

CAN MANAGEMENT IMPROVE LEARNING? EXPERIMENTAL EVIDENCE FROM EDUCATION*

Felipe Galvão Puccioni, Tiago Cavalcanti

This study provides experimental evidence of the power of management to boost the learning production of public schools. We developed a new management programme to improve school management at the cost of USD PPP GDP 15.22 per pupil per year. No financial incentive. No personnel or system changes. No teaching time increase. The programme’s ATE estimates were 0.916 standard deviations (SD) for school management, 0.226 (0.059) SD for reading and 0.237 (0.059) SD for mathematics (cost-effectiveness of 1.95 ‘Additional SD per \$100 (PPP)’). The treatment intensity analysis (IV estimates) showed that the impact of increasing the school management by one score point (on a scale from one, worst management, to five, best management) was 0.680 (0.245) SD for reading and 0.714 (0.265) SD for mathematics (cost-effectiveness of 15.61 ‘Additional SD per \$100 (PPP)’). Schools that improved their management level by one score point due to the treatment delivered in two years to their pupils the learning equivalent to what is learnt in more than four years in a control school, i.e., high implementation schools were around 110% more productive in providing learning to their pupils. Furthermore, the IV analysis showed that an increase of one SD in management has a causal effect of 0.24 SD in pupils’ learning. *JEL Codes:* C93, H83, I20, J24 and M10.

I. INTRODUCTION

Educated people are generally healthier, have fewer comorbidities, live longer and earn more than people with less education (Card 1999; Chetty, Friedman, and Rockoff 2014; Davies et al. 2018; Krueger 2003). Evidence

*We are thankful to Luiz Antonio Guaraná and Renan Ferreirinha for their help and support. This research would not have been achievable without our collaboration with the Court of Accounts of Rio de Janeiro (TCMRio), Rio de Janeiro’s City Hall, and the Municipal Secretariat of Education of Rio de Janeiro (SMERJ). We also thank the programme’s implementation team and all school managers who participated in this project. All errors are the sole responsibility of the authors.

suggests that education affects occupational prestige, success in the labour market, marriage happiness, children's success, decision-making and many other non-pecuniary factors (Oreopoulos and Salvanes 2011). Not without reason, education has been a fundamental human right since the Universal Declaration of Human Rights (1948).

Even though many factors may be partially responsible for driving pupils' educational performance, such as family background, culture, country wealth, investment in education, etc., how the resources available to education are used may be crucial (Abdulkadiroğlu et al. 2011; Angrist et al. 2010, 2012; Barros et al. 2019, 2021; Beg, Fitzpatrick, and Lucas 2023; Bloom et al. 2020, 2012a; Curto and Fryer 2014; Dobbie and Fryer 2011; Fryer 2017; Fryer 2014; Gosnell, List, and Metcalfe 2020; Tavares 2015). We focus on understanding how a public school can be more efficient in providing pupils' learning. How can public schools provide the most learning for pupils from the resources available? Can specific management practices improve the learning productivity of public schools?

Through a large-scale randomised field experiment, this study shows that a novel Agile management programme can significantly improve pupils' learning. We developed this intervention to deliver the 'best' management practices to school managers through one-to-one coaching sessions and on-the-job training. These specific managerial practices have been discussed extensively in the literature and are detailed in (Bloom et al. 2015). Yet, instead of contracting a consulting company to provide the treatment, the programme's implementation team was formed by civil servants who worked autonomously through iterations (sprints) to continuously deliver small programme pack-

ages following the Agile project management approach. We call the intervention the Science and Management for Education Programme (SMEP).

The experiment was conducted with 31,760 pupils from 80 schools randomly sampled from the city of Rio de Janeiro's population of grades 1-9 public schools, one of the largest public education networks in Latin America.¹ Muralidharan and Niehaus (2017) found that in the top-five economic journals published between 2001 and 2016, only 31% of the randomised control trials (RCTs) showed that units were drawn from a larger population. From the random experimental sample, 40 schools were randomly assigned to treatment and 40 to control through a pair-matching procedure.

The average treatment effect (ATE) estimates of the impact of the Science and Management for Education Programme on pupils' educational outcomes were 0.226 (0.059) standard deviations (SD) for reading and 0.237 (0.059) SD for mathematics. Since pupils' learning (our outcome of interest) comes from assessments that measure all pupils from first to ninth grade against the same scale of competencies and scores, it is possible to translate the ATE estimates to years of learning. Pupils in treatment schools are $\frac{3}{4}$ of a school year of learning (reading and mathematics) ahead of pupils from schools that did not receive the treatment. Treatment schools were 37.5% more productive in providing reading and mathematics learning to their pupils than non-treated schools. Moreover, the treatment had a statistically significant impact of 19% on pupil absenteeism reduction.

We conducted two double-blind management surveys before and after the experiment. The surveys measured the management of the schools regarding

¹Information from <https://dl.acm.org/doi/abs/10.1145/3614321.3614379>

23 specific management practices that the programme aimed to improve at the treatment schools. The surveys adopted the World Management Survey (WMS) methodology (Bloom and Van Reenen 2007, 2010; Bloom et al. 2012b; Bloom et al. 2015). The surveys were classified as double-blind because school managers were unaware their schools were being scored, and interviewers were unaware of the schools' educational performance. The estimate of the programme's causal effect (ATE) on school general management was 0.916 (0.268) standard deviations (SD). This estimate is statistically significant at the 1% level.

Since the surveys assessed each school regarding each of the 23 practices, it was also possible to measure the intervention's impact on each practice. The treatment had a positive impact (ATE) on most of the 23 management practices. This fact confirms that the intervention successfully achieved its goal of improving treatment school management regarding the 'best' management practices. However, the impact was heterogeneous across the practices.

The heterogeneity of the programme's impact across the practices is best understood by grouping the practices into five sets: target setting, leadership, operation, monitoring and people management. The impact of treatment (ATE) was 1.530 standard deviations (SD) for the target-setting group of practices that aggregates Target balance (t10), Target time horizon (t12), Target interconnection (t11), Target clarity/comparability (t14) and Target stretch (t13), and 1.069 SD for the leadership group of management practices that is formed by Leadership vision (l21), Leadership accountability (l22) and Clearly defined roles (l23). Both estimates are statistically significant at the 1% level.

The impact of the programme was 0.370 SD for the operation group of management practices, which is formed by Adopting best practices (o4), Data-driven planning (o3), Instruction personalisation (o2), and Planning standardisation (o1). Regarding the monitoring group of management practices, which aggregates Performance tracking (m6), Performance review (m7), Continuous improvement (m5), Performance dialogue (m8) and Consequence management (m9), the causal effect of treatment was 0.324 SD. Both estimates were not statistically significant.

Three of the six people management practices - Managing talent (p18), Attracting employees (p20) and Retaining talent (p19) management practices - were positively impacted by the treatment. However, the estimates were not statistically significant. The other three practices from the people management group, Fixing poor performers (p16), Promoting high performers (p17) and Rewarding high performers (p15), did not change due to the treatment. This fact was expected since Brazilian legislation for civil servants, such as public school teachers, makes it very difficult to remove or dismiss poor performers and promote or reward high performers.

The treatment intensity analysis that used the random assignment as the instrumental variable (IV) reveals that one score point improvement in school management — on a scale that goes from one, worst management, to five, best management — due to the treatment had a striking impact on pupils' educational outcomes: 0.680 (0.245) SD for reading and 0.714 (0.265) SD for mathematics. If the causal endogenous variable, school management, is standardised to have a mean of zero and a standard deviation of one regarding the control group, a one standard deviation change in school management

had a causal effect of 0.247 (0.089) SD for reading and 259 (0.096) SD for mathematics. It is the first study to present the causal effects of management, measured following the WMS methodology, on pupils' educational outcomes. Surprisingly, these findings are very similar to the correlation estimates found by Bloom et al. (2015).

The IV estimates can be translated into years of learning as done with the ATE estimates. Pupils in high implementation schools - schools that improved their management level by one score point due to the treatment - are more than two academic years of learning ahead of pupils from schools that did not receive the treatment. Two years in these high implementation schools taught pupils the equivalent of what is learnt in more than four years in a Rio de Janeiro school. These schools were around 110% more productive in providing learning to their pupils than non-treated schools.

The ATE and IV estimates were statistically significant at the 1% level across subjects, a rare fact in the related academic literature. Importantly, the differences in the programme's impact are very small and not statistically significant within the different subgroups: race, gender, poverty level, school segments (I: grades 1-5 schools vs II: grades 6-9 schools), school size (pupils) or pupils' baseline performance groups. Furthermore, despite the treatment's large effects on management and learning, the intervention had a per year cost per pupil of USD (PPP GDP) 15.22.

Two years of programme implementation and two deep management surveys conducted before and after the programme in 2021 and 2023 have clarified our understanding of public schools in Brazil. Bureaucracy and daily 'emergencies' drag school managers' routines down. Moreover, Brazil gener-

ates many incentives through civil society, media, and watchdog organisations such as the courts of accounts for school managers to focus on building facilities, repairs, pupils' food, security, etc., which is surely important. However, little incentive is given to school managers to focus on managing to improve pupils' learning.

In addition to the lack of incentives to focus on the management that can improve pupils' learning, school managers lack information on the best management practices to improve pupils' education and the skills needed to implement these practices. Lack of information - managers may not know that they are not performing well and not know what to do to perform better — and lack of motivation — managers not being incentivised to enhance or accountable for enhancement - are identified as possible causes for the persistence of poor management practices in organisations by the management theory Gibbons and Henderson (2012).

Based on our diagnoses of the management of the public schools, we can explain the success of the Science and Management for Education Programme. Firstly, the programme filled the informational gap on how to manage a public school better to improve pupils' learning, providing school managers with very detailed, adapted-to-their-reality information on the best management practices. Secondly, the programme not only helped school managers implement these practices in their schools through simple tools but also developed the skills needed to do so. We tried at maximum to make things easy for school managers following a large body of evidence on the impact of "make it easy" on programme participation and engagement (Thaler 2021). Thirdly, the programme succeeded in bringing school man-

agers' attention to ways, through better management, to improve pupils' learning. Lastly, all these three possible causes for the programme's success rely on the impact of the managerial practices provided to schools on pupils' learning.

This study contributes to several strands of literature. The treatment effects discussed in this work on pupils' learning are amongst the largest in the education intervention literature according to Kraft (2020) that analysed the distribution of 1,942 effect sizes from 747 experiments evaluating education interventions with standardised test outcomes. Moreover, our cost-benefit analysis has shown that the programme has the highest internal rates of return (IRR), 115% (ATE) and 163% (High Implementation), compared with 17 relevant US interventions in education (Fryer 2017; Krueger 2003). The present value of the expected future Brazilian earnings inflows that our intervention (ATE) generates for a typical pupil is around USD (PPP GDP) 133,086.24 or R\$ 343,362.19 where R\$ means Brazilian Reais. The present value achieves USD (PPP GDP) 400,625.58 or R\$ 1,033,794.60 for a typical pupil from a school that improved its management level by one score point due to the treatment.

The cost-effectiveness analysis that followed the J-PAL approach discussed in Dhaliwal et al. (2013) shows that compared with 27 other relevant interventions to improve pupils' learning, the Science and Management for Education Programme (SMEP) has one of the highest measures of cost-effectiveness. The programme (ATE) cost-effectiveness is 1.95 'Additional SD per \$100 (PPP)'. Furthermore, for schools that changed their management by one score point due to the treatment (High Implementation), the inter-

vention cost-effectiveness achieves a striking 29.44 ‘Additional SD per \$100 (PPP)’, showing the power of management to improve school productivity (pupils’ learning).

A key factor in understanding the pronounced results achieved by the programme in the cost-benefit and cost-effectiveness analyses is its low per pupil per year cost of USD (PPP GDP) 15.22. The programme’s low cost can be attributed to the following facts: the intervention was implemented without changing any existing systems or personnel and without providing any financial incentives, and the programme’s implementation only involved working with school managers, without any additional work conducted with teachers or pupils, which helped to keep the programme’s implementation team at a reduced number of professionals.

This work also contributes to the literature by solving the conflicting evidence on the impact of management on productivity, mainly on education. On one hand, Abdulkadiroğlu et al. (2011), Angrist et al. (2010, 2012), Barros et al. (2019, 2021), Beg, Fitzpatrick, and Lucas (2023), Bloom et al. (2020, 2012a), Bruhn, Karlan, and Schoar (2018), Curto and Fryer (2014), Dobbie and Fryer (2011), Fryer (2017), Fryer (2014), Gosnell, List, and Metcalfe (2020), and Tavares (2015) have discussed the positive causal effects of management on productivity. On the other hand, Hoyos, Ganimian, and Holland (2019), Muralidharan and Singh (2020), and Romero et al. (2022) have reported no effect of management on pupils’ learning.

