THE ECONOMIC JOURNAL

The Economic Journal, 1–20 https://doi.org/10.1093/ej/ueab087
Advance Access Publication Date: 11 November 2021

BEYOND SHORT-TERM LEARNING GAINS: THE IMPACT OF OUTSOURCING SCHOOLS IN LIBERIA AFTER THREE YEARS*

Mauricio Romero and Justin Sandefur

Outsourcing the management of ninety-three randomly-selected government primary schools in Liberia to eight private operators led to learning gains of 0.18σ after one year, but these effects plateaued in subsequent years (reaching 0.2σ after three years). Beyond learning gains, the programme 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. Despite facing similar contracts and settings, some providers produced uniformly positive results, while others presented trade-offs between learning gains, access to education, child safety, and financial sustainability.

Public–private partnerships in education are common around the world (Patrinos *et al.*, 2009; Aslam *et al.*, 2017). While canonical results from contract theory might suggest education is well suited for outsourcing (Hart *et al.*, 1997), the required assumptions may not always hold in practice. Governments may struggle to observe key outputs (e.g., child safety or non-cognitive skills) and a narrow focus on raising test scores may result in a multitasking problem (Holmstrom and Milgrom, 1991). In rural areas where alternatives are limited, outsourcing may generate private monopolies. While dynamic incentives in repeated contracting may overcome some of these pitfalls (Banerjee and Duflo, 2000; Corts and Singh, 2004), governments may be unwilling or unable to sanction private contractors' bad performance.

In 2016, the Liberian government outsourced the management of 93 randomly selected public schools, comprising 8.6% of public school students, to eight private providers. The programme

* Corresponding author: Justin Sandefur, Center for Global Development, 2055 L St NW 5th floor, 20036, Washington, DC, USA. Email: jsandefur@cgdev.org

This paper was received on 15 January 2021 and accepted on 3 November 2021. The Editor was Steffen Huck.

The data and codes for this paper are available on the Journal repository. They were checked for their ability to reproduce the results presented in the paper. The replication package for this paper is available at the following address: https://doi.org/10.5281/zenodo.5579799.

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 programme. 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 Innovations for Poverty Action team, 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 randomised controlled trials registry entry is available at: https://www.socialscienceregistry.org/trials/1501 as well as the pre-analysis plan. Replication data and code is available at https://doi.org/10.7910/DVN/5OPIYU and https://doi.org/10.5281/zenodo.5579799. 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 views expressed here are ours, not those of the Ministry of Education of Liberia or our funders. All errors are our own.

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.

bundled private management with, in theory, a doubling of education expenditure per child. After one academic year, students in outsourced schools scored 0.18σ higher in English and math. However, costs surpassed original projections and some providers engaged in unforeseen and harmful behaviour, including mass removal of students and efforts to conceal sexual abuse allegations, complicating any assessment of long-term welfare gains (Romero *et al.*, 2020).

In this paper, we study how the short-term effects changed after three years, focusing on heterogeneity in contracted and non-contracted outcomes across providers. There are reasons to expect that the programme's performance would have changed after three years. Both learning-by-doing and selective contract renewal—in which successful providers were rewarded with more schools—point toward larger impacts on the programme's primary goal of raising learning outcomes. Furthermore, interim evaluation results in Romero *et al.* (2020) and media coverage of various programme failings in the first year (Rosenberg, 2016; Kristof, 2017; Pilling, 2017; *The Economist*, 2017; Tyre, 2017) provided opportunities to revise contracts and mitigate unintended consequences in later years. However, reduced media scrutiny or gradual recognition by providers that the government would not sanction failures, may have led to deteriorating quality.

Treatment effects on test scores remained statistically significant after three years, but plateaud after the first year—reflecting both limited decay for those who graduated after one or two years and limited additional gains for those exposed for all three years. Beyond learning outcomes, outsourcing increased dropout rates among the students originally in treatment schools. While the outsourcing programme reduced corporal punishment, it did not change rates of sexual violence perpetrated by school staff. Finally, average effects concealed different results across providers in all dimensions. While some providers produced uniformly positive results, even on non-contracted outcomes, others presented stark trade-offs between learning gains and other goals.

In terms of additional direct costs of the programme, the ministry expects providers to operate for US$50 or less in the long term. While self-reported unit costs have fallen for most providers over time, they remain at least three times as much as the government target for some providers.

We complement the large literature on US charter schools (see Betts and Tang, 2014; Chabrier *et al.*, 2016, for reviews), and join the growing literature studying outsourcing in other settings (e.g., Barrera-Osorio, 2007; Bonilla, 2010 in Colombia; Eyles and Machin, 2019 in the UK; Barrera-Osorio *et al.*, Forthcoming in Pakistan). We also add to the literature highlighting some of the unintended consequences of multitasking (Holmstrom and Milgrom, 1991) and contract incompleteness (Hart *et al.*, 1997), showing that in our setting outsourcing leads to increased dropout rates.[1] Finally, while heterogeneity in charter school performance has been documented elsewhere, earlier work has emphasised differences in students and contracts (Chabrier *et al.*, 2016). Here, under the same contract and in similar contexts, private contractors' identity matters, suggesting that selecting providers aligned with the public interest (à la Akerlof and Kranton, 2005; Besley and Ghatak, 2005) may be key for public–private partnerships in education.

1. Research Design

Below we summarise the most important features of the programme and the experimental design. Romero *et al.* (2020) provides further details.

[1] For example, Hsieh and Urquiola (2006) show vouchers induce sorting in Chile; and Bergman and McFarlin (2018) document discrimination in admissions in charter schools in the United States.

1.1. *The Programme*

1.1.1. *Context*

The government's primary motivation for the outsourcing programme was the low levels of learning in public schools. At baseline, ∼25% of pupils enrolled in fifth grade could not read a single word. In addition, access remains an unresolved issue. The last nationally representative household survey before the experiment reported net primary enrolment at 38%, partially explained by high levels of over-age enrolment (Liberia Institute of Statistics and Geo-Information Services, 2016).

Public primary school is nominally free in Liberia, though informal fees are standard. In contrast, fees are permitted for pre-primary classes. At the baseline, government spending on public primary schools was ∼US$50 per pupil, almost entirely allocated to teacher salaries.

1.1.2. *Intervention*

The Liberian Education Advancement Programme (LEAP)—formerly, the Partnership Schools for Liberia (PSL) programme—is a public–private partnership for school *management*. Under the programme, the government delegated 93 public schools' management, covering 8.6% of all public school students, to eight private organisations. Providers were paid on a per-pupil basis and forbidden from charging fees or screening students based on ability.

Of the eight providers, three are for-profit and two (both non-profits) are Liberian organisations. See Table 6 in the Online Appendix for a list of providers and sample sizes.

In contrast to some other public–private partnerships in education (e.g., US charter schools), teachers (and school principals) in outsourced Liberian public schools remained civil servants and on the government payroll.

