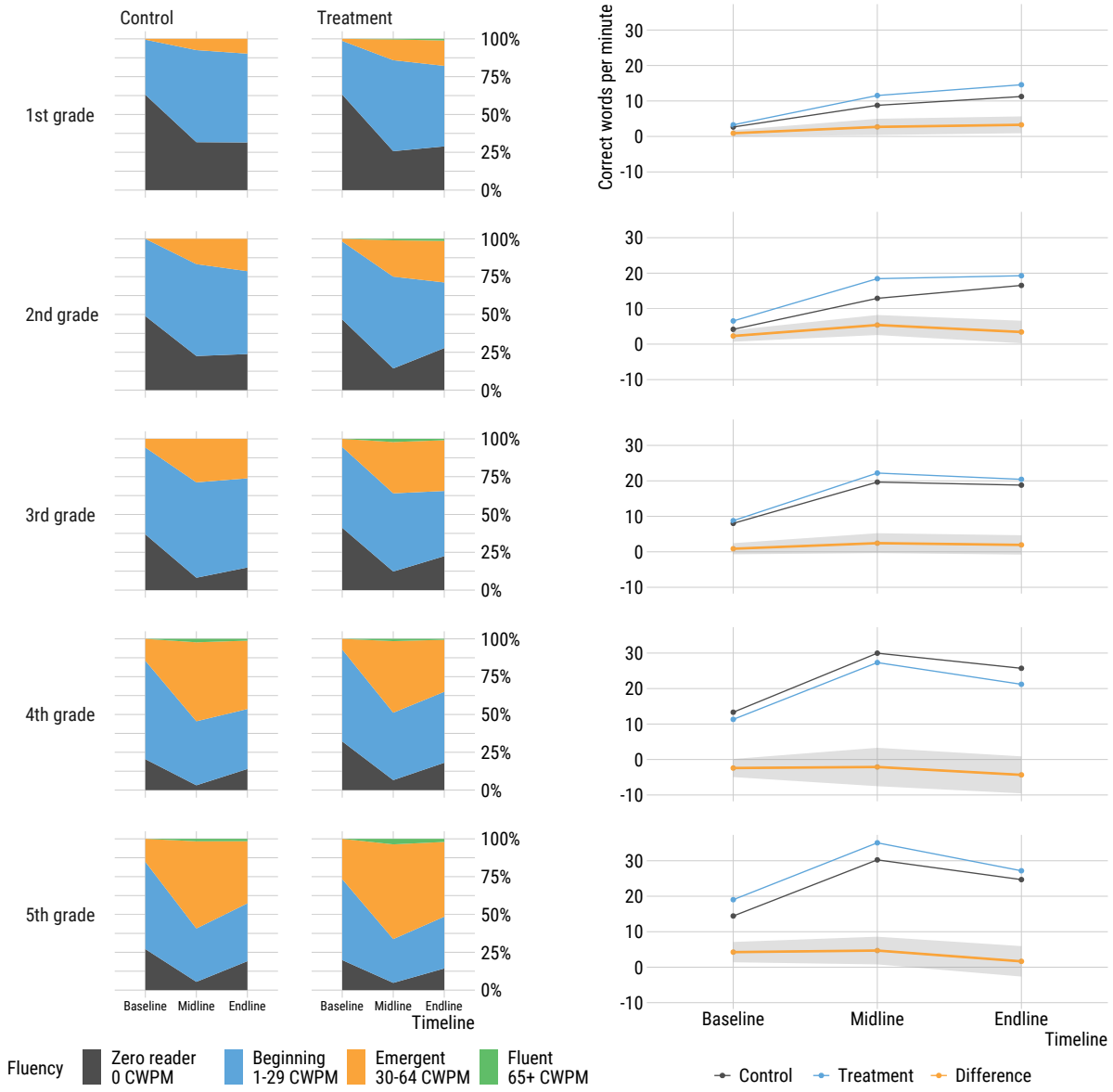


(b) Treatment-on-the-treated effect (ToT)



Note: These figures show the treatment effect using randomization inference for both midline and endline. The randomization inference uses 5000 iterations to calculate p-values based upon the distribution of squared tstatistics following A. Young (2018). Figure A.3a shows the intention-to-treat (ITT) effect, while Figure A.3b 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.