Based on the evidence brought by this study, three factors can explain the null effects literature. Interventions that deliver only information or training do not necessarily change organisational management practices. Also, small

changes in the management level of organisational units are not enough to affect productivity. Our programme had to change the management of the schools by 0.916 SD to change pupils' learning by around 0.23 SD. The treatment intensity analysis has shown that a standard deviation (SD) change in the management of the schools due to treatment has a 0.24 SD effect on pupils' educational outcomes. Lastly, not all managerial practices impact productivity. Moreover, even a practice that can affect productivity may fail if implemented at a low level. The Science and Management for Education Programme not only successfully provided schools with the right managerial practices to affect pupils' learning but also implemented these practices at a level capable of impacting pupils' educational performance.

School management is likely to matter not only for Rio de Janeiro schools but for all Brazilian public schools. Despite culture, wealth, public investment in education, and other differences across Brazilian cities, public schools have similar organisational structures to be managed. For example, each school has a principal, supervisor (s) or pedagogical coordinator (s) and teachers. The organisation of schools are standardised even worldwide according to Dobbie and Fryer (2013), Fryer (2017), and Fryer (2014). Therefore, this work can have far-reaching implications for Brazil's educational policy.

Brazil has serious problems in equipping its pupils with quality education. We show that following its efficiency in education, even increasing by 100% Brazil's cumulative per-pupil public expenditure on education (from 6 to 15 years old) - from USD PPP 37,954.00 (2018) to USD PPP 75,908.00 - the country would still have an educational performance based on PISA results

below Chile that had a cumulative per pupil expenditure on education of USD PPP 50,149.00 in 2018.² That is, even if Brazil spent 50% more than Chile, doubling Brazil’s public expenditure on education, the country would not achieve Chile’s educational performance.

However, our programme (ATE estimates) could make Brazil as efficient in education as Chile or European countries such as Italy, Belgium or Norway almost without changing the public per pupil expenditure in education through our low-cost intervention of USD (PPP GDP) 15.22 per pupil per year. Furthermore, if Brazil implemented the programme at a high level so that the management of Brazilian schools improved by one score point, Brazil would advance 68.0 (0.68 SD) score points in reading and 71.4 in mathematics in PISA. The best managerial practices delivered by the programme could close the perverse and persistent educational gap between Brazil and European countries.

Our work has substantial implications for global policy since it contributes to the priority global issue of low-quality education, providing an intervention that can leverage pupils’ learning at a low cost and scale. Most interventions are ineffective in improving learning, or if effective, they are expensive (Angrist et al. 2020; Fryer 2017; J-PAL 2020; Kraft 2020). For example, Curto and Fryer (2014) show that attending a SEED (School for Educational Evolution and Development) school, which combines a ‘No Excuses’ charter model with a 5-day-a-week boarding programme, increases achievement by 0.211 SD in reading and 0.229 SD in maths per year; however, the per year per pupil cost is USD 51,904.45. Our treatment delivers higher impacts but at a per

²PISA 2018 Results (Volume I) - © OECD 2019

year per pupil cost of USD (PPP GDP) 15.22. Also, a review of 150 education impact evaluations in low- and middle-income countries showed that half of the interventions had no effect on learning (Angrist et al. 2021). This study highlights the power of management to improve education. Better-managed schools can boost learning for pupils worldwide at a very low cost.

II. BACKGROUND AND PROGRAMME DETAILS

II.A. Institutional Background - Rio de Janeiro Education

The municipality of Rio de Janeiro has a population of around 6.2 million people and is the capital city of the Federated State of Rio de Janeiro in Brazil.³ The municipality has one of Latin America’s largest educational systems, responsible for 1,549 educational units and 615,338 pupils in 2023.⁴ This large system of state-run schools and pupils is managed by the Municipal Secretariat of Rio de Janeiro (SMERJ). Since Brazil has only public or private schools, the first being funded and managed by the government, we use public or state-run schools interchangeably in this study.

Most Rio de Janeiro public schools offer some combination of grades 1-9 (991 schools in 2023) and are responsible for 441,842 pupils. Other units are nurseries and schools offering special and adult education. After grade 9, pupils are no longer in the municipality of Rio de Janeiro’s education system. The Federated State of Rio de Janeiro provides high school education (upper secondary).

Principals are elected for two-year terms in the public schools under the

³<https://www.ibge.gov.br/cidades-e-estados/rj/rio-de-janeiro.html>

⁴<https://educacao.prefeitura.rio/educacao-em-numeros/>

responsibility of the city of Rio de Janeiro.⁵ Only teachers who are civil servants from the city with at least five years of teaching experience are eligible. The school staff, pupils and their parents can vote. The aim is to involve the school community in the decision-making process. It is common for principals to be re-elected several times for the same school. Approximately 85% of the principals who served in the 2020-2021 term were re-elected for the 2022-2023 term. Principals are elected with a deputy principal and can choose their staff, including a pedagogical coordinator.

Brazil's Basic Education Development Index (Ideb), shown in Table A1 from Appendix A, was created in 2007 to measure the quality of education in the country. It considers two factors: the rate of pupil approval and the average performance in reading and mathematics assessments (Saeb). Ideb is calculated for 5th and 9th-grade pupils separately. Rio de Janeiro Federated State, also known as RJ, comprises 92 municipalities, including its capital, Rio de Janeiro, while Brazil has 5,568 municipalities.

Rio de Janeiro's city performance in Ideb is compared with 92 municipalities from Rio de Janeiro Federated State (RJ) and all Brazil's municipalities in Table A1 from Appendix A. This table shows that Rio de Janeiro municipality's public expenditure per pupil increased by 102% while the Brazilian municipalities' average increased by 63% from 2007 to 2019. According to the 2020 Court of Accounts of Rio de Janeiro (TCMRio) Special Report⁶, the average public school teacher's monthly wage in the municipality of Rio de Janeiro was, on average, R\$ 3,741.00 (values updated to 2019) in 2007

⁵Rio de Janeiro City's Law n.504/1984. The law can be accessed here.

⁶It can be accessed here.

and reached R\$ 6,300.00 in 2019. However, the educational performance of the municipality of Rio de Janeiro, as measured by Ideb, has significantly declined compared to other Brazilian municipalities. In the 5th-grade Ideb assessment ranking, Rio de Janeiro fell from 1423rd among all Brazilian municipalities in 2007 to 2529th in 2019. In the 9th-grade Ideb assessment ranking, Rio de Janeiro dropped 586 positions from 2007 to 2019.

It is relevant to mention that Brazilian municipalities' average increase of 63% in public expenditure per pupil on education from 2007 to 2019 contrasts with the unchanged performance in mathematics, reading, and sciences of Brazil's 15-year-old pupils in PISA at least since 2009 (countries performance in PISA are discussed in Appendix .A) . These numbers raise the question: What can be done if increased expenditure on education does not necessarily lead to significant improvements in education outcomes?

II.B. Science and Management for Education Programme (SMEP)

Given strong evidence of how management practices can enhance organisations' efficiency, we designed a public policy to improve schools' management practices. Policy-wise, our aim is to implement a programme to enhance pupils' educational outcomes through a low-cost intervention that can be easily adapted to different realities and expanded to large school networks. On the research side, our goal is to investigate the causal effect of management in schools on pupils' learning. Therefore, we developed a management intervention to be implemented as the treatment arm of a randomised field experiment. We name it as the Science and Management for Education Programme (SMEP).

To make this experiment possible, we convened several meetings with the Municipal Court of Accounts of Rio de Janeiro (TCMRio) and Municipal Secretariat of Education of Rio de Janeiro (SMERJ) to discuss the terms and conditions for conducting a randomised field experiment in state-run schools overseen by SMERJ. It is important to highlight that the Courts of Accounts in Brazil have the constitutional duty of analysing the efficiency and effectiveness of public policies. A partnership agreement was signed in June 2021 with the participating organisations, SMERJ and TCMRio, to formalise the collaboration. The agreement established the confidentiality of the experiment. Also, based on the collaboration, civil servants from the SMERJ and TCMRio were kindly allocated to form the programme implementation team. The team’s job was to deliver the SMEP treatment to schools.

In September 2021, we randomly selected eighty schools from the municipality of Rio de Janeiro population of 992 grades 1-9 state-run schools.⁷ From this random sample of eighty state-run schools, forty schools were randomly assigned to treatment and forty to control using a block pair-matching procedure detailed in Section III.A. It is relevant to mention that the control schools received no support. The collaboration agreement established that no one apart from the researcher would know the control schools’ identification to avoid any possible bias. More details are in Section III.A.

SMEP provided specific management practices to school managers in their day-to-day operations through one-to-one coaching sessions and on-the-job training. School managers are principals, deputy principals and pedagogical

⁷Educational units exclusively responsible for special or adult education and nurseries were excluded from the population of schools

coordinators (or supervisors). The intervention design was not based on traditional training programmes but on one-to-one coaching sessions and on-the-job training. The programme goal was to provide schools with the 23 ‘best’ school management practices described in detail by Bloom et al. (2015). These practices can be grouped into five groups: operations, monitoring, target-setting, people management and leadership. See Appendix B for more details about the description of these practices.

The programme implementation relies on the Agile project management methodology principles. We chose the Agile project management methodology for our project because of its ability to handle uncertainties common in real-world interventions such as a field experiment. The Agile approach makes it possible to implement the programme incrementally and interactively, delivering ‘small packages’ of the policy at each iteration (sprint) (Institute 2021, 2017).

These management practices for schools are similar to those outlined by the World Management Survey (WMS) across different sectors, such as manufacturing, retail and health care, but with changes to adjust the framework to the school context (Bloom and Van Reenen 2007, 2010; Bloom et al. 2014; Bloom et al. 2015). For each of the 23 ‘best’ management practices, the WMS has defined the worst (score one), the midway (score 3) and the best management scenario (score five). The team formed by civil servants helped to address the challenging issue of gaining the trust of public school managers. Certainly, state-run school managers are supposed to have much more in common with other civil servants who share similar knowledge and experience on Rio de Janeiro public administration than with private consulting

teams or researchers.

Although the team planned and reviewed actions together at each iteration, the programme was implemented individually or in pairs. Each team member was responsible for the same school for at least six months to ensure a good involvement (trust) with each school under their responsibility. Mutual trust between the team members and the school managers can facilitate the adoption of the best management practices. However, a very close friendship between team members and school managers could prevent our plans to maintain some level of school manager discomfort needed to generate movement. Therefore, we limited the member responsibility for the same school to at most one year.

Through one-to-one coaching sessions with principals (deputy principals) and pedagogical coordinators (supervisors), the team discussed the needs, goals, and possibilities to improve the school based on the best management practices. When the team felt the school managers could not discuss and implement determined practice because of a lack of knowledge or skills, the team delivered on-the-job training about that specific issue. The on-the-job training was done through one-to-one sessions at the school or through practical thematic seminars. The goal was to give school managers knowledge, tools, and procedures to improve their management skills.

Since the programme implementation occurred through one-to-one coaching sessions and on-the-job training with principals (deputy principals) and pedagogical coordinators, no contact was made between the implementation team and teachers or pupils. Although participation in the programme was not compulsory, the implementation team could contact the schools when-

ever necessary to influence them to increase participation. For example, even the more resistant schools received team contact and visits for two years to discuss the best management practices. The adherence to the programme was very heterogeneous. We suppose that with some institutional mechanisms for persuading the schools to participate, the intensity of participation would be larger, mainly for the less engaged school managers.

Principals, deputy principals and pedagogical coordinators (or supervisors) manage the school's learning production for their pupils. However, school managers do not deliver educational services directly to pupils; teachers do. Thus, better teacher management is critical to improving pupils' educational outcomes. Therefore, we designed the Science and Management for Education Programme (SMEP) to support school managers conducting one-to-one meetings with their teachers. One-to-one teacher coaching has proven to be a fundamental tool to improve education (Kraft, Blazar, and Hogan 2018). The programme provided support and training in teacher coaching to school managers followed Bambrick-Santoyo (2018). These meetings were the main channel for implementing many best management practices within the schools. The SMEP required that school managers conduct one-to-one meetings with teachers at least every two months.