There are three noteworthy features of the evolution of the programme since it started in 2016. In 2017, the programme expanded to an additional 98 schools. These schools were not experimentally assigned and are not included in our analysis (see Figure 2 in the Online Appendix). Second, the programme changed some of its operating rules. All providers were given uniform contracts (unlike the first year, when Bridge had a different contract) and the Ministry of Education did not allow capping class sizes.[2] Finally, the country had a presidential election in late 2017.[3] The new administration, which took office in early 2018, claims it stopped prioritising treatment schools in the assignment of teachers or those in the process of bringing existing teachers onto the payroll.

Providers must teach the Liberian national curriculum but have flexibility in defining the intervention. They may choose how to use school resources (e.g., providing remedial programmes, prioritising 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 what activities the provider engaged in. We summarise teachers' responses in Figure 3 in the Online Appendix, which shows considerable variation in the specific activities and providers' total activity level.

On paper, the Liberian government's financial obligation to treatment schools is the same as any other public school: it provides teachers and maintenance valued at about US$50 per student. In addition, providers receive *extra* funding (US$50 per student), coordinated by the Ministry of Education but paid by third-party philanthropy. Providers have complete autonomy over the use

[2] This had little effect as student expulsions in the first year meant few classes remained above the cap.
[3] Sandholtz (2020) examines whether the policy created incentives for politicians to adopt it.

of these funds. Providers may also raise more funds on their own (see Section 3 for more details on providers' costs).

The contracts between the government and the providers specified a set of key performance indicators (KPIs) through which providers' performance was to be measured. While these KPIs included learning outcomes and access (e.g., enrolment and retention), they did not mention child safety (e.g., corporal punishment or sexual abuse) or sustainability (e.g., cost per pupil or teacher retention). In addition, while access was technically part of the KPIs, in practice the focus placed by the government and by outside donors was on learning outcomes (e.g., Ministry of Education—Republic of Liberia, 2016; Werner, 2017a,b; Werner and Malkus, 2017; Starr, 2017; 2020). Both the Ministry and outside donors disregarded news that Bridge International Academies was enforcing enrolment caps in many schools, forcing many children to find a new school (e.g., Senah, 2016; Mukpo, 2017; Werner, 2017a). Thus, we treat learning as the primary contracted outcome and enrolment, child safety, and sustainability as non-contracted outcomes.[4]

1.2. *Experimental Design*

1.2.1. *Sampling and random assignment*

Two key features of the sampling and randomisation process are that (1) providers agreed to a list of schools they would be willing to serve before random assignment and (2) pupils were sampled from lists made before the programme began and tracked regardless of where they went.

Based on providers' preferences and requirements, 185 eligible schools was non-randomly partitioned across providers. The schools allocated to each provider were paired based on their infrastructure quality. Within each pair, we randomly assigned schools to treatment or control. Providers did not manage all the schools initially assigned to treatment, and we treat these schools as non-compliant, presenting results in an intention-to-treat framework (Table 6 in the Online Appendix provides more details).

Treatment assignment may change the student composition across schools. To prevent differences in students' composition from driving differences in outcomes, we sampled 20 students per school (from K1 to grade 5) from enrolment logs from the 2015/2016 school year, before the treatment was introduced. We associated each student with their 'original' school, regardless of what school (if any) they 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 unbiased estimates of the programme's intention-to-treat (ITT) effect within the student population originally attending study schools.

1.2.2. *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 2019. Figure 4 in the Online Appendix provides a detailed timeline.

[4] The contracts stated that 'The Government of Liberia reserves the right to revoke the contract of any school or operator deemed to be below standard' and that 'Performance of these schools, as measured by the KPIs and internal GoL quality assurance process will form the basis of any extension of contract length and expansion or reduction in the number of schools allocated to any one provider'.

1.2.3. *Test design*

In all three rounds of data collection, we conducted one-on-one tests because literacy cannot be assumed at any grade level. All students took the same adaptive test, regardless of the grade. However, the test items are different across survey rounds. Stop rules instructed enumerators to skip higher-order skills if students cannot answer questions related to more basic skills. We estimate a separate item response theory (IRT) model for each round of data collection. We normalise the IRT scores relative to the control group.

1.2.4. *Additional data*

Given the concerns about child safety in programme schools raised by the sexual abuse scandals involving two of the providers (Baysah, 2016; Young, 2018), we added a sexual violence module to the student survey during the endline (in 2019). Sexual abuse is difficult to measure and rarely reported through official channels in Liberian schools. We collected data via an anonymous survey of students twelve years old and above. Enumerators asked students 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. Students could opt out—the response rate is close to 90% and balanced between treatment and control schools (see Table 7 in the Online Appendix). Online Appendix B provides more details on the survey protocol and instrument.

See Romero *et al.* (2020) for details of the other teacher and school survey modules, including classroom observation.

1.2.5. *Balance and attrition*

Romero *et al.* (2020) show that time-invariant school and student characteristics, as well as pre-treatment administrative measures of school characteristics, are balanced across treatment and control schools.

Students were tracked to their homes and tested there. Attrition from our original sample is balanced between treatment and control in both follow-up rounds and is below 4% in the second wave and below 3% in the third wave (see Table 8 in the Online Appendix).

2. Overall Policy Impact

We estimate the programme's overall impact before turning to specific providers' impact in the following section. In both cases, we focus on four margins: (1) *learning*, as measured by test scores; (2) *sustainability*, which hinges, in part, on whether the programme effects come from increases in material inputs or staffing versus improvements in school management; (3) *access*, defined as impacts on enrolment and grade attainment; and (4) *child safety*, as measured by pupil surveys on corporal punishment and sexual abuse. We estimate the effects via ordinary least squares, clustering the standard errors at the school level. Replication data is available at Romero *et al.* (2018) and Romero *et al*. (2021).

The Liberian government and its philanthropic backers contracted private operators to manage public schools with the explicit, primary goal of raising learning outcomes. In practice, test-score gains plateaued after the first year. We explore why this may be the case at the end of this section, but a change in teacher behaviour may explain this levelling off. While teacher behaviour and pedagogy are still better in treatment schools, the pedagogical advantage attenuated after the first

year. In addition, several of the private operators' expenditures per pupil fell dramatically after the first year.

The government and funders' intense focus on learning gains may have contributed to adverse or null effects of the programme on non-contracted outcomes (access, child safety, and sustainability). Providers who expelled students en masse in the first year were not sanctioned. The government and funders did not revise the key performance indicators to include child safety, despite credible sexual abuse allegations against some providers' staff. Perhaps unsurprisingly, early signs of deleterious effects on access, detected after one year, are confirmed after three years, and the programme did not reduce sexual abuse, despite the influx of resources and external oversight. We do, however, observe a decrease in corporal punishment, another non-contracted outcome.