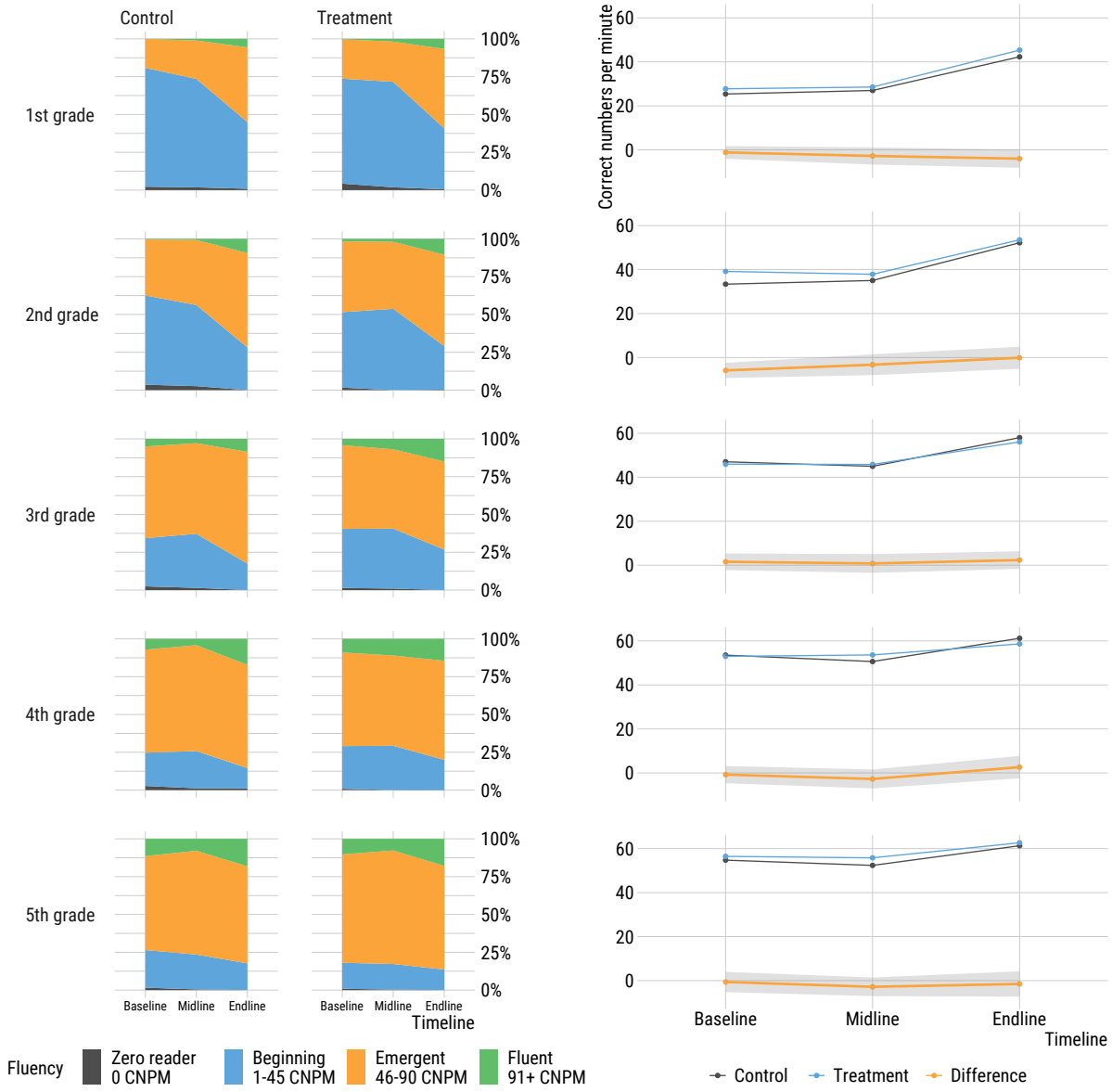
Figure A.4: Correct words per minute by grade - all operators



(a) Correct words per minute by fluency groups (b) Difference between treatment and control

Notes: The figure presents the correct words per minute grouped by fluency for treatment and control groups, and the estimated difference in each wave.