Figure I shows five photos of the Science and Management for Education Programme in action. The top left photograph shows a principal organising pupils' educational performance data for one-to-one coaching sessions with the teachers. The top right photo highlights a team member coaching the principal and pedagogical coordinator on the target-setting group of management practices. The centre picture depicts the TCMRio auditorium filled

with principals and pedagogical coordinators after a long day of hands-on training on leadership. The bottom left photo shows a principal presenting information provided by the programme to pupils' parents on the returns of education. The bottom right photograph highlights two programme members coaching the principal while being observed by the pedagogical coordinator.

One last point concerns the challenge of accessing schools in areas of Rio de Janeiro controlled by criminals, where even the police have a hard time to enter in. Despite all the efforts to reach these schools, sometimes it was impossible to access some of them, as in the case of an armed conflict just happening near the school. When schools were inaccessible due to violence, the solution was to bring school managers of these schools to where the implementation team was based. They reported a strong feeling of being valued by our approach. This challenge highlights the very heterogeneous school realities across the city of Rio de Janeiro space. The following section outlines the Science and Management for Education Programme's timeline from January 2022 to December 2023.

II.C. Programme Implementation - 2022 to 2023

SMEP worked through iterations (sprints). The first sprint or phase lasted two months, from January to February 2022. It consisted of an opening seminar, one-to-one coaching sessions, and on-the-job training on two related management practices: 'Standardisation of Instructional Processes' and 'Data-Driven Planning and Student Transitions'.

The opening seminar introduced the management programme and encouraged principals, deputy principals, and pedagogical coordinators to par-

Figure I: Science and Management for Education Programme (SMEP)



Notes: The top left photo shows a principal organising pupils' educational performance data for one-to-one coaching sessions with the teachers. The top right photo highlights an implementer coaching school managers on the target-setting group of management practices. The central picture depicts the TCMRio auditorium full of treatment school managers after a long day of hands-on leadership training. The bottom left photo shows a principal presenting information provided by the programme on the returns of education to pupils' parents. The bottom right image highlights two implementation team members coaching a principal and being observed by the pedagogical coordinator.

ticipate in the initiative since participation was not compulsory. The team provided schools with feedback sessions on their management level based on a survey conducted at the end of 2021 (before the random assignment) to diagnose schools' management levels. This survey assessed the schools' management regarding the 23 best management practices. More details on the survey conducted in 2021 are in Section III.B.

In this sprint, there was clear resistance to using pupil data to direct actions. The team had to work carefully to overcome the situation. Furthermore, there was an evident lack of standard data analysis skills across the schools. For example, most school managers were not able to work with spreadsheets. Consequently, the programme provided more intense on-the-job training sessions to help managers to be able to analyse pupil data.

The second sprint started in March 2022 and aimed to provide schools with the following related management practices: 'Target Balance', 'Target Inter-Connection', 'Time Horizon of Targets', 'Target Stretch', and 'Clarity and Comparability of Targets'. The programme team chose to work with these practices to help school managers develop action plans demanded by the Secretariat of Education (SMERJ). Since the team worked using the Agile management project methodology, adapting to the Rio de Janeiro school network realities was expected.

The team used the SMART methodology to set goals and targets. A SMART goal should be specific, measurable, achievable, realistic and time-bound (Doran et al. 1981). The implementation team's goal was to help schools set their goals and targets and break down these goals and targets to teachers' and class levels. As part of the iteration, the team conducted a

practical seminar to help principals, principal deputies, and pedagogical coordinators with common doubts about the target-setting group of management practices.

In May 2022, the third sprint took place. The focus was on reviewing previous practices since the implementation team realised that some school managers had not fully absorbed the on-the-job training and were not confident in their skills to conduct one-to-one meetings with teachers. Also, the team continued to assist school managers in developing their school action plans as required by the Secretariat of Education (SMERJ). At this point, most but not all schools were already engaged with the programme and even seeking team support.

At the end of May, the team started a new iteration focusing on all management practices from the monitoring group: ‘Continuous Improvement’, ‘Performance Tracking’, ‘Performance Review’, ‘Performance Dialogue’, and ‘Consequence Management’. This sprint started with a workshop for pedagogical coordinators (supervisors) focusing on the monitoring group of management practices to ensure pupils’ learning. The team again highlighted the relevance of using one-to-one coaching sessions with teachers to improve the schools’ management.

After the mid-year school holidays in Brazil (July), the team continued developing the monitoring group of practices together with the target-setting management practices with the schools (August 2022). The team also started to discuss ‘The Personalisation of Instruction and Learning Practice’. This iteration lasted one month.

In September and October 2022, the team changed the focus to leadership

practices. First, principals, deputy principals and pedagogical coordinators were invited to participate in a full-day practical workshop on leadership. They had to solve school practical problems related to three management practices: ‘Leadership Vision’, ‘Clearly Defined Accountability for School Leaders’, and ‘Clearly Defined Leadership and Teacher Roles’. During the workshop, school managers were also encouraged to consider the management practice ‘Adopting Educational Best Practices’ from the operating group of practices.

Following the workshop, the two-month sprint included one-to-one sessions with school managers to deepen their understanding of leadership practices and ‘Adopting Educational Best Practices’. The team provided school managers with simple slides on the returns to education. School managers were orientated to present the slides to parents and pupils to make them more aware of the importance of education. The team informed school managers that giving information to parents and children on the returns to education could improve pupils’ frequency and performance (Nguyen 2008). The feedback from school managers regarding the slides was very positive. School managers said parents were very interested in the slides and even asked to receive them in print.

During the November and December 2022 iteration, the team coached schools regarding all the aspects discussed throughout the year. The team also discussed ways to help school managers implement the following people management practices: ‘Rewarding High Performers’, ‘Removing Poor Performers’, ‘Promoting High Performers’, ‘Managing Talent’, ‘Retaining Talent’, and ‘Attracting Talent’. However, there was very little space to develop

these practices. In Brazil, the legislation makes it difficult to promote, reward, remove or fire civil servants, such as public school teachers, based on their performance.

The programme implementation finished in December 2022 with the year's last seminar, coinciding with the end of the school year in Brazil. The team invited school managers who best implemented the programme, some from very challenging local settings, to present their experiences to other treatment schools. The goal was to share experiences across treatment school managers on the programme. The goal was also to influence less engaged schools, showing the success of school managers who were working in very challenging realities.

The event caught schools' attention and encouraged them to participate even more in the SMEP. A post-event survey showed that almost all school managers were satisfied with the seminar and the programme. Since the local settings are very different across the municipality of Rio de Janeiro, school managers were pleased to see that there was a way to improve school management practices even in the most challenging scenarios.

The Science and Management for Education Programme (SMEP) resumed in January 2023, focusing on deepening the school's adherence to the 23 best management practices. In the first year, school managers were extensively exposed to the best management practices and necessary tools; in contrast, the second year focused on deepening the implementation of these practices. During the first year, on-the-job training was more prevalent than one-to-one coaching sessions, but the second year primarily relied on the latter. The team continued to provide on-the-job training whenever necessary,

but the focus was mainly on coaching sessions with school managers. With the new skills acquired by school managers from the first year summed to a less data-resistant mindset, the coaching meetings with the programme team improved significantly in 2023.

Apart from the sprints to coach school managers regarding the best management practices extensively presented in 2022, two relevant seminars were conducted in 2023. Both utilised the same dynamic from the last seminar of 2022, with the most engaged school managers, mainly from challenging school realities, presenting their experiences to advance the management in their schools. Again, the post-event surveys showed that school managers were very satisfied with the experiences shared by them.

III. METHODS AND DATA

III.A. The Selection of Schools for the Management Experiment

Our study complies with all critical research protocols. It received approval from the University of Cambridge’s Ethical Committee and is registered on the AEA RCT registry (AEARCTR-0007669 on 15 May 2021; <https://www.socialscisceregistry.org/trials/7669>). The intervention was not sensitive. Rio de Janeiro authorities and institutions responsible for education in the city also approved the experiment.

We conducted a clustered pair-matching randomised field experiment with eighty schools randomly selected from the municipality of Rio de Janeiro population of grades 1-9 schools to analyse the impact of our management programme on pupils’ educational outcomes. It is clustered because we ran-

domly assigned schools, not pupils, to treatment. It is pair-matching because we formed school pairs based on previous educational achievements, and within each pair, we randomly assigned one school to treatment and the other to control. Field because the experiment was conducted as an ordinary pilot government project within the Brazilian public administration.

Our sample of eighty schools was randomly selected from the 2021 Rio de Janeiro population of 992 grades 1-9 public schools. These schools have two main types: first-segment schools,⁸ responsible for 1st to 5th grade⁹, and second-segment schools,¹⁰ responsible for 6th to 9th grade. There are 70.9% of first-segment schools and 29.1% of second-segment schools under the Municipal Secretariat of Education of Rio de Janeiro (SMERJ).

The schools participating in the programme were unaware they were part of an experiment, and the sample identification was kept confidential. The participating organisations signed a confidentiality agreement to protect the experiment's integrity. Only one member of the programme implementation team had access to the sample identification, and he was required to keep the information entirely confidential. The term 'experiment' was avoided in the programme implementation vocabulary. The project was presented to the schools as a typical initiative conducted in public administration. These measures were implemented to minimise the risk of biases and spillovers.

One of the differences between first-segment and second-segment schools is their teaching approach. First-segment schools have generalist teachers

⁸They are also known as primary schools in the UK and US.

⁹As a rule, schools from first-segment have classes from first to fifth grades; however, some first-segment schools have the 6th grade

¹⁰They are also known as lower secondary schools in the UK and US.

who teach multiple subjects to one class, while second-segment schools have specialist teachers who teach one subject to various classes. This means that second-segment school managers have more teachers to manage, which can make their job more complex. Our random sample of 80 schools had 56 schools from segment I (grades 1-5) and 24 from segment II (grades 6-9). Thus, we conducted separate random assignments for segments I and II schools to ensure equal representation in both experimental groups, control and treatment. This method allowed us to assess the causal effect of our program across both segments and within each segment. Block randomisation also may generate smaller standard errors without downsides for the analysis (Gerber 2012).

Using information correlated to the outcome of interest in the random assignment can avoid the emergence of large and significant random differences in the outcome between treatment and control groups before the experiment starts (Abadie and Imbens 2011; Fryer 2017). Our outcome of interest is pupils' learning. Unfortunately, at the time of the random assignment, we could not access the scores from the standardised reading and mathematics tests, Rio Assessments, applied by the Secretariat of Education (SMERJ) to all pupils in 2021. The only pupils' learning assessment results available in October 2021 were the 2019 Brazilian Assessment System for Basic Education or Saeb 2019. The Saeb 2019 consisted of mathematics and reading tests intended to be applied nationally to all 5th and 9th-grade pupils.¹¹

We adopted a pair-matching procedure to conduct our random assign-

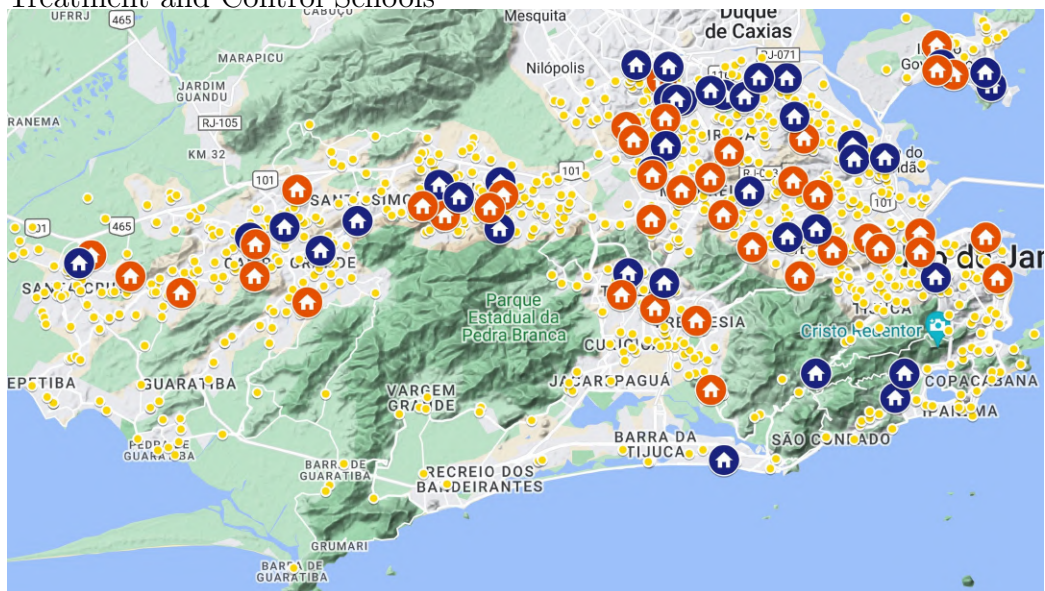
¹¹Only the scores from schools where more than 80% of their 5th/9th-grade students participated in the 2019 Saeb were accessible.

ment within each school block (segment). We started the random assignment procedure by creating a school ranking by block, first-segment and second-segment schools, based on the descending order of the average school score in the Saeb 2019 (sum of reading and mathematics school averages divided by two). First-segment schools were represented by their 5th-grade average score, and second-segment schools by their 9th-grade average score.