2.1. *Learning*

As noted in the introduction, there are reasons to anticipate either increased impacts of the programme over time (e.g., learning-by-doing and selective contract renewal of the best operators) and the opposite (e.g., reduced media scrutiny or gradual recognition by providers that the government would not sanction failures).

Following our pre-analysis plan, we report intention-to-treat (ITT) estimates on learning gains from two specifications. The first specification amounts to a simple comparison of post-treatment outcomes for treatment and control individuals controlling for matched-pair fixed effects (i.e., stratification-level dummies). The second specification controls for time-invariant characteristics measured at the individual and school level (Table 17 in the Online Appendix provides a list of controls).

The ITT effect of the programme after three academic years is 0.16σ for English (p-value < 0.001) and 0.21σ for math (p-value < 0.001), as shown in Table 2 (Panel A, column 8). Treatment effects plateau after one year (when the treatment effects on English and math were 0.18σ and 0.18σ, as shown in column 5). The 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 columns 7 and 8. Our results are robust across different measures of student ability (see Table 18 in the Online Appendix for details).

The pattern of results across cohorts suggests limited learning decay after leaving treatment schools (and, conversely, limited marginal benefits from continued exposure). While the overall ITT effect includes students scheduled to graduate after one or two years, if we focus on students originally enrolled in lower grades (exposed for three years), the overall treatment effect is virtually unchanged (see Table 2, Panel B).[5]

While we focus on the ITT effect as the key policy-relevant parameter, we also report treatment-on-the-treated (ToT) estimates on learning outcomes. There are two sources of non-compliance in our experiment: school-level non-compliance when providers failed to take control of all of their assigned schools, and student-level non-compliance when students left their original school, either voluntarily or because providers excluded them. Given that non-compliance is unlikely to be random, we use the random assignment as an instrument for compliance to estimate the ToT. While the assumptions required for ToT estimates—monotonicity and the stable unit treatment value assumption (SUTVA)—appear reasonable at the school level, they may not be at the student

[5] These estimates combine differences in lengths of exposure and persistence as well as heterogeneous treatment effects by grade.

level. As shown below and in Romero *et al.* (2020), treatment caused some students to leave (voluntarily or not) their assigned schools, violating monotonicity. Also, a student's peers (and related peer effects) may have changed as some peers were forced out of school by the treatment, violating SUTVA at the student level. Thus, while we present ToT at the school level (see Table 2, Panels A and B) and ToT estimates at the student level (see Table 19 in the Online Appendix), the latter should be interpreted with caution.[6]

The ToT effect is 0.18σ for English (p-value < 0.001) and 0.24σ for math (p-value < 0.001), as shown in Table 2 (column 9). These school-level ToT estimates average over both student-level non-compliance and different exposure lengths. However, comparing the first- and third-year results suggests that most of the learning gains are accrued after one year of treatment.

An important concern when interpreting these results expressed in standard deviations is how much learning they represent. In general, standard deviations are not comparable across contexts (see Singh, 2015 for a discussion). Moreover, when counterfactual learning levels are low (and have low variance), large treatment effects in standard deviations could reflect modest learning in absolute terms (e.g., Eble *et al.*, 2021; Fazzio *et al.*, 2021). We use correct words per minute (WPM) as a benchmark. Students enrolled in Grade 1 in 2015/2016 in control schools can read 11 WPM on average in 2019. Their counterparts in treatment schools can read 15 WPM. For students enrolled in Grade 5 in 2015/2016, the difference between treatment (27 WPM) and control (25 WPM) in 2019 is less than 2 WPM. As a benchmark, to understand a simple passage students should read 45–60 WPM (Abadzi, 2011). Figure 5 in the Online Appendix provides the evolution of correct words per minute across survey rounds for different cohorts.

2.2. *Sustainability*

The outsourcing programme changed the management of treated schools, while also increasing the resources available to them. The sustainability of the programme 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 Subsection 2.3) and underperforming teachers to non-programme schools. Overall, while programme schools had more resources (e.g., more and better teachers), observable management practices also improved.

2.2.1. *Inputs and resources*

In the first year, the Ministry of Education agreed to release some underperforming teachers from programme schools, replace those teachers, and provide additional new teachers (Romero *et al.*, 2020). The net result was that programme schools had 2.6 more teachers on average (p-value < 0.001) after one year. After three years, providers still have 2.2 more teachers on average (p-value < 0.001; see Table 1, Panel C).

The composition of teachers also remains different in programme schools compared to control schools after three years due to the reshuffling of teachers in the first year (see Table 1, Panel D). The average teacher in a programme school is younger, less experienced, more likely to have private school experience and has higher test scores (based on memory, math, word association and abstract thinking tests).

[6] Apart from identification concerns, a student-level ToT estimate may also be less policy relevant, as the ethics of forcing students/parents to enrol in treatment schools or forbidding them to move are questionable.

Table 1. *ITT Treatment Effects on Enrolment and Teacher Characteristics.*

	Year 1		Year 3	
	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
Panel A: Enrolment and attendance data (school-level data)				
Enrolment 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: Enrolment and attendance (student-level data)				
% enrolled in the same school	83.16	0.71	41.04	2.13
	(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	3,622
Panel C: Teachers (school-level data)				
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 D: Teacher characteristics (teacher-level data)				
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 SDs	−0.01	0.14**	−0.03	0.21***
	(0.99)	(0.06)	(0.99)	(0.06)
Observations	489	1,167	478	1,142

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 treatment effect and its standard error (in parentheses) taking into account the randomisation 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 enrolment (taken from enrolment logs) and student attendance measure by our enumerators during a spot check. If the school was not in session during a regular school day we mark all students as absent. Panel B presents student-level data including whether the student is still enrolled in the same school, whether they are enrolled in school at all, and whether they have missed school in the previous week (conditional on being enrolled in school). Panel C has school-level outcomes related to the number of teachers. Panel D presents teacher-level outcomes including their score in tests conducted by our survey teams. Standard errors are clustered at the school level. Statistical significance at the 1, 5 and 10% levels is indicated by ***, **, and *, respectively.