Figure A.5: Correct numbers per minute by grade - all operators



(a) Correct numberes per minute by fluency groups (b) Difference between treatment and control

Notes: The figure presents the correct numbers per minute grouped by fluency for treatment and control groups, and the estimated difference in each wave.

B Gender based violence survey

In each school students who are 12 years old or older, and who are part of the original sample, were surveyed regarding gender roles and sexual abuse. This module had its own assent form and it is independent of the student evaluation (i.e., a student can decide to participate in the student assessment and refuse to participate in the sexual violence survey).

This survey was a shorter and adapted version of a survey previously used in primary schools of Liberia (Postmus et al., 2015; Steiner et al., 2018) (see Figure B.6). Students were instructed to not provide name or any other personal identifiable information. In addition, they were told that they do not need to take part of the survey if they do not want to and that they can skip questions if they want to. Since the purpose of this is to obtain school level averages, rather than link student performance to sexual abuse, this data did not have student identifiers and will never be linked to any other data set.

An "answer key" was distributed to the students; those who wish to not participate received a referral pathways and the enumerator concluded the interview. These answer keys did not have any identifying information on the students. A member of the IPA team read the questions and gave enough time for students to complete the answer key anonymously. Once all of the questions had been read, the enumerator instructed the student to put the anonymous sheet in a specially marked and sealed envelope.

Note that some of the questions are not related to sexual violence and were intended to serve as measures of whether the students were understanding the instructions of the survey module or not. Lastly, note that "men and women business" is used through the survey instead of sexual intercourse an appropriate phrasing for children in Liberia.

In addition to this, all students receive referral information about different organizations that they can contact in case that they want to report any occurrences of sexual violence.

Figure B.6: GBV survey instrument

Now, I want to talk to you about men and women business by force. Girls and boys, and women and men, may experience unwanted sexual contact by people they know well, such as a romantic/love partner, a trusted adult, and a family member or friend, or by strangers. We also want to talk about your experiences related to sexual violence that other students, like yourself, have reported to us, so I will ask you some questions about your personal experience. Remember that we won't tell anybody your answers – your teachers won't know, your family won't know, your classmates won't know, and you can skip any questions that you don't feel comfortable answering. Even I won't know how you answered. Remember that each time I say "Since coming to this school" I am referring to your most recent school, even if you are no longer attending and when I say "I" I am talking about you.

Questionnaire A

		YES/NO
1.	Do you live in America?	
2.	6+5 equals 11	
3.	Do you smile at people every time you meet them?	
4.	Do you always keep your promises?	
5.	Would you ever lie to people?	

Since coming to [this school] has a teacher or staff member:		YES/NO
6.	done men and women business with you?	
7.	Touched you on your body when you didn't want it? (penis, butt for boys) (breasts, butt, vagina for girls)	
8.	Forced you to do men and women's business when you didn't want to?	

Since coming to [this school] has a student:		YES/NO
9.	Touched you on your body when you didn't want it? (penis, butt for boys) (breasts, butt, vagina for girls)	
10.	Forced you to do men and women's business when you didn't want to?	

Since coming to [this school] has someone in your house:		YES/NO
11.	Touched you on your body when you didn't want it? (penis, butt for boys) (breasts, butt, vagina for girls)	
12.	Forced you to do men and women's business when you didn't want to?	

Question 13:

1. I live in America
2. I am older than 11

Question 14:

1. I am alive
2. I live in Ghana
3. I am six years old

Question 15:

1. I miss school days many times
2. I brushed my teeth today

Question 16:

1. I am wearing socks
2. My favorite color is blue

Question 17:

1. I always tell the truth
2. I enjoy studying
3. I am a female

Question 18:

1. I think people should pray
2. Since coming to this school I have done men and women's business with a teacher or staff member
3. I prefer banana over pineapple