Within each block (school segment), we arranged pairs of schools in descending order of the Saeb 2019 average and randomly assigned one school to the treatment group and the other to the control group. This process was repeated for each block until all forty pairs were assigned. We generated four groups of schools: treatment and control groups for first-segment schools (28 schools each) and treatment and control groups for second-segment schools (12 schools each). This method ensured balance regarding the school segment and Saeb 2019 school average score across the treatment and control groups.

Figure II shows the map of the municipality of Rio de Janeiro with the geographical distribution of their population of 992 grades 1-9 state-run schools in 2021. Blue and red circles represent the treatment and control schools, respectively, and yellow dots represent the remaining schools. Treatment and control schools' distribution followed the population distribution pattern. The next section details the process to generate data on the schools' management level before (September 2021) and after (December 2023) the experiment in 2021.

Figure II: Geographical View of the City of Rio de Janeiro Schools, Including Treatment and Control Schools



Notes: Blue and red circles are treatment and control schools, respectively. Yellow dots represent the remaining grades 1-9 public schools in the city of Rio de Janeiro. Source: Created using My Map from Google Maps. Schools' address information provided by the Municipal Secretariat of Education of Rio de Janeiro (SMERJ).

III.B. Data I - Management Level of Schools in 2021 and 2023

We conducted two surveys with the 80 selected schools. The first survey was conducted just before the random assignment in September 2021. The second survey was conducted in December 2023, after the intervention. The surveys were double-blind since school managers were unaware their schools were being scored, and interviewers were unaware of the schools' educational performance. We followed closely the survey methodology developed by the World Management Survey - WMS (Bloom and Van Reenen 2007, 2010; Bloom et al. 2014; Bloom et al. 2015). The only difference between our survey and the WMS methodology is that we conducted face-to-face interviews instead of telephone interviews.

The 2021 survey aimed to diagnose the management level of the schools in our sample and to train the implementation team to deliver the treatment to schools. The 2023 survey had three goals: (i) to analyse the efficacy of the treatment in providing the 23 best management practices; (ii) to assess the treatment's impact on the schools' management level; and (iii) to analyse the causal effect of school management practices on pupils' learning.

We used the WMS questionnaire and scoring grid developed by Bloom et al. (2015) to measure the management of our sample schools regarding the 23 best management practices. The questionnaire and scoring grid are in Appendix H. The questionnaire has 23 parts, each one representing a management practice. Based on the school manager's answers, each of the 23 management practices should be scored against a grading score that ranges from 1, worst management level, to 5, best management level.

For instance, in practice 1) of the questionnaire, 'Standardisation of the Instructional Processes', a school receives a score of one if 'No clear or institutionalised instructional planning processes or protocols exist; little verification or followup is done to ensure consistency across classrooms', a score of three if 'School has defined instructional planning processes or protocols to support instructional strategies and materials and incorporate some flexibility to meet students' needs; monitoring is only adequate', and a score of five if 'School has implemented a clearly defined instructional planning process designed to align instructional strategies and materials with learning expectations and incorporate flexibility to meet student needs; these are followed up on through comprehensive monitoring or oversight'.

The WMS questionnaire comprises open-ended questions that avoid lead-

ing respondents towards a particular answer. For example, the initial inquiry in the performance monitoring aspect is, ‘What kind of main indicators do you use to track school performance?’ rather than a close-ended query like, ‘Do you evaluate your school’s performance using the academic scores of students [yes/no]?’ (Bloom et al. 2015). The initial open-ended inquiry is succeeded by additional questions such as ‘How frequently are these indicators measured?’, ‘Who gets to see this data?’ and then ‘If I were to walk through your school, what could I tell about how you are doing against your indicators?’. The interviewers could ask follow-up questions whenever necessary. The goal is to give scores regarding the actual management practices of a school and not on the school managers’ aspirations of what the school’s management should be.

The WMS also developed an updated and more detailed version of the scoring grid contained in the original version of the questionnaire. It is called D-WMS.¹² Both scoring grids, original WMS and D-WMS, are compatible. Both scoring grids go from one, worst management, to five, best management. We tested the D-WMS scoring grid in some schools outside our experimental sample. Regarding interviewers’ training, the D-WMS tool was much more time-consuming and difficult to use. Moreover, scoring schools was more time-consuming and difficult since the interviewers needed to learn and think about many more possibilities to score a school than the original version. Therefore, we opted to use the original WMS questionnaire for practical matters.

The programme implementation team members conducted the manage-

¹²here.

ment survey in September 2021. Since the random treatment assignment had not yet occurred, there were no concerns about biases from the team at that time. However, after two years of intervention, the team was strongly involved with the treatment schools, which could be a clear source of bias in the second survey. Therefore, we requested one of our partners, the Court of Accounts of Rio de Janeiro (TCMRio), if seven auditors, not involved with the experiment, could conduct the management survey in December 2023. The Court agreed to allocate them from November to December 2023 for training sessions and for conducting the 2023 management survey in the 80 sample schools.

The interviewers in 2021 and 2023 had an intensive two-week training on the WMS methodology. The 2021 and 2023 survey teams knew well about the functioning of the Rio de Janeiro state-run schools since they were bureaucrats specialised in education and public management from the Rio de Janeiro public administration. This knowledge was crucial to make practices observed in schools adequately codified along the lines proposed by Bloom and Van Reenen 2007, 2010; Bloom et al. 2014; Bloom et al. 2015. As part of the training sessions, the teams also had practical experience applying the WMS survey tool to some schools selected from outside the study sample. After the training surveys, the interviewers discussed the scoring procedure. The goal was to align the scoring patterns across the members of the survey teams.

For both surveys, the Municipal Secretary of Education of Rio de Janeiro (SMERJ) contacted the selected schools and informed them that interviews would be conducted with school principals to discuss local school operations.

No more information was given to schools. School principals have enough knowledge of school management practices and are involved in the day-to-day school activities. Other senior managers, such as the secretary, superintendents, and region coordinators, are supposedly working with more strategic actions far from regular school activities. The SMERJ was informed not to disclose the research nature of the survey. The word interview was adopted to obfuscate the idea that participant schools would be graded in line with the practice proposed by Bloom and Van Reenen 2007.

We applied all procedures recommended by the WMS to avoid potential bias from our survey (Bloom and Van Reenen 2007; Bloom et al. 2015). Each team member received a list of schools with only their name and contact details. Selected schools were distributed randomly to interviewers but no interviewer could apply the survey in the school in her neighbourhood. This is a mechanism to ensure that schools were not known to interviewers. In addition, interviewers were instructed to not get more information about their allocated schools. Interviewers were oriented to contact their allocated schools and to arrange face-to-face interviews with school principals. They were also advised to not discuss pupils' educational results.

One relevant point stressed by the WMS methodology is that the survey should focus on managerial practices and not on principals' views on how schools should be managed. The interviewers were allowed to make annotations, but they were not supposed to score the school in front of the principal. The coding process happened after the interviews and outside school facilities. As the interviews were recorded, the interviewers could consult the recording. The interviews in both surveys lasted, on average, 75 minutes.

We achieved a 100% response rate in both surveys, i.e., the principals of the 80 schools participated in the two surveys. This is a much higher response rate than in other well-known management studies conducted across different sectors by the World Management Survey (WMS), which reached, on average, a 41% survey response rate according to Bloom et al. 2015.

We double-scored all interviews in both surveys, such that each management practice in each school had two different scores given by two different coders (also called raters). The interviewer (first coder or rater) gave each school's first (primary) scores. The second scoring round for each school was conducted using the interview recordings by survey team members chosen randomly. The goal was to assess the inter-rater reliability or inter-rater agreement among raters.

It is important to ensure that raters scored schools based on a systematic management practice assessment following the WMS methodology. The management level of a school should be an attribute of the school and not an attribute of any survey team member (Gwet 2021). The Appendix C shows high agreement coefficients in both surveys, indicating that scores assigned to each of the 23 management practices for each school are independent of the specific survey team member (rater or coder) who evaluated the school.¹³ The evidence suggests that the management scores given to schools are not subjective views of raters but actual school characteristics.

The scores given by the interviewers (raters 1) represent the schools'

¹³Table A3 in Online Appendix C reports for the 2021 survey a Brennan-Prediger coefficient of 0.720 and a Gwet's AC2 of 0.760. The 2023 survey presents higher agreement coefficients, such as a Brennan-Prediger of 0.788 and a Gwet's AC2 of 0.811. These values are considered by Landis and Koch (1977) to be relatively high or close to perfect agreement between raters.

management in three different formats. Firstly, the general management of a school is the average of the scores reached by a school regarding each of the 23 management practices surveyed. Secondly, since the 23 practices can be grouped into five groups - target setting, leadership, operation monitoring and people management - each group in each school is represented by the average of the scores reached by that school regarding the practices that form the specific group. Since each management practice is scored against a scale ranging from one, worst management, to five, best management, the general management and group of practices averages also follow the same scale from one to five. Thirdly, analyses are also conducted separately with each of the 23 management practices.

III.C. Data II - Other Relevant Variables

The Municipal Secretariat of Education of Rio de Janeiro (SMERJ) provided the following pre-treatment data at the individual level: pupil's race, gender, grade (1-8), 2021 Rio Assessments scores, absence rate, whether the pupil's family in 2021 was a beneficiary of the Brazilian cash transfer program (i.e., Bolsa Família); race, gender, age, education, and experience for principals and pedagogical coordinators; and gender and education for teachers. At the school level, the SMERJ provided information on school size by number of pupils and school segment (segment I, 1st to 5th grade or segment II, 6th to 9th grade). We also generated data on the management level of schools in 2021 through the management survey conducted in 2021. All these covariates were collected before January 2022 and are considered pre-treatment covariates. Regarding the 2021 Rio Assessments, the scores are

based only on the Classical Test Theory (rate of right answers).

Unfortunately, around a third of pupils' scores from the Rio Assessments applied in 2021 are missing. Many pupils were not assessed due to COVID-19 issues. To deal with missing data, not only in the 2021 pupils' scores but also in the other pre-treatment covariates, we use the Missing Indicator Method (Kayembe et al. 2022; Zhao and Ding 2022).

Another issue with the 2021 Rio Assessments was related to where the tests were taken by pupils. Although the Rio Assessments were designed to be taken at school, they were also allowed to be taken from home due to COVID-19 pandemic, making it difficult to control for interference.

The primary outcome variable comes from Rio Assessments, which are applied four times each academic year - April, June, September, and December — to all first—to ninth-grade pupils enrolled in a school from SMERJ.¹⁴ Rio Assessments were developed by a specialised centre from a federal Brazilian university and started to be applied in the municipality of Rio de Janeiro in April 2021. An external organisation sets the exams, but schools apply the tests.

The fact that schools administer the Rio Assessments exams may raise questions related to teachers' external influence on treatment relative to control pupils' scores. However, this is unlikely, given how exams are administered. Typically, the teachers administering the exams are not the same as those teaching the classes. Additionally, answer sheets sent to schools are uniquely identified for each student. If two different options are marked, the answer is considered incorrect. Furthermore, since answer sheets contained

¹⁴This is called in Portuguese *Atividade Diagnóstica em Rede* (ADR).

the identification of each pupil, it is improbable that school managers could make any changes directly without the involvement of teachers.

In order to artificially increase their pupils' scores in Rio Assessments, school managers would need to make teachers agree to cheat, as teachers are accountable for administering exams. School managers do not have ways to force teachers to act wrongly since they cannot promote, reward, remove, or fire teachers in Brazilian public schools. Teachers have strong stability in Brazil's public administration to resist illegal or unethical demands. Also, there were no additional rewards for principals if school performance improved.

In the design of our experiment, treatment and control schools were unaware that they were part of a scientific experiment. A strict confidentiality agreement was in place. Additionally, schools were informed that they were not being ranked or compared to other schools and that all management data collected as part of the program implementation were anonymised to prevent any school identification. Therefore, there was no strong incentive for treatment schools to engage in unethical and illegal behaviour of artificially and systematically altering their pupils' scores.