2.2.2. *School management*

The treatment effects on school management are similar after one and three years of the programme (see Table 3, Panel D). Programme schools are more likely to be in session (i.e., school

Table 2. *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)
Panel A: All students									
English	0.09*	0.07**	0.08**	0.17***	0.18***	0.21***	0.15***	0.16***	0.18***
	(0.05)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Math	0.07	0.05	0.06*	0.19***	0.18***	0.22***	0.20***	0.21***	0.24***
	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Abstract	0.05	0.03	0.04	0.05	0.05	0.06	0.02	0.03	0.03
	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.03)	(0.03)	(0.04)
Composite	0.08*	0.06*	0.07*	0.19***	0.18***	0.22***	0.19***	0.20***	0.22***
	(0.05)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)
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
Panel B: Cohorts exposed for three years									
English	0.10**	0.06*	0.06*	0.18***	0.17***	0.19***	0.19***	0.17***	0.19***
	(0.05)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)
Math	0.06	0.02	0.02	0.18***	0.17***	0.18***	0.23***	0.21***	0.24***
	(0.05)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)
Abstract	0.04	0.01	0.01	0.01	0.02	0.02	0.00	0.00	0.00
	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
Composite	0.07	0.03	0.03	0.19***	0.17***	0.19***	0.22***	0.20***	0.23***
	(0.05)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Observations	2,598	2,598	2,598	2,579	2,579	2,579	2,590	2,590	2,590

Notes: In all regressions, the outcome is student test scores in standard deviations normalised to have mean zero and standard deviation of one in the control group in each wave of data collection. Columns 1–3 are based on the first wave of data and show the difference between treatment and control schools taking into account the randomisation 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 randomisation 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 randomisation 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 school is in fact treated. In Panel B the sample is restricted to students enrolled in Grade 3 or below in 2015/2016 (and thus who may still be enrolled in primary school by the third wave of data collection). Abstract refers to the score from a Raven's style module to measure the students' abstract thinking abilities. Standard errors are clustered at the school level. Statistical significance at the 1, 5 and 10% levels are indicated by ***, **, and *, respectively.

Table 3. *ITT Treatment Effects on Teacher Behaviour and School Management.*

	Year 1		Year 3	
	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
Panel A: Spot checks on teachers (school-level data)				
% 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 on teacher behaviour (student-level data)				
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 (school-level data)				
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
Panel D: School management (school-level data)				
% school in session at spot check	83.70	8.66*	75.82	12.41**
	(37.14)	(4.52)	(43.05)	(5.33)
Instruction time (hours/week)	14.69	3.17***	18.91	3.63***
	(4.04)	(0.65)	(5.74)	(0.86)
Principal's working time (hours/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 treatment effect and its standard error (in parentheses) taking into account the randomisation 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 during a school day. Panel B presents data from our panel of students where we asked them about their teachers' behaviour. 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 20 in the Online Appendix has the results without imputing values for schools not in session. Panel D presents data on management practices. The index of good practices is the first component of a principal component analysis (PCA) of the variables in Table 21 in the Online Appendix. The good practices index is normalised to have mean zero and standard deviation of one in the control group. The management index is based on DWMS-style questions, but restricted to multiple-choice questions. The management index is normalised to have mean zero and standard deviation of one in the control group. Standard errors are clustered at the school level. Statistical significance at the 1, 5 and 10% levels are indicated by ***, ** and *, respectively.

is open, students and teachers are on campus, and classes are taking place) during a regular school day: 8.7 percentage points (p-value = 0.058) after one year, and 12 percentage points (p-value = 0.022) after three years. School days are also longer: 3.2 more hours per week of instructional time (p-value = 0.0011) after one year and 3.6 more hours per week (p-value < 0.001) after three years.

Finally, management practices (as measured by a 'good practices' principal component analysis index normalised to a mean of zero and standard deviation of one in the control group) are 0.4σ (p-value = 0.0011) higher in programme schools after one year and 0.55σ (p-value < 0.001) higher after three years. In 2019 we also measure management practices using a Development World Management Survey (DWMS) style survey (Lemos and Scur, 2016). According to this index, management practices improved by 0.68σ (p-value < 0.001) after three years.

2.2.3. *Teacher behaviour*

A possible explanation for why treatment effects on test scores plateau after the first year (i.e., the gap between treatment and control remained constant) is that the treatment effects on teacher behaviour dissipate after the first year. We conducted unannounced spot checks of teacher attendance and collected student reports of teacher behaviour (see Table 3, Panels A and B). We also measured teacher time use and classroom management using the Stallings classroom observation instrument (see Table 3, Panel C).

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, this is no longer true after three years when the treatment effects are smaller and statistically insignificant. This does not seem to be driven by control schools improving teacher attendance.

Classroom observations also show a reduction in the treatment effects on teacher behaviour after three years. After one year teachers were 25 percentage points (p-value < 0.001) less likely to be off-task during class time; this effect decreased to 14 percentage points (p-value = 0.013) after three years. The reduction in treatment effects could be explained by control schools improving over time, where the time off-task went from 56% in 2016 to 40% in 2019.

2.3. *Access*

The programme reduced enrolment and increased dropout for the sample of students enrolled initially in treatment schools. Simultaneously, the programme had a positive treatment effect on total enrolment in treatment schools, implying that they pulled in new students from outside our baseline sample.[7]

Students enrolled in treatment schools in 2015/2016 are 3.3 percentage points (p-value = 0.0042) less likely to be enrolled in any school after three years, from a base of 85% (see Table 1, Panel B). This is not driven by students removed from their schools in the first year due to large class sizes (see Table 10 in the Online Appendix; see also Romero *et al.*, 2020, for a discussion of this issue). Instead, the effect seems to be driven by older students and by girls (see Table 11 in the Online Appendix).

Increased dropout is partially driven by pregnancies. Students originally enrolled in treatment schools are 2.33 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 (see Table 12 in

[7] Providers do not seem to engage in cream skimming. There is no evidence that any group of students is systematically excluded from treatment schools after three years (see Table 9 in the Online Appendix).

the Online Appendix).[8] This result need not imply providers increased the rate of teen pregnancy; alternatively, providers may enforce more the national policy requiring pregnant girls to drop out of school until childbirth (Martinez and Odhiambo, 2018).

The programme also reduced transition to secondary school. The effect is driven by students enrolled in Grade 4 and 5 in 2015/2016 (see Table 13 in the Online Appendix) who are 2.21 percentage points (p-value = 0.02) less likely to be enrolled in secondary school, from a base of 18% (see Table 14 in the Online Appendix). Notably, three treatment schools—assigned to Bridge International Academies at the company's request—had a secondary school on the same premises before the programme. After Bridge took control of these schools, it reassigned the classrooms and teachers assigned to secondary grades to primary, shutting down the secondary school. Students may have been less likely to progress to secondary education simply because the nearest secondary school was closed. This is consistent with Bridge driving the overall negative effect on school attainment (see Section 3). However, due to small sample sizes, any estimate of the impact of shutting down these grades is very imprecise.[9]

Despite the adverse impact on dropout, the programme had a net positive effect on enrolment in treatment schools after three years equivalent to 36 students per school (p-value = 0.052). This treatment effect comes from enrolment shrinking less in treatment schools because overall enrolment fell across the board during this period. This may be due, in part, to a reduction in fees. The likelihood that principals reported charging fees in primary decreased in programme schools by 21 percentage points (p-value = 0.0025) after three years (see the Online Appendix, Table 16, Panel A). Student attendance (measured during a spot check by our enumerators) was higher in treatment schools by 11 percentage points (p-value = 0.017, see Table 1, Panel A).