The Rio Assessments are formed by reading and mathematics tests that are based on the Common National Curriculum Base (BNCC).¹⁵ Our primary analysis relies on the scores from the Rio Assessments that use the Item Response Theory (IRT) to score pupils against the same cross-grade scale of competencies as the Brazilian Assessment System for Basic Educa-

¹⁵See <http://basenacionalcomum.mec.gov.br>.

tion (Saeb).¹⁶ Pupils from different grades are scored based on the same competency scale, allowing comparability across grades.

Based on the December 2023 Rio Assessments and considering only control pupils, we observe that the average difference in pupils' scores by grade was 13.14 score points for reading and 14.73 for mathematics.¹⁷

Figure A5 in Appendix D shows the control pupils' scores by grade for reading and mathematics in the December 2023 Rio Assessments. Boxplots represent the distribution of scores for each grade. The average score from the end of the 2nd to the end of the 9th grade varies around 100 points in reading and mathematics. With seven years of schooling, pupils acquire, on average, 100 score points in reading and mathematics in Rio de Janeiro. However, the variation across grades is very heterogeneous. The largest reading and mathematics learning gains happened through the 4th and 5th grades.

Table A4 in Appendix D provides data on pupils' missing reading and mathematics scores between 2022 and 2023. The fixed sample data consists of 31,760 pupils from the 80 sample schools. The first two columns of the table, labelled 'Before adjustment', show the number of missing scores and the percentage of these missing scores, considering the complete sample of 31,760 pupils, before any changes were made to fill in the missing data. Missing data can be attributed to many possible reasons, such as some pupils being transferred to schools outside the sample, some being unable to attend

¹⁶Saeb scale here

¹⁷Only pupils from grades 1 to 8 at the beginning of the 2022 school year were considered in our fixed sample. However, some pupils failed to go to the next grade in 2023, and 25 pupils in grade 1 in 2022 remained in the same grade in 2023. Since only these 25 pupils represent the 2023 1st grade in our sample, we do not use the 2023 grade 1 in this specific analysis.

school on the test day, and others moving from a treatment school to a control school or vice versa. The fourth and fifth columns, labelled ‘After adjustment’, show the missing data information after filling in each pupil’s missing performance data with their standardised scores from previous tests on the same subject.

For instance, if a pupil does not have the standardised reading score from the December 2023 Rio Assessments, we use the standardised reading score from the September 2023 Rio Assessments to replace this missing score. If the September 2023 Rio Assessments reading score is also missing, we use the standardised score from June 2023 Rio Assessments. If necessary, we continue this process until April 2022 (the first Rio Assessment after the beginning of the treatment). We only replace missing data with previous standardised scores to maintain the pupil’s relative position in their population grade in a specific subject and Rio Assessments application date. Remember that the scores were standardised by subject, grade and application date using the control group as the reference. Replacing missing data with previous pupils’ performance information is a conservative approach since pupils’ older standardised scores are from when the pupils had been exposed to the programme for less time.

Based on this adjustment approach, our attrition rate is almost null, i.e., there is only 1.10% of pupils’ missing data for reading and 1.12% for mathematics. Regressions of a dummy representing missing data after the adjustment on a treatment dummy show that the attrition rate differences between treatment and control are 0.16% for reading and 0.18% for mathematics; both differences are not statistically different from zero at standard

confidence levels. Even before the adjustment, the differences were 0.34% for reading and 0.13% for mathematics, which are not statistically significant.

From SMERJ, we also had access to the scores from the Rio Assessments based on the Classical Test Theory (CTT), i.e., the scores were As an outcome variable, we also use the simple proportion of right answers, i.e., the Classical Test Theory (CTT). We replaced missing data following the same procedure we have used for the IRT scores, filling in each pupil’s missing performance data with their standardised scores from previous Rio Assessments on the same subject.

III.D. Pre-Treatment Summary Statistics

The school year in Brazil runs from January to December, with the regular enrolment period taking place between November and December of the previous year. We focus on pupils enrolled for the 2022 academic year during the standard enrolment period from November to December 2021. We did not consider pupils enrolled later in the 2022 academic year. As pre-treatment variables, we collected pupil characteristics at the end of 2021.

As previously mentioned, we excluded pupils enrolled in the 9th grade for the 2022 academic year since they were supposed to attend upper-secondary schools (high schools) in 2023, which is outside the scope of the study. The treatment was planned for two years, from January 2022 to December 2023. Consequently, the pupil sample in this study consists of 1st- to 8th-grade pupils enrolled (during November/December 2021) in a sample school for the 2022 school year.

Table I: Pre-Treatment Statistics - Pupils sample

	Valid obs	Control (mean)	Treatment (difference/SE)
A. Characteristics			
Female (%)	31,427	0.488	-0.003 (0.006)
White (%)	28,743	0.335	0.019 (0.019)
Brown (%)	28,743	0.543	-0.015 (0.013)
Black (%)	28,743	0.120	-0.003 (0.010)
Bolsa família (% receiving)	31,427	0.340	-0.060 (0.033)
B. Educational outcomes			
Reading - September 2021	22,702	0.000	0.088 (0.048)
Mathematics - September 2021	22,663	0.000	0.086 (0.050)
Reading - December 2021	23,581	0.000	0.081 (0.053)
Mathematics - December 2021	23,544	0.000	0.063 (0.054)
Pupils total (Schools)	31,760 (80 schools)	16,134 (40 schools)	15,626 (40 schools)

Notes: This table presents information on the characteristics and educational outcomes of pupils enrolled (during November/December 2021) in a sample school for the 2022 school year. The column ‘Valid obs’ shows the number of observations with data available. The ‘Treatment’ column shows the difference between the treatment and control averages with the standard errors (SE) of the differences reported in parenthesis. ‘SE’ means clustered robust standard errors with clusters defined as the school pairs used for the random assignment. Each difference between averages and standard error comes from a regression analysis where relevant variables in the table are regressed on a treatment dummy. Pupils’ scores from the 2021 Rio Assessments were standardised by subject, grade, and test application date using the control group as the reference.

We have 31,760 pupils from 80 schools randomly sampled from the population of the city of Rio de Janeiro state-run schools. In the 2022 school year, 16,134 and 15,626 grade 1-8 pupils were enrolled in control and treatment schools, respectively,

during the regular enrolment period from November to December 2021.

Table I presents pre-treatment information on the characteristics and educational outcomes of grade 1-8 pupils enrolled in a sample school for the 2022 school year. The ‘Treatment’ column shows the difference between the treatment and control pupils’ averages with the standard errors (SE) of the differences reported in parenthesis. ‘SE’ means clustered robust standard errors with clusters defined as the school pairs used for the random assignment. Each difference between averages and standard error comes from a regression analysis where the variables in the table are regressed on a treatment dummy.

The educational outcomes in Table I come from Rio Assessments, which are standardised reading and mathematics tests taken by all pupils across two different dates: September and December 2021. Scores from the 2021 Rio Assessments were standardised by subject, grade, and test application date using the control group as the reference. The column ‘Valid obs’ shows the number of observations with data available.

According to Table I, black pupils in the control group are 12%, while they are 11.7% in the treatment group. Brown pupils are 54.3% and 52.8% in the control and treatment groups, respectively. The control group has 34% of pupils in beneficiary families in Brazil’s Bolsa Familia cash transfer, while the same fraction in the treatment group is 28%. Regarding pupils’ educational outcomes, pupils’ scores are slightly higher in the treatment group; surely, these are differences raised by chance since we randomised the treatment. Since the control group was the reference group for the standardisation procedure, the average scores of the control group score were zero.

Table II: Pre-Treatment Statistics - Schools sample

	Valid obs	Control (mean)	Treatment (difference/SE)
Principals			
Female (%)	80	0.900	-0.100 (0.080)
White (%)	69	0.459	-0.053 (0.121)
Brown (%)	69	0.405	0.126 (0.121)
Black (%)	69	0.135	-0.073 (0.072)
College degree (%)	80	0.725	0.025 (0.100)
Experience (years)	70	6.4	1.0 (1.5)
Pedagogical Coordinators			
Female (%)	75	0.973	-0.078 (0.057)
White (%)	58	0.645	-0.127 (0.131)
Brown (%)	58	0.323	0.011 (0.126)
Black (%)	58	0.032	0.116 (0.077)
College degree (%)	74	0.811	0.027 (0.090)
Experience (years)	60	3.1	0.0 (0.7)
Teachers (average by school)			
Female (%)	80	0.814	-0.002 (0.034)
College degree (%)	80	0.816	0.035 (0.032)

Notes: This table presents information on the characteristics of principals, pedagogical coordinators and teachers in a sample school for the 2022 school year. The column ‘Valid obs’ shows the number of observations with data available. The ‘Treatment’ column shows the difference between the treatment and control averages with the standard errors (SE) of the differences reported in parenthesis. ‘SE’ means clustered robust standard errors with clusters defined as the school pairs used for the random assignment. Each difference between averages and standard error comes from a regression analysis where relevant variables in the table are regressed on a treatment dummy.

Table II reports the 2022 average characteristics of principals, pedagogical coordinators, and teachers for treatment and control schools (data collected at the end of 2021). The column ‘Valid obs’ shows the number of observations with data available. The ‘Treatment’ column shows the difference between the averages of the treatment and control groups with the standard errors (SE) from the differences in parenthesis. For all variables, the average difference is not statistically different from zero at usual confidence levels.

According to Table II, most principals and pedagogical coordinators are white females. In addition, the average experience of principals is 6.4 and 7.4 years for control and treatment schools, respectively, and the average experience for pedagogical coordinators is 3.1 years for both groups. Most principals, pedagogical coordinators and teachers have a college degree in both groups.

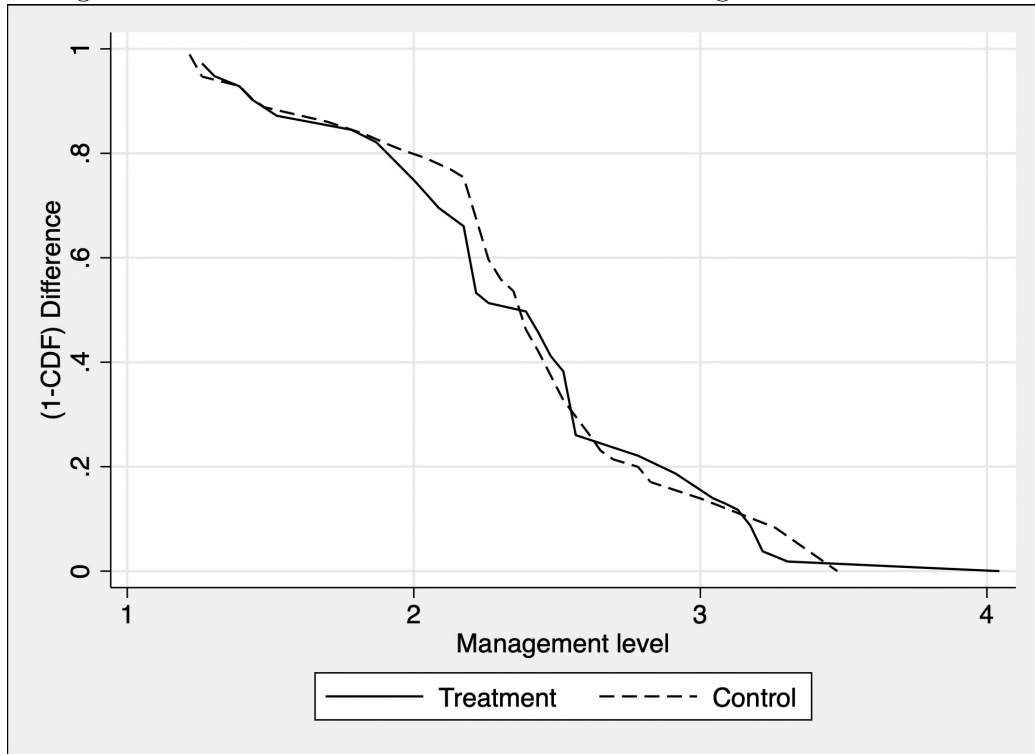
We also conducted a management survey in 2021 to evaluate the management level of our experimental sample of 80 schools randomly selected from the Rio de Janeiro population of grades 1-9 schools. Each school was assessed based on the 23 best management practices identified and discussed by Bloom et al. (2015). Each practice was scored based on a scoring grid that goes from one, worst management, to five, best management. An average management score was calculated for each school, averaging the score reached in each of the 23 practices.

A regression of the 2021 school management level on a treatment dummy using robust standard errors shows a not statistically significant difference of 0.027 (0.857) score points between treatment and control school management levels¹⁸. The 2021 management level of control schools was 2.407, and that of treatment schools was 2.379. This management level means that Rio de Janeiro public schools had the 23 best management practices implemented at a relatively low level, not

¹⁸If the management is standardised to have a mean of zero and a standard deviation of one regarding the control group, the difference is 0.048 (0.267).

achieving even the midway management level of 3 and far from the highest level of implementation of the best practices, 5.

Figure III: Treatment and Control Schools Management Level in 2021

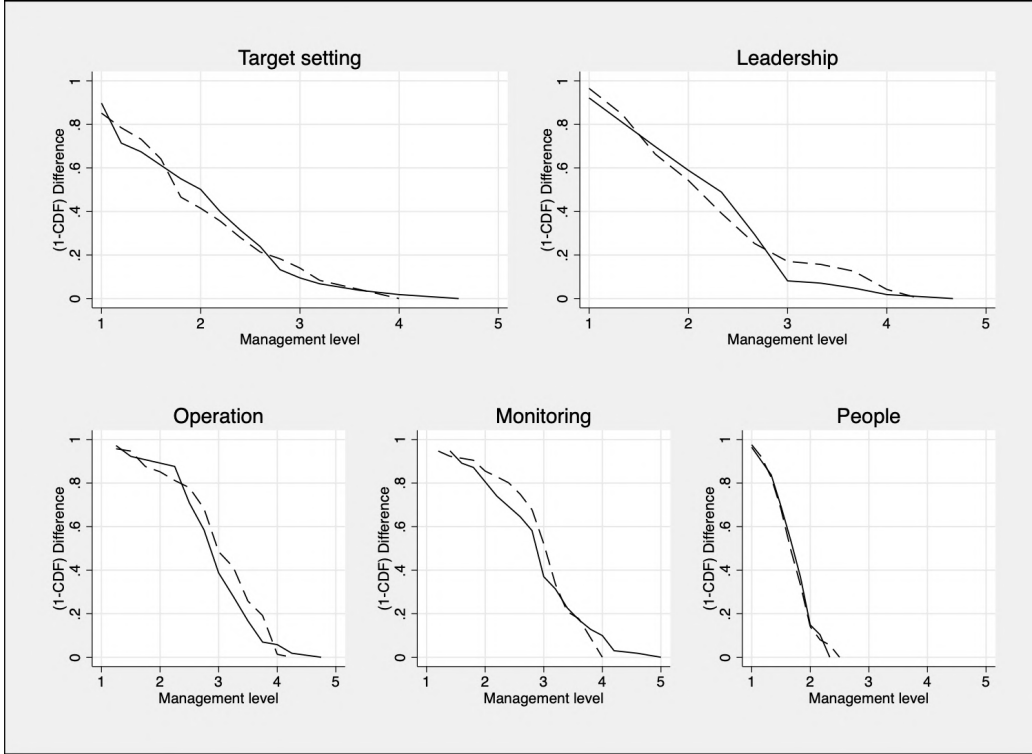


Notes: The figure illustrates the difference between treatment and control schools regarding the probability of a school having its management level greater than a specific management level on the x-axis (i.e., one minus the CDF). Although the management level scale can go from one, worst management, to five, best management, the x-axis shows only the scale range where there are observations to make the graph better to visualise. The solid line represents treatment schools, and the dashed lines control schools.

Figure III reports for treatment and control schools the probability of a school having its management level, based on the 23 best management practices, greater or equal to a specific management level on the x-axis (i.e., one minus the CDF). The solid line represents treatment schools, and the dashed lines control schools. This figure shows that treatment and control schools had very similar management levels in 2021 based on such cumulative probability.

Figure IV reports similar probability by grouping the 23 management practices

Figure IV: Treatment and Control Schools by Management Groups in 2021



Notes: The figure illustrates the difference between treatment and control schools regarding the probability that a specific group of management practices of a school is greater or equal than a specific management level on the x-axis (i.e., one minus the CDF). The management level on the x-axis goes from one, worst management, to five, best management. The solid line represents treatment schools, and the dashed lines control schools.

into five groups: target setting, leadership, operation, monitoring and people management. Once more, the solid line represents treatment schools, and the dashed lines control schools. Almost 90% of the schools did not reach level 2 of the people management practices group. Around 50% of schools have levels lower than 2 in target setting and leadership. Rio de Janeiro schools were relatively better managed in operation and monitoring groups of management practices. As expected from a random assignment procedure, the control and treatment reported cumulative probabilities were very similar across the five groups of management practices.

Based on Table I, Table II, Figure III and Figure IV, we can conclude the treatment and control groups are, on average, very similar, as expected from a random assignment procedure. Our experimental sample is also very similar, on average, to the population of pupils and schools in Rio de Janeiro, as we can see in our Appendix E.

III.E. Econometric Specifications

We measured school management through a survey conducted in December 2023 with our 80 sample schools following the WMS methodology. The schools were assessed regarding each of the 23 best management practices discussed in this study. Each practice was scored against a score grade that goes from one, worst management, to five, best management. Therefore, each school has a management score for each of the 23 practices. Detailed information about the management survey can be found in Section III.B.

The general management of a school after the treatment is the simple average of the scores reached in each of the 23 practices in our December 2023 management survey. Our outcome variable M_s can represent one specific management practice, a group of management practices or all 23 management practices of the school s . Let Z_s be a dummy that is one if the school s was randomly assigned to treatment and zero if the school s was randomly assigned to control. Let n_s be the number of sample pupils in each school s . Since the number of pupils per school varies from schools with 156 pupils to schools with 1346 pupils, it is important to weigh this school-level analysis with the number of pupils per school. Therefore, the average treatment effect (ATE), λ_{ATE} , can be identified using the following weighted regression:

$$M_s\sqrt{n_s} = \iota\sqrt{n_s} + \lambda_{ATE}Z_s\sqrt{n_s} + \varepsilon_s\sqrt{n_s}. \quad (1)$$

Equation 1 identifies the average treatment effect (ATE) of the programme on the management of the schools. Equation 1 can be used to identify the ATE of the programme on a specific management practice or even in a group of management practices. Standard errors are robust standard errors (Chaisemartin and Ramirez-Cuellar 2024). It is relevant to mention that the attrition rate is null since all the eighty sample schools participated in the December 2023 management survey.

We also aim to estimate the impact (the average treatment effect - ATE) of the Science and Management for Education Programme (SMEP) on pupils' educational outcomes. The analysis is done separately by subject. The treatment lasted two years, from January 2022 to December 2023.

It is essential to mention that the attrition rates regarding pupils' scores in our sample are only 1.1% of the reading and mathematics scores from the December 2023 Rio Assessments. This is due to the approach of replacing December 2023 missing outcome data with pupils' standardised scores¹⁹ from previous tests. It is a conservative approach since we used information from pupils exposed to the treatment for less than two years. Moreover, the attrition rate differences between treatment and control are 0.16% for reading and 0.18% for mathematics; both differences are not statistically significant. The differences are the same before or after the adjustment. More details on the approach used to deal with missing outcome data can be found in Section III.C.

For each pupil i , let Y_{isp} be the outcome of interest: individual reading or mathematics score from the December 2023 Rio Assessments standardised by sub-

¹⁹Scores were standardised by grade, subject and test application date using the control group as the reference.

ject and grade using the control group as the reference. Thus, considering the control group, the scores have a mean of 0 and a standard deviation of 1 in each grade of each subject. Let Z_{isp} be an indication variable that takes value one when the pupil was enrolled (during November/December 2021) in a treatment school for the 2022 academic year and zero when the pupil was enrolled (during November/December 2021) in a control school for the 2022 school year. The subscript s represents the school in which the pupil i was enrolled for the 2022 school year, and p is the pair of schools used for the random assignment that included the school s . Therefore, the average treatment effect (ATE), β_{ATE} , can be identified for each subject using the following causal equation:

$$Y_{isp} = \alpha + \beta_{ATE}Z_{isp} + \epsilon_{isp}. \quad (2)$$

The unadjusted regression shown in Equation 2 identifies the programme’s average treatment effect (ATE) on pupils’ educational outcomes for each subject. The ATE identification and estimation relies on the use of clustered robust standard errors with clusters defined as the school pairs used for the random assignment plus the non-inclusion of pair-fixed effects in the regression (Chaisemartin and Ramirez-Cuellar 2024).

The previous equations focused on the average causal effect of the treatment on pupils’ learning and school management. We also want to understand the impact of the intervention on pupils’ educational achievements through the intensity of the treatment delivered. The causal variable of interest here is the management level of the schools measured by the management survey conducted in December 2023. The management level of an experimental school is the simple average of the 23 scores received by the school in each of the 23 management practices (dimensions) surveyed. As said before, each management dimension is scored against a grid of

score points from one, worst management, to five, best management level. More details about the 2023 management survey can be found in Section III.B.

Since our independent endogenous variable management is continuous, the average causal response theorem (ACR) and its continuous corollary (Angrist and Imbens 1995) give the path to identify and estimate the causal effect of the intensity of the treatment, the management change driven by the random assignment (programme) in each school, on pupils' learning.

Based on the ACR theorem and its continuous corollary, the IV estimation using a variable treatment intensity produces a weighted average derivative along the length of a possibly nonlinear causal response function (Angrist and Imbens 1995). This study's possibly nonlinear causal response function is the relation between pupils' learning and school management. Also, comparing the CDF of the endogenous variable (treatment intensity) with the instrument turned on and off helps to understand where the action comes from with the instrument. In this context, IV recovers the average derivative over the range of the school management levels where the instrument (random assignment) shifts the CDF of the endogenous variable (management) mostly sharply.

We can assume that our instrumental variable, the random treatment assignment, is independent of any factor (Independence Assumption). It is also assumed that improving school management is the only channel for changing pupils' learning from the instrument (Exclusion Restriction). The programme was delivered only to treatment school managers; therefore, the instrument could only directly affect school managers. Also, school managers can only affect pupils' educational performance through the school's management since school managers do not teach pupils. Consequently, the treatment assignment can only reach pupils' learning through improving the school management, which is driven directly by the school

managers and indirectly by the programme.

Furthermore, since there was a strict confidentiality agreement to protect the identification of the control and treatment schools, it is unlikely that any action due to the random assignment information has been directed to experimental schools apart from the programme provided for treatment schools or the actions that would have been taken regardless of the experiment.

On the one hand, our programme provided the 23 best management practices identified and discussed in the literature to the treatment schools. On the other hand, after two years of the programme, a survey that followed the WMS methodology measured our sample schools' management level regarding the 23 best management practices. As explained before, our fixed pupils' sample considers only pupils registered for grades 1-8 in a sample school for the 2022 school year. Thus, we only consider pupils who enrolled for 2022 during the regular period of enrolment, November and December 2021.

The analysis is done separately by subject: reading and mathematics. Let Y_{isp} and Z_{isp} be the same variables defined when discussing the ATE. Let M_{isp} be the general management level of the school s from the school pair p in which the pupil i registered for the 2022 school year (enrolment occurred in November and December 2021). M_{isp} represents a school's average score regarding the 23 management practices surveyed by our 2023 survey. Let β_{IV} be the causal parameter of interest, i.e., the impact of the treatment intensity (school management changes driven by the random assignment/the programme) on pupils' educational outcomes. β_{IV} can also be interpreted as the causal effect of a one-score point change in the general management of the schools (the scale goes from one, worst management, to five, best management) due to the instrument on pupils' educational outcomes. The parameter, β_{IV} , can be identified from the following Two-Stage Least Square

(2SLS) procedure that uses the random assignment as the instrumental variable (IV):

$$Y_{isp} = \alpha + \beta_{IV} \hat{M}_{isp} + \epsilon_{isp}, \quad (3)$$

where the \hat{M}_{isp} is the fitted M_{isp} from the following first-stage equation

$$M_{isp} = \omega + \nu Z_{isp} + \xi_{isp}.$$

Standard errors are clustered robust standard errors, with clusters being the school pairs formed for the random assignment and the non-inclusion of school pairs' fixed effects in the regressions (Chaisemartin and Ramirez-Cuellar 2024).

We also discuss the programme's impact on management practices and pupils' educational achievements as well as the causal effect of management on pupils' learning with full regression adjustment models using baseline covariates (adjusted models) such as pupils' scores from tests applied before the experiment began and the level of management of the schools measured also before the experiment began. Details in the Appendix I.

IV. RESULTS

IV.A. Treatment Causal Effects

Table III shows the impact of our treatment, the Science and Management for Education Programme, under different specifications. All the outcome variables were generated in December 2023, i.e., two years after the beginning of the programme implementation. The sample includes all pupils (31,760) enrolled in grades 1 to 8 for the 2022 school year in one of the 80 sample schools randomly selected from the Rio de Janeiro population of schools. These pupils registered to a sample school in the regular period of enrolment, November and December 2021,

before the experiment began.