While the effect on access to education (i.e., enrolment, as opposed to learning) is negative for the children in our sample, the net effect on access to education overall depends on the (unobserved) share of new students in partnership schools who were previously unenrolled versus enrolled in other schools.

2.4. *Child Safety*

The starkest example of a non-contracted outcome is the protection of child safety, which was not measured in the first year evaluation nor covered in Romero *et al.* (2020). Both government and operators initially resisted the inclusion of this module because it was outside the programme's scope, despite the revelation of child abuse scandals involving more than one operator.[10]

We measure two aspects of child safety: corporal punishment and sexual abuse (see Table 4). Corporal punishment is widespread: 51% of students in control schools report being hit by their teachers at least occasionally. The programme reduced this rate by 4.6 percentage points (p-value = 0.043).

Sexual abuse rates in our data are lower than those reported in previous studies (Postmus *et al.*, 2015; Steiner *et al.*, 2021): 3.6% of students in control schools report having had sex with

[8] The survey options were: (1) left school to work, (2) pregnancy, (3) could not afford school fees and (4) others.

[9] The negative treatment effect on secondary school transition could be partially driven by providers holding back students to ensure they are ready for secondary school by the time they finish sixth grade. While there is some evidence of this behaviour (see Table 15 in the Online Appendix), it cannot explain the overall negative treatment effect on enrolment.

[10] These scandals refer to abuse that occurred before the launch of the programme but not fully revealed to the public until after the programme had begun.

Table 4. *Gender Based Violence Since Enrolling in the Current School (Measured in 2019).*

	All		Girls		Boys	
	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)	Control mean (5)	Treatment effect (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: Column 1, 3 and 5 present the percentage of children who report abuse by different parties (teachers, family members, and other students), as well as the standard deviation (in parentheses) for the control group. Columns 2, 4, and 6 present the treatment effect on these variables and its standard error (in parentheses) taking into account the randomisation design (i.e., including 'pair' fixed effects). Columns 1–2 include all students, columns 3–4 restrict the sample only to girls, and columns 5–6 restrict the sample only to boys. Standard errors are clustered at the school level. Statistical significance at the 1, 5 and 10% levels is indicated by ***, ** and *, respectively.

a teacher (statutory rape) since attending their current school.[11] The programme had a small (−0.41 percentage points) and statistically insignificant (p-value = 0.46) impact on this measure. However, the programme increased reported forced sexual intercourse by peers by 1.7 percentage points (p-value = 0.0042) from a base of 3.6%. Is possible that the likelihood of *reporting* an incident may have increased in programme schools. Indeed, reported cases of forced sexual intercourse at home—where the true rate is unlikely to be affected by the programme—increased by 1 percentage points (p-value = 0.071) from a base of 2.9%.

In summary, the programme reduced but came far from eradicating corporal punishment in schools. Despite multiple credible reports of sexual abuse in schools run by providers involved in this programme, we fail to find any significant change in self-reported sexual abuse resulting from the programme. Sexual abuse remains widespread, and private providers fail to use an influx of new resources and external oversight to reduce its incidence.

2.5. *Discussion: What Explains the Main Effects?*

The neglect of non-contracted outcomes documented above, including harmful effects on dropout and null effects on sexual abuse, is consistent with what a basic multitasking model would predict. However, this begs the question of why contracted outcomes, i.e., test-score gains, plateaued after the first year.

Some potential mediating variables, including teacher behaviour, deteriorated over time. Here we push that analysis one step further, to test whether any observed changes in intermediate outcomes can explain treatment effects on primary outcomes. To do so, we estimate a separate

[11] In a companion paper, Johnson *et al.* (2019) compare the survey protocol that we used with a protocol identical to the one used by Postmus *et al.* (2015) and Steiner *et al.* (2021) and find similar rates of sexual abuse across protocols.

treatment effect for each of the 93 matched pairs in our sample. We can do this for learning outcomes and other outcomes (e.g., teacher attendance, school management, and enrolment). As an exploratory analysis, we show the correlation between these treatment estimates in Table 22 in the Online Appendix.

The correlations point toward a decisive role for new—and unsustainable, if applied beyond the programme—staffing arrangements early in the programme as explanatory factors for initial (and sustained) success. Specifically, treatment effects in learning outcomes are significantly associated across matched pairs with treatment effects in two other outcomes: keeping pupils (enrolled at baseline) in school and removing teachers (employed at baseline). The first may be somewhat mechanical: if school delivers any learning, keeping kids in school boosts the test-score treatment effect. The second suggests that replacing underperforming teachers was an important (if unauthorised) tactic used by some operators. Its effect is somewhat confounded, though, given that operators who removed a high share of teachers also lobbied for an even larger number of new, better-qualified teachers.

Turning from learning outcomes to access: what explains operators' success in getting or keeping kids in school? Treatment effects on overall enrolment (not just pupils present at baseline) are uncorrelated with treatment effects on learning. This may suggest that school quality does not generate demand from parents, or that schools achieved learning gains precisely by capping class sizes. Hiring new teachers was the only factor significantly associated with enrolment gains.

Increases in student attendance are correlated, unsurprisingly, with treatment effects on teacher attendance, teacher time on task, and (mechanically) whether the school was in session. Notably, higher rates of sex between teachers and students are negatively associated with pupil attendance. No such correlation exists between pupil attendance and corporal punishment.

3. Provider Level Heterogeneity

Do identical contracts in similar contexts produce the same results? Our experimental design amounts to having eight experiments, one for each provider, with identical contracts after year 1, the same principal (the government of Liberia) and heterogeneous agents (from local non-profits to multinational companies).

Before making comparisons across providers we must address two technical obstacles. First, providers managed schools in different counties with somewhat different baseline conditions. Thus, while each provider's treatment effects are internally valid (see Table 5), they are not comparable without further assumptions. However, Romero *et al.* (2020) finds baseline heterogeneity does little to explain heterogeneity in treatment effects. This is still the case, and Table 24 in the Online Appendix provides results after controlling for baseline school and student characteristics. Second, sample sizes for each provider are small. We use randomisation inference, which provides exact tests of sharp hypotheses regardless of the sample size (Young, 2019).

We find heterogeneity in treatment effects across providers in our main outcomes: learning, sustainability, access and child safety. Furthermore, 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.