The pupils' reading and mathematics scores (IRT) from the December 2023 Rio Assessments represent their learning in Table III. The scores were standardised to have a mean of zero and a standard deviation of one in each grade and subject regarding the control group. 'Mgmt ([1,5])' represents the general management of each school. It is the simple average of the 23 scores reached by a school regarding the 23 best management practices in our December 2023 survey. For each school, 'Mgmt ([1,5])' goes from one, worst management, to five, best management. 'Mgmt (SD)' is the 'Mgmt ([1,5])' standardised to have a mean of zero and a standard deviation of one regarding the control group. From 31,760 sample pupils, there are 31,412 and 31,405 valid pupils' scores for reading and mathematics, respectively. All 80 sample schools have valid management scores.

No baseline covariates were included in regressions that generated the estimates shown in columns (1) and (2) of Table III. Coefficients shown in columns (3) and (4) come from regressions that included the following centred (demeaned) baseline covariates and their interactions with the treatment variable: pupils' reading and mathematics scores standardised by grade using the control group as the reference from two different Rio Assessments applied in September and December 2021 when the outcome variable is reading or mathematics, and the schools' average management score from our 2021 management survey when the outcome variable is the management of the schools. Standard errors are reported in parentheses. Clustered robust standard errors with clusters set to be the pairs used for the random assignment procedure are used when the outcome variable was reading or mathematics. Standard errors are robust but not clustered when the outcome variable is the management level of the schools. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table III: The Treatment Causal Effects

	No covariates		With covariates	
	(1)	(2)	(3)	(4)
	ATE OLS	Mgmt ([1,5]) IV	ATE RA	Mgmt ([1,5]) IV/RA
Reading (SD)	0.226*** (0.059)	0.680*** (0.245)	0.187*** (0.045)	0.564*** (0.182)
Maths (SD)	0.237*** (0.059)	0.714*** (0.265)	0.208*** (0.049)	0.627*** (0.212)
Mgmt (SD)	0.916*** (0.268)	–	0.928*** (0.260)	–

Note: The table shows the impact of the treatment, SMEP, under different specifications. All the outcome variables were generated in December 2023, two years after the beginning of the treatment. The sample includes all pupils (31,760) enrolled in grades 1 to 8 at the beginning of 2022 in one of the 80 sample schools randomly selected from the Rio de Janeiro population of schools. These pupils registered to a sample school in the regular period of enrolment, November and December 2021, before the experiment began. The pupils’ reading and mathematics scores (TRI) from the December 2023 Rio Assessment represent their learning. The scores were standardised to have a mean of zero and a standard deviation of one in each grade and subject regarding the control group. ‘Mgmt ([1,5])’ represents the schools’ average management score that goes from one, worst management, to five, the best management. ‘Mgmt ([1,5])’ comes from a survey conducted in December 2023 that adopted the World Management Survey (WMS) methodology. ‘Mgmt (SD)’ is the ‘Mgmt ([1,5])’ standardised to have a mean of zero and a standard deviation of one regarding the control group. From 31,760 sample pupils, there are 31,412 and 31,405 valid pupils’ scores for reading and mathematics, respectively. Columns (1) and (3) report average treatment effects (ATE) estimates from regressions without and with baseline covariates, respectively. Columns (2) and (4) show IV estimates generated from regressions without and with baseline covariates, respectively. IV estimates can be interpreted as the causal effects of a one-score point change (on a management scale that goes from one, worst management, to five, best management) in school management, driven by the random assignment/treatment, on pupils’ educational outcomes. Columns (3) and (4) report ATE and IV estimates from regressions with the following centred baseline covariates (and their interaction with treatment): pupils’ reading and mathematics scores from two different Rio Assessments that were applied in September and December 2021, and schools’ average management score from our 2021 management survey. Standard errors are reported in parentheses. They are clustered robust standard errors when the outcome variable is reading or mathematics. The clusters for these standard errors are the school pairs used for the random assignment procedure. Standard errors are robust but not clustered when the outcome variable is the management level of the schools. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Column (1) of Table III reports the average treatment effect (ATE) on pupils' learning and the management level of the schools. The ATE estimates are 0.226 standard deviations (SD) for reading and 0.238 SD for mathematics, both significant at the 1% level. Furthermore, the ATE estimate is 0.916 SD for school management with significance at the 1% level. Column (3) shows slightly lower ATE estimates from regressions with centred baseline covariates and their interaction with the treatment.

The estimated causal effects of the improvement in the management of the schools, generated by the random assignment, on pupils' learning are shown in column (2) of Table III. The IV estimates of the impact of a one-score point change (on a scale from one to five) in school management, driven by the treatment, are 0.680 SD for reading and 0.714 SD for mathematics, both significant at the 1% level. Column (4) reports slightly lower IV estimates from regressions with centred baseline covariates and their interaction with the treatment.

The treatment also had a statistically significant impact on pupils' absenteeism reduction. This analysis is based on regressions of the pupils' absence rate on a treatment dummy. We used clustered robust standard errors with the clusters defined as the school pairs used for random assignment. Pupils from treatment schools were 19% less absent than pupils from control schools in 2023 ²⁰. In 2022, treatment pupils were 18% less absent than control pupils. In 2021, treatment and control pupils had almost the same absence rate. The difference was 0.2% and not statistically significant. It is also relevant to mention that there were no significant differences between control and treatment schools regarding pupils who abandoned school (around 0.37% of pupils abandoned school in 2023 in both groups).

Section III.C shows that a typical pupil in a Rio de Janeiro school acquired, on

²⁰The absence rate was 9.16% for control schools and 7.5% for treatment schools

average, 13.14 score points in reading and 14.73 in mathematics by grade (and year) based on the December 2023 Rio Assessments. This analysis used control schools to represent Rio de Janeiro schools. Furthermore, considering the December 2023 Rio Assessments (control pupils), the average standard deviation by grade was 45.30 for reading and 42.50 for mathematics.²¹ It is possible to transform the average score acquired by grade in standard deviations (SD) by simply dividing the average score acquired by grade by the average standard deviation across grades.

Thus, each year of schooling or each grade in Rio de Janeiro schools gives children, on average, a 0.290 (13.14/45.30) SD of learning in reading and a 0.343 (14.73/42.50) SD of learning in mathematics. We need only to divide ‘the causal effects estimates (SD)’ by ‘the average learning acquired by grade (SD)’ to translate our impact estimates from SD to years of learning.

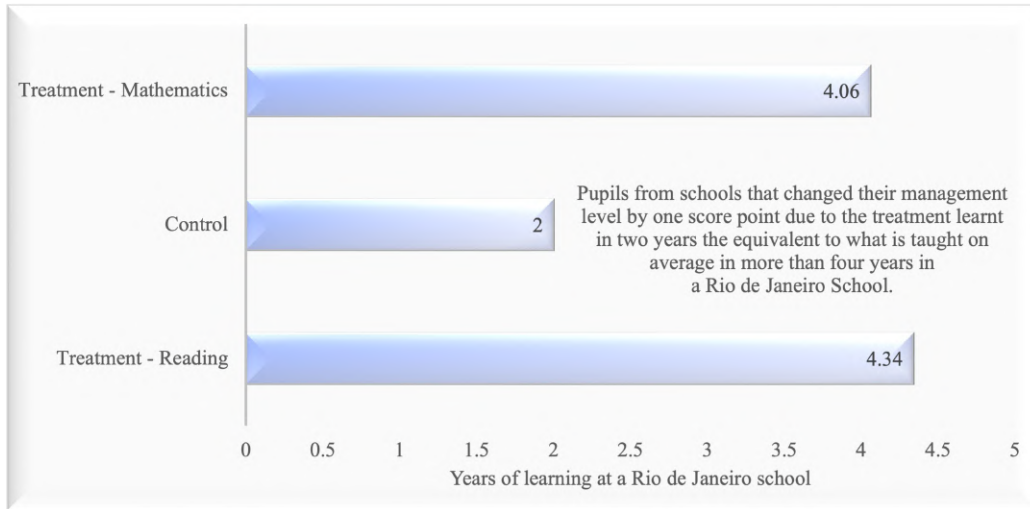
The ATE estimates in terms of years of learning (in a Rio de Janeiro school) are $\frac{0.226SD}{0.29SD} = 0.78$ years for reading and $\frac{0.237SD}{0.347SD} = 0.68$ years for mathematics. Pupils in treatment schools are around 3/4 of a school year ahead of pupils from schools that did not receive the treatment regarding learning acquired in reading and mathematics. Two years in a treatment school gave pupils the learning in reading and mathematics equivalent to what is learnt by pupils in 2 years and 3/4 of a school year in a Rio de Janeiro school. Treatment schools are 39% more productive in providing learning in reading and 34% in mathematics than non-treatment schools.

Figure V shows the IV estimates of the impact of the high implementation schools - schools that changed their management level by one score point (on a scale from one, worst management, to five, best management) due to the treatment - on pupils’ educational performance. The ‘Control’ bar represents the learning

²¹Firstly, we calculated the standard deviation within each grade by subject. Secondly, we averaged these standard deviations regarding the grades within each subject.

acquired on average by pupils in two years at a Rio de Janeiro school. The IV estimates are $\frac{0.680SD}{0.290SD} = 2.34$ years of learning for reading and $\frac{0.714SD}{0.347SD} = 2.06$ years of learning for mathematics.

Figure V: Management Impact (IV) in Terms of Years of Learning



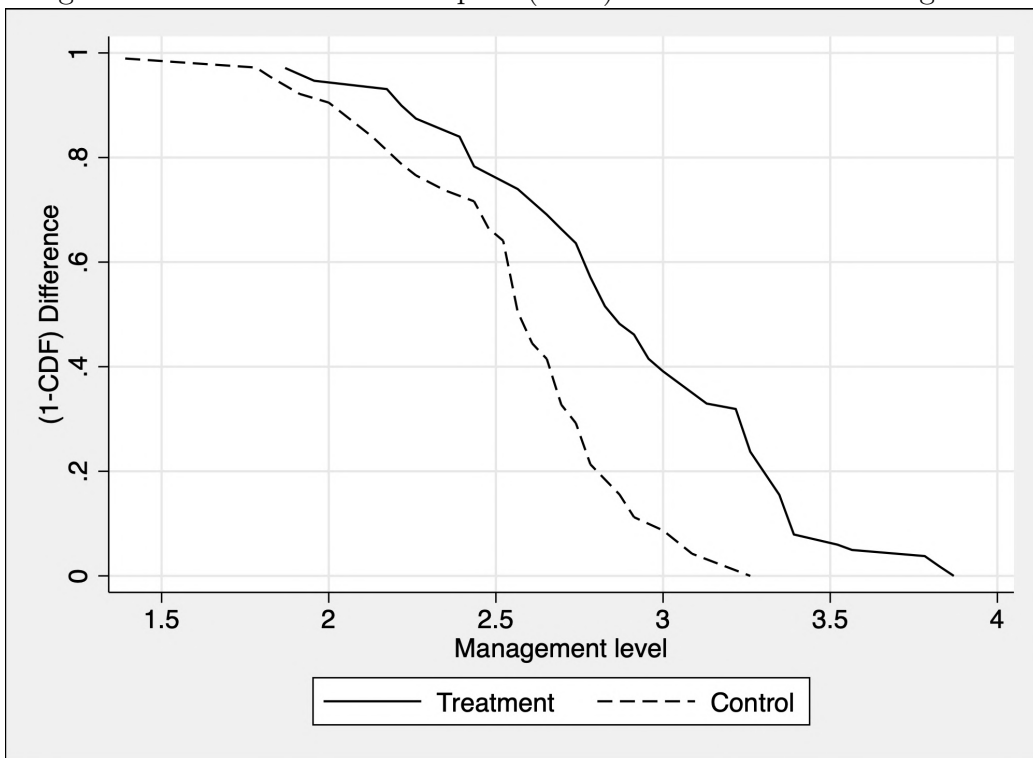
Notes: The figure shows the IV estimates of the impact of the high implementation schools - schools that changed their management level by one score point (on a scale from one, worst management, to five, best management) due to the treatment - on pupils' educational performance. The 'Control' bar represents the learning acquired on average by pupils in two years at a Rio de Janeiro school. Pupils in high implementation schools are more than two academic years of learning (reading and mathematics) ahead of pupils from schools that did not receive the treatment. Two years in a high implementation school taught pupils reading and mathematics equivalent to what is learnt in more than four years in a Rio de Janeiro school. Treatment schools were 117% more productive in providing learning in reading and 103% in mathematics than non-treatment schools.