Table 5. *Treatment Effects by Provider (Randomisation Inference).*

	BRAC (1)	Bridge (2)	MtM (3)	Omega (4)	Rising (5)	St. Child (6)	Stella M (7)	LYONET (8)
Panel A: Student test scores (ITT)								
English	0.088	0.223	0.294	−0.057	0.317	0.245	−0.201	0.625
	[0.284]	[0.088]	[0.015]	[0.593]	[0.195]	[0.050]	[0.511]	[0.123]
Math	0.056	0.393	0.448	−0.099	0.433	0.286	0.076	0.285
	[0.519]	[0.002]	[0.006]	[0.311]	[0.063]	[0.116]	[0.789]	[0.127]
Composite	0.058	0.348	0.424	−0.083	0.417	0.285	−0.042	0.419
	[0.473]	[0.007]	[0.006]	[0.430]	[0.063]	[0.071]	[0.972]	[0.127]
Panel B: Student test scores (ToT)								
English	0.088	0.223	0.519	−0.063	0.405	0.284	0.000	0.625
	[0.283]	[0.090]	[0.016]	[0.591]	[0.190]	[0.045]	[0.0]	[0.123]
Math	0.056	0.393	0.790	−0.110	0.553	0.331	0.000	0.285
	[0.520]	[0.001]	[0.001]	[0.310]	[0.063]	[0.101]	[0.0]	[0.125]
Composite	0.058	0.348	0.748	−0.092	0.532	0.330	0.000	0.419
	[0.472]	[0.006]	[0.003]	[0.432]	[0.062]	[0.064]	[0.0]	[0.126]
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]	[0.345]	[0.002]	[0.021]	[0.000]	[0.125]
Age in years (teachers)	−4.085	−11.442	−8.668	−5.860	−10.420	−1.912	−8.065	0.798
	[0.000]	[0.000]	[0.000]	[0.000]	[0.073]	[0.364]	[0.001]	[0.751]
Test score (teachers)	0.245	0.246	−0.051	0.259	0.294	0.275	−0.290	0.284
	[0.048]	[0.158]	[0.749]	[0.172]	[0.386]	[0.278]	[0.388]	[0.016]
Panel D: Enrolment and access								
Δ enrolment	57.275	25.045	69.900	26.789	75.000	−5.182	−1.625	92.625
	[0.083]	[0.683]	[0.026]	[0.383]	[0.215]	[0.900]	[0.917]	[0.054]
Student attendance	20.351	4.921	44.265	18.158	29.513	20.853	6.259	13.573
	[0.001]	[0.421]	[0.001]	[0.229]	[0.000]	[0.030]	[0.490]	[0.000]
% still attending any school	−1.101	−6.525	−3.808	−4.844	0.934	−0.439	3.175	−0.074
	[0.646]	[0.088]	[0.342]	[0.163]	[0.890]	[0.934]	[0.662]	[0.918]
% still attending same school	3.496	−4.488	4.002	5.028	11.779	7.341	8.969	−0.393
	[0.481]	[0.367]	[0.741]	[0.471]	[0.517]	[0.218]	[0.377]	[0.877]
Panel E: Child Safety								
Teacher never hits students (%)	6.35	−2.11	19.39	−0.66	9.57	14.55	9.59	−5.23
	[0.157]	[0.677]	[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 randomisation 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. Panel A presents data on intention-to-treat estimates on students' test scores (measured in standard deviations relative to the control group distribution). Panel B presents data on treatment-on-the-treated estimates, taking into account school-level non-compliance, on students' test scores (measured in standard deviations relative to the control group distribution). Panel C presents data related to the pool of teachers in each school. Panel D presents data related to school enrolment. Panel E presents data related to child safety, measured as the percentage of children who report corporal punishment or sexual abuse from teachers. The *p*-values from randomisation inference are presented in brackets. The randomisation inference uses 5,000 iterations to calculate *p*-values based on the distribution of squared *t*-statistics following Young (2019). Number of schools refers to the number of schools in both treatment and control groups.

3.1. *Learning*

While the programme as a whole raised composite learning outcomes by 0.2σ, three of the eight providers produced negligible and statistically insignificant learning gains, while the other five generated similar ITT effects of $\approx 0.4\sigma$ (Table 5). Of this last group, four of five (Rising Academies, More than Me, Bridge, and Street Child) have statistically significant learning effects. The ToT effects (Table 5, Panel B) shows a higher variance across providers, with effects as high as 0.75σ for More Than Me and as low as -0.09σ for Omega Academies.

3.2. *Sustainability*

The programme's main cost, beyond the fixed price agreed per pupil, was extra staffing (paid in part by the government) and other subsidies (covered by providers or their independent fundraising).

Private operators demanded freedom to dismiss teachers and priority in the allocation of extra staff from the government, which, as noted above, was correlated with learning gains but potentially unsustainable. Large-scale dismissal of teachers was driven by two operators, Bridge and More than Me (Table 5, Panel C). Most providers received several new teachers, except Omega Academies, Stella Maris, and the Liberian Youth Network. While weeding out bad teachers is important, reallocating weak teachers to other schools and giving preference to select schools in the recruitment of new teachers is unlikely to raise average performance in the system as a whole. Note that we are unable to verify whether the teachers dismissed from programme schools were reassigned to other public schools.

Consistent with the idea that high costs in the first year were driven by start-up investments and fixed costs that would decline as the programme grew, unit costs have fallen since the first year, but remain above the original projections. The ministry expects providers to operate for US$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 US$300 per pupil, with some providers spending the target amount (US$50 per pupil) and one other spending over US$600 per pupil. After three years, the average (self-reported) expenditure has fallen to US$119 per pupil. However, Bridge International Academies and More Than Me continued to spend at least three times as much as the government target (see Figure 1).

3.3. *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 $= 0.342$). Dropout due to pregnancy is a major factor in Bridge (see Table 23 in the Online Appendix). As mentioned in Subsection 2.3, another explanation may be that Bridge effectively shut down the nearest secondary school for some students.

The treatment effect on school size—i.e., attracting new students, not in our baseline sample—is positive for six of eight providers, and significantly so for four (Table 5, Panel D). Furthermore, five of the eight providers have a positive and statistically significant treatment effect on student attendance.

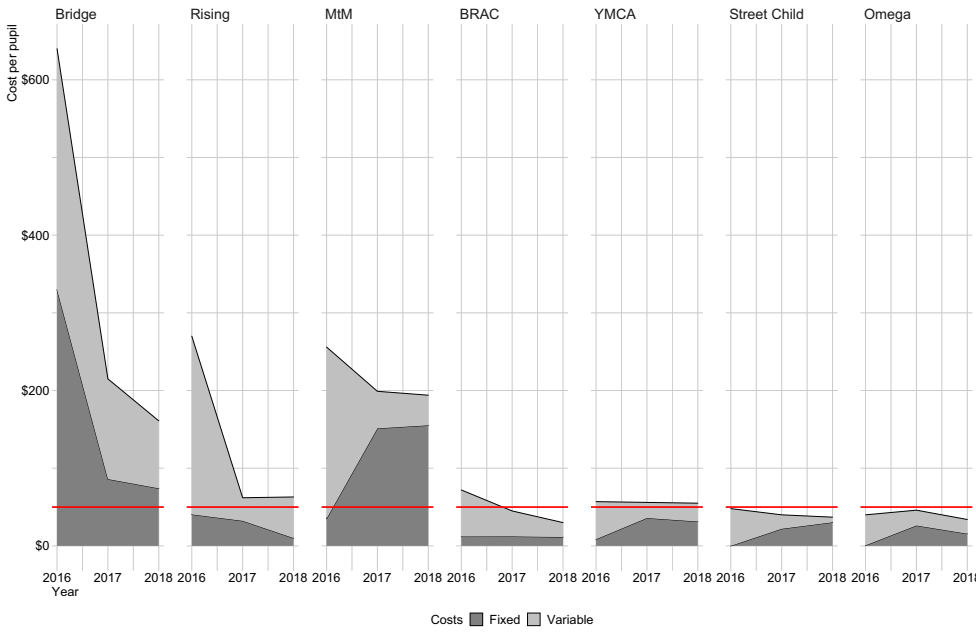

Fig. 1. *Per Pupil Cost. Numbers are based on budgets submitted to Social Finance, who managed the pool of funds that paid providers the per-pupil subsidy. Stella Maris did not provide budget data. Numbers do not include the cost of teaching staff borne by the Ministry of Education. The horizontal line represents US$50 per pupil. While we report fixed and variable costs, it appears providers fail to distinguish between these categories. For instance, Street Child and the Liberian Youth Network claim nearly zero variable cost, implying the potential for costless expansion of the programme, which is unrealistic.*

3.4. *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), positive and insignificant for four other providers, and negative and insignificant for the remaining three.

In terms of students reporting sexual abuse by teachers, we observe a negative treatment effect (i.e., lower rates) for BRAC (p-value = 0.053), negative and insignificant point estimates for four other providers, and positive and insignificant estimates for three others.

4. Conclusions

Liberia's initiative to outsource the management of 93 randomly selected government primary schools to eight private providers led to short-term learning gains, partially offset by high costs and early signs of harmful side effects on the broader school system (Romero *et al.*, 2020). In this paper, we summarise impacts (1) over a longer time horizon, (2) on a range of outcomes beyond test scores and (3) distinguishing the average impact of the outsourcing policy. Treatment effects plateau for the primary contracted outcome, learning gains, after the first year. Among outcomes

that were not contracted or explicitly de-emphasised by the government, the programme had null or harmful effects. Treatment reduced enrolment and increased dropout for the sample of students originally enrolled in treatment schools. While the programme reduced corporal punishment in schools, sexual abuse did not decline and remains widespread in treatment schools, despite an influx of new resources and external oversight.

Beyond these average effects, we document substantial heterogeneity in impacts across the eight private providers which managed schools as part of the overall programme. Heterogeneity exists not only in learning but also in access to education, child safety and the sustainability of providers' models. Complicating the policy analysis, impacts on these various dimensions are not perfectly correlated. While some providers show almost uniformly positive effects (even on non-contracted outcomes), others present trade-offs for policymakers who must decide on the weights to attach to positive gains in one dimension and losses in another. Notably, firm observable characteristics do not predict outcomes. Nevertheless, private contractors' identity matters and selecting providers aligned with the public interest (Akerlof and Kranton, 2005; Besley and Ghatak, 2005) may be key for public–private partnerships in education.

Some of our results are readily explained by a multitasking framework, combined with the overall weakening of performance accountability in the programme over time. At first, a focus on learning gains led to a positive treatment effect on test scores (the contracted output) and the neglect of other outcomes. The government gave operators explicit, public signals regarding what behaviour would and would not be sanctioned, e.g., ignoring early signs of harmful effects on access. Similarly, revelations of prior sexual abuse scandals involving operators led to no serious effort to improve child safety. As the political and media pressure to deliver waned over the three years, learning gains plateaued and other outcomes stagnated or continued to decay.

Sanctioning underperforming providers is an obvious policy recommendation of our results. In practice, even providers who did nothing in the first year (e.g., Stella Maris) expelled large numbers of students (Bridge International Academies), or who were implicated in serious sexual abuse cases (More than Me) were rewarded with more schools when the programme expanded. Foreign donors played an accessory role here. Many providers invested heavily in the Liberian programme in the first year, expecting that good performance would generate long-term funding (in Liberia and elsewhere) and good publicity internationally. Yet, even as evidence emerged of harmful unintended consequences, foreign philanthropy continued to fund the organisations responsible. The arbitrary allocation of external funding allowed less efficient operators to persist. The competitive market dynamics that the public–private partnership was trying to recreate (e.g., market discipline) did not materialise.

ITAM, Mexico
Center for Global Development, USA

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

References

Abadzi, H. (2011). 'Reading fluency measurements in EFA FTI partner countries: Outcomes and improvement prospects', GPE Working paper, World Bank.

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

Aslam, M., Rawal, S. and Saeed, S. (2017). *Public–Private Partnerships in Education in Developing Countries: A Rigorous Review of the Evidence*, Report, London: Ark Education Partnerships Group.

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

Barrera-Osorio, F. (2007). 'The impact of private provision of public education: Empirical evidence from Bogota's concession schools', Working paper, World Bank.

Barrera-Osorio, F., Blakeslee, D.S., Hoover, M., Linden, L., Raju, D. and Ryan, S.P. (Forthcoming). 'Delivering education to the underserved through a public–private partnership program in Pakistan', *The Review of Economics and Statistics*, published online ahead of print.

Baysah, A.M., Jr. (2016). 'Liberia: Police charge youth activist for sodomy', *The New Republic*, 2 November. https://web.archive.org/web/20161103182507/https://allafrica.com/stories/201611020824.html (last accessed: 24 November 2021).

Bergman, P. and McFarlin, I., Jr. (2018). 'Education for all? A nationwide audit study of school choice', Working paper, National Bureau of Economic Research.

Besley, T. and Ghatak, M. (2005). 'Competition and incentives with motivated agents', *The American Economic Review*, vol. 95(3), pp. 616–36.

Betts, J.R. and Tang, Y.E. (2014). 'A meta-analysis of the literature on the effect of charter schools on student achievement', Working paper, Society for Research on Educational Effectiveness.

Bonilla, J.D. (2010). 'Contracting out public schools for academic achievement: Evidence from Colombia', Ms., University of Sao Paulo.

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

Corts, K.S. and Singh, J. (2004). 'The effect of repeated interaction on contract choice: Evidence from offshore drilling', *Journal of Law, Economics, and Organization*, vol. 20(1), pp. 230–60.

DIVA-GIS. (2016). 'Liberia administrative areas', http://biogeo.ucdavis.edu/data/diva/adm/LBR_adm.zip (last accessed: 15 August 2021).

Eble, A., Frost, C., Camara, A., Bouy, B., Bah, M., Sivaraman, M., Hsieh, P.T.J., Jayanty, C., Brady, T., Gawron, P., Vansteelandt, S., Boone, P. and Elbourne, D. (2021). 'How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia', *Journal of Development Economics*, vol. 148, article 102539.

Eyles, A. and Machin, S. (2019). 'The introduction of academy schools to England's education', *Journal of the European Economic Association*, vol. 17(4), pp. 1107–46.

Fazzio, I., Eble, A., Lumsdaine, R.L., Boone, P., Bouy, B., Hsieh, P.-T.J., Jayanty, C., Johnson, S. and Silva, A.F. (2021). 'Large learning gains in pockets of extreme poverty: Experimental evidence from Guinea Bissau', *Journal of Public Economics*, vol. 199, article 104385.

Hart, O., Shleifer, A. and Vishny, R.W. (1997). 'The proper scope of government: Theory and an application to prisons', *The Quarterly Journal of Economics*, vol. 112(4), pp. 1127–61.

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

Hsieh, C.-T. and Urquiola, M. (2006). 'The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program', *Journal of Public Economics*, vol. 90(8), pp. 1477–503.

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

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

Lee, D.S. (2009). 'Training, wages, and sample selection: Estimating sharp bounds on treatment effects', *The Review of Economic Studies*, vol. 76(3), pp. 1071–102.

Lemos, R. and Scur, D. (2016). 'Developing management: An expanded evaluation tool for developing countries', Working paper, RISE.

Liberia Institute of Statistics and Geo-Information Services. (2016). Liberia—Household income and expenditure survey 2014–2015, Survey, World Bank.

Martinez, E. and Odhiambo, A. (2018). *Leave No Girl Behind in Africa: Discrimination in Education Against Pregnant Girls and Adolescent Mothers*, Technical report, New York: Human Rights Watch. https://www.hrw.org/sites/default/files/report_pdf/au0618_insert_webspreads.pdf (last accessed: 24 November 2021).

Ministry of Education—Republic of Liberia. (2016). 'Request for expression of interest for provision of primary partnership school services'. http://www.emansion.gov.lr/doc/MOE_1.pdf (last accessed: 6 August 2017).

Mukpo, A. (2017). 'In Liberia, a town struggles to adjust to its new charter school', World Education Blog. https://gemreportunesco.wordpress.com/2017/04/12/in-liberia-a-town-struggles-to-adjust-to-its-new-charter-school (last accessed: 28 July 2017).
Patrinos, H.A., Barrera-Osorio, F. and Guáqueta, J. (2009). *The Role and Impact of Public–Private Partnerships in Education*, Washington, DC: World Bank Publications.
Pilling, D. (2017). 'Liberia is outsourcing education. Can it work?', *Financial Times*, 21 April. https://www.ft.com/content/291b7fca-2487-11e7-a34a-538b4cb30025 (last accessed: 13 September 2017).
Postmus, J.L., Hoge, G.L., Davis, R., Johnson, L., Koechlein, E. and Winter, S. (2015). 'Examining gender based violence and abuse among Liberian school students in four counties: An exploratory study', *Child Abuse & Neglect*, vol. 44, pp. 76–86.
Romero, M., Sandefur, J. and Sandholtz, W. (2018). 'Partnership schools for Liberia', Harvard Dataverse. DOI: 10.7910/DVN/5OPIYU.
Romero, M., Sandefur, J. and Sandholtz, W.A. (2020). 'Outsourcing education: Experimental evidence from Liberia', *American Economic Review*, vol. 110(2), pp. 364–400.
Romero, M., Sandefur, J. and Rodriguez-Ramirez, R. (2021). Replication package for: 'Beyond short-term learning gains: The impact of outsourcing schools in liberia after three years'. https://doi.org/10.5281/zenodo.5579799.
Rosenberg, T. (2016). 'Liberia, desperate to educate, turns to charter schools', *The New York Times*, 14 June. http://www.nytimes.com/2016/06/14/opinion/liberia-desperate-to-educate-turns-to-charter-schools.html (last accessed: 20 July 2016).
Sandholtz, W.A. (2020). 'Do voters reward service delivery? Experimental evidence from Liberia', Ms., Mimeo.
Senah, G. (2016). 'At Kendeja public school, more than 300 students left unenrolled', The Bush Chicken, 9 September. http://www.bushchicken.com/at-kendeja-public-school-more-than-300-students-left-unenrolled (last accessed: 28 July 2017).
Singh, A. (2015). 'How standard is a standard deviation? A cautionary note on using SDs to compare across impact evaluations in education', World Bank Blogs, 13 January. http://blogs.worldbank.org/impactevaluations/how-standard-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations (last accessed: 31 July 2017).
Starr, K. (2017). 'Let the man do his job!', Mulago, 26 July. https://www.mulagofoundation.org/articles/let-the-man-do-his-job (last accessed: 24 May 2021).
Starr, K. (2020). 'A messy start, solid research, and huge potential', Mulago, 15 January. https://www.mulagofoundation.org/blogs/a-messy-start-solid-research-and-huge-potential (last accessed: 24 May 2021).
Steiner, J.J., Johnson, L., Postmus, J.L. and Davis, R. (2021). 'Sexual violence of Liberian school age students: An investigation of perpetration, gender, and forms of abuse', *Journal of Child Sexual Abuse*, 30, pp. 21–40.
The Economist. (2017). 'Liberia's bold experiment in school reform', *The Economist*, 25 February. https://www.economist.com/news/middle-east-and-africa/21717379-war-scorched-state-where-almost-nothing-works-tries-charter-schools-liberias (last accessed: 13 September 2017).
Tyre, P. (2017). 'Can a tech start-up successfully educate children in the developing world?', *The New York Times*, 27 June. https://www.nytimes.com/2017/06/27/magazine/can-a-tech-start-up-successfully-educate-children-in-the-developing-world.html (last accessed: 27 June 2017).
Werner, G.K. (2017a). 'Liberia has to work with international private school companies if we want to protect our children's future', *Quartz Africa*, 3 January. https://qz.com/876708/why-liberia-is-working-with-bridge-international-brac-and-rising-academies-by-education-minister-george-werner (last accessed: 20 July 2017).
Werner, G.K. (2017b). 'Opinion: Liberia's battle to educate our next generation', *Devex*, 6 October. https://www.devex.com/news/opinion-liberia-s-battle-to-educate-our-next-generation-91202 (last accessed: 24 May 2021).
Werner, G.K. and Malkus, N. (2017). 'Education reform in Africa—with George Werner, Liberian Minister for Education', American Enterprise Institute, 29 August on YouTube. https://www.youtube.com/watch?v=6sEpkdkvRZg (last accessed: 30 November 2017).
Young, A. (2019). 'Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results', *The Quarterly Journal of Economics*, vol. 134(2), pp. 557–98.
Young, F. (2018). 'Unprotected', *ProPublica*, 11 October. https://features.propublica.org/liberia/unprotected-more-than-me-katie-meyler-liberia-sexual-exploitation (last accessed: 15 December 2018).