Pupils in high implementation schools are more than two academic years of learning (reading and mathematics) ahead of pupils from schools that did not receive the treatment. Two years in a high implementation school taught pupils reading and mathematics equivalent to what is learnt in more than four years in a Rio de Janeiro school. Treatment schools were 117% more productive in providing learning in reading and 103% in mathematics than non-treatment schools.

Figure VI shows the effect of the instrument variable (random assignment)

on the schools' management levels. The figure illustrates the instrument-induced difference between treatment and control schools in the probability of a school having its management level regarding the 23 best management practices greater or equal to a specific management level on the x-axis (i.e., one minus the CDF). Although the management level scale can go from one, worst management, to five, best management, the x-axis of Figure VI shows only the scale range where there are observations to make the graph better to visualise.

Figure VI: The Treatment's Impact (ATE) on the Schools' Management



Notes: The effect of the instrument variable (random assignment) on the schools' management levels. The figure illustrates the instrument-induced difference between treatment and control schools in the probability of a school having its management level regarding the 23 best management practices greater or equal to a specific management level on the x-axis (i.e., one minus the CDF). Although the management level scale can go from one, worst management, to five, best management, the x-axis shows only the scale range where there are observations to make the graph better to visualise. The solid line represents treatment schools, and the dashed lines control schools.

The solid line in Figure VI represents treatment schools, and the dashed lines control schools. There are differences across the entire range of the x-axis (schools' management levels). For example, while the chance that a control school had a management level greater than or equal to 3 is around 10%, the chance for a treatment school is greater than 40%.

The results in Table III and the differences reported in Figure VI show that the programme successfully changed the overall management of the treatment schools compared to the control schools. Since the 23 best management practices can be grouped into five groups: target setting, leadership, operations, monitoring and people management, as discussed in Section II.B, it is valuable to analyse how the treatment impacted each of them.

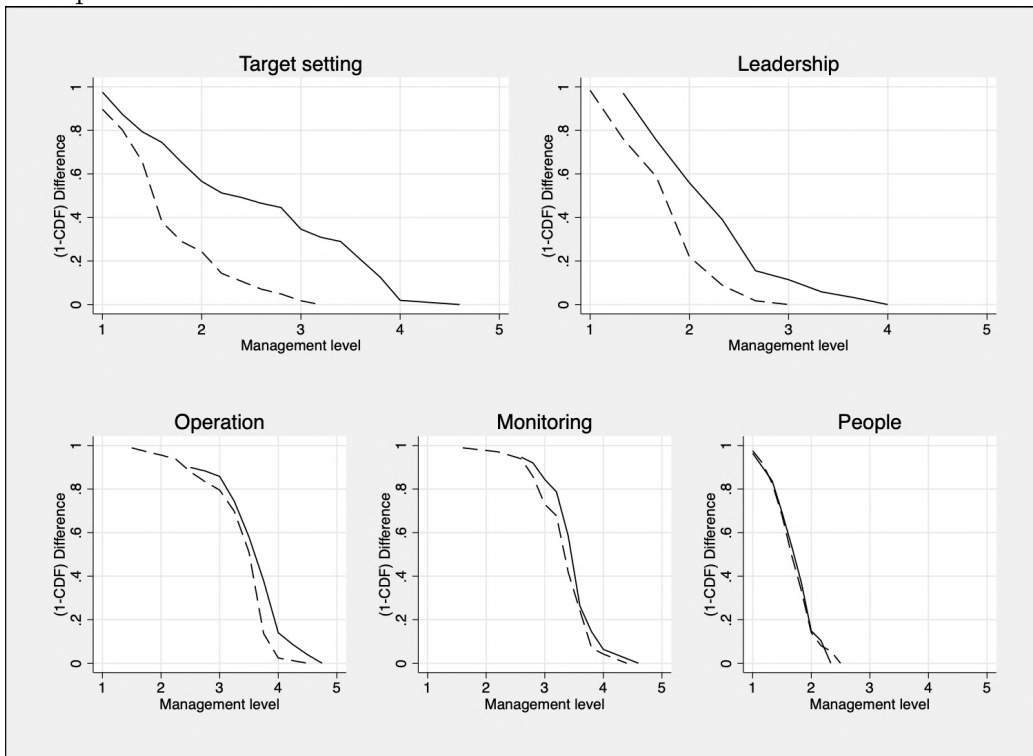
As discussed in Section III.B, for each school, we generated an average score regarding each group of practices by simply averaging the scores of the practices that belong to each group. Segregating the general management into groups of management practices helps to understand how the impact of the treatment is distributed across the five groups.

Figure VII shows the impact of the treatment (instrument) across the five groups of management practices. The figure reports the instrument-induced difference between treatment and control schools in the probability that a specific group of management practices of a school is greater or equal to a particular management level on the x-axis (i.e., one minus the CDF²²). The management level scale on the x-axis goes from one, worst management, to five, best management. The solid line represents treatment schools, and the dashed lines control schools.

Figure VII shows that most of the impact driven by the treatment on the schools' general management is concentrated in the target setting and leadership

²²Weighted CDF. The weights used are the number of grades 1-8 pupils in each school, as discussed in Section III.E.

Figure VII: The Treatment's Impact (ATE) on the Schools' Management Groups of Practices



Notes: The figure illustrates the instrument-induced difference between treatment and control schools regarding the probability that a specific group of management practices of a school is greater or equal to a specific management level on the x-axis (i.e., one minus the CDF). The effect of the instrument variable (random assignment) on each group of management practices (target setting, leadership, operation, monitoring and people management) can be seen through the difference between control and treatment schools CDFs (one minus CDF). The management level scale on the x-axis goes from one, worst management, to five, best management. The solid line represents treatment schools, and the dashed lines control schools.

groups. For example, while the probability that a control school had a target setting group of management practices level greater than or equal to 3 is zero, the treatment school chance is around 35%. Moreover, while the chance that a control school had its leadership practices average score greater than or equal to 2 is 20%, the treatment school chance is around 60%. The differences between operation and monitoring groups are smaller. For instance, while the probability that a control school had an operation group level greater than or equal to 3.75 is 20% in Figure VII, the treatment school chance is around 40%.

Yet, while the chance that a control school had an operation average score greater than or equal to 3 is around 70%, the treatment school chance is around 85%. However, the differences between treatment and control schools regarding operation and monitoring groups are much smaller than those encountered with the management practices' target setting and leadership groups. Figure VII reports no differences between control and treatment schools regarding the people management group of practices.

It is relevant to remember that treatment schools' CDF was slightly behind the control schools' CDF regarding operation and monitoring group of practices in 2021, according to Figure III. In 2023, treatment schools' CDF was ahead of the control schools' CDF regarding operation and monitoring group of practices. Thus, we can consider the difference between treatment and control schools regarding operation and monitoring presented in Figure VII to be a little larger if we look at the differences in 2021 shown in Figure III. More details about the impact of the treatment on the groups of management practices can be found in Appendix J.

So far, we have discussed the schools' management, grouping all the 23 best practices or grouping them into five groups (averaging the practices by group); we also want to know if the programme successfully provided schools with the 23 best

management practices as planned. Moreover, it is relevant to understand how the treatment’s impact spread across the 23 practices.

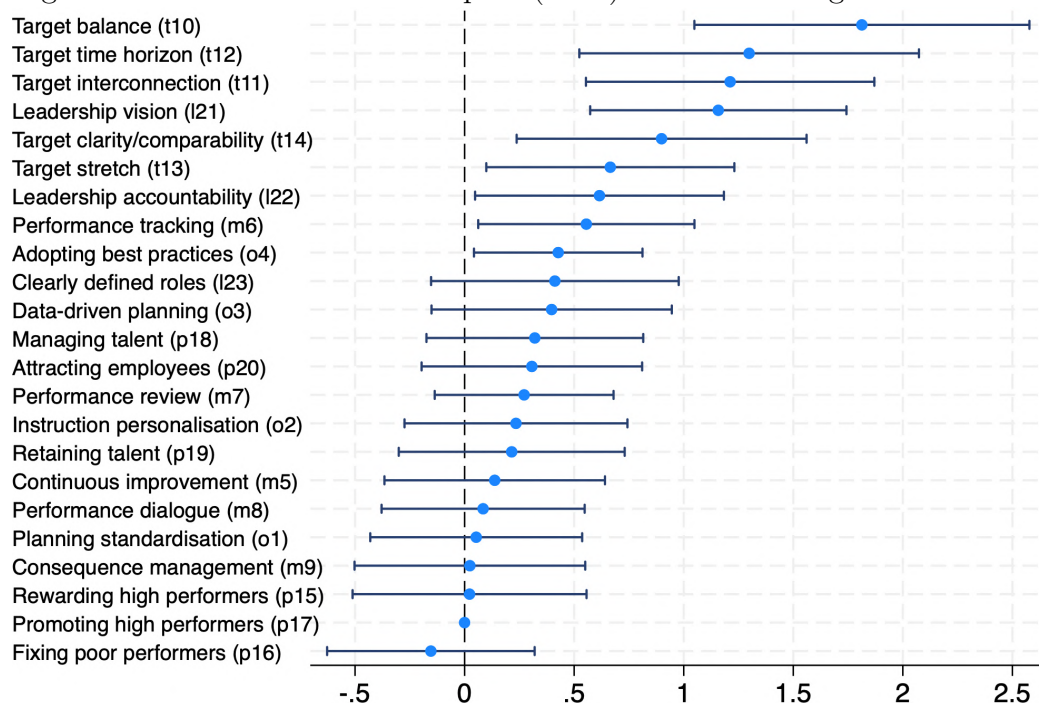
Figure VIII shows the programme’s impact on each of the 23 best management practices. The figure displays the ATE point estimates and their 95% confidence intervals. Each point estimate and confidence interval comes from a weighted regression at the school level of one of the 23 management practices on a treatment dummy that turns one when the school was randomly assigned to treatment and zero otherwise, as discussed in Section III.E. Robust standard errors were used. Each of the 23 management practices used as outcome variables in the regressions was standardised to have a mean of zero and a standard deviation of one regarding the control group. Thus, the x-axis scale is in standard deviations (SD).²³

Figure VIII reports that the treatment positively changed the management of the schools across almost all the 23 best management practices. This indicates that the programme’s goal of providing schools with the 23 best management practices was achieved. It is possible to note that the treatment effects were heterogeneous across the practices.

Figure VIII shows the impact of the treatment was statistically significant at the 5% level and larger than one standard deviation (SD) in four practices: Target balance (T10), Target time horizon (T12), Target interconnection (T11), and Leadership vision (L21). The ATE estimates were statistically significant at the 5% level and between 0.4 and 0.9 SD on five practices: Target clarity/comparability (T14), Target stretch (T13), Leadership accountability (L22), Performance tracking (M6) and Adopting best practices (O4). Although not statistically significant, the treatment impact was positive on the following ten practices: Clearly defined roles (L23), Data-driven planning (O3), Managing talent (P18), Attracting

²³The weights used are the number of grades 1-8 pupils in each school, as discussed in Section III.E

Figure VIII: The Treatment’s Impact (ATE) on Each Management Practice



Notes: The figure shows the programme’s impact (ATE) on each of the 23 best management practices. The figure displays the ATE point estimates and their 95% confidence intervals. Each point estimate and confidence interval comes from an unadjusted regression (school level) of one of the 23 management practices on a treatment dummy that turns one when the school was randomly assigned to treatment and zero otherwise, as discussed in Section III.E. Robust standard errors were used. Each of the 23 management practices used as outcome variables in the regressions was standardised to have a mean of zero and a standard deviation of one regarding the control group. Thus, the x-axis scale is in standard deviations (SD).

employees (P20), Performance review (M7), Instruction personalisation (O2), Retaining talent (P19), Continuous improvement (M5), Performance dialogue (M8) and Planning standardisation (O1). Four practices did not change due to the treatment: Consequence management (M9), Rewarding high performers (P15), Promoting high performers (P17), and Fixing poor performers (P16).

The results shown in Figure VIII from a more detailed analysis of the 23 management practices confirm the results from the grouped analysis reported in Figure VII. However, although Figure VII and Table A8 show no difference between

treatment and control schools regarding the people management group of practices, Figure VIII shows differences (not statistically significant) between control and treatment schools regarding Managing talent (P18), Attracting employees (P20) and Retaining talent (P19) management practices.

IV.B. Heterogeneous Treatment Effects

In this section, we want to investigate the heterogeneous impact of the treatment on pupils' learning across pupils' gender, pupils' race, pupils' families receiving Bolsa Familia, pupils' grades (1 to 8) in 2022, school segment, school size (numbers of pupils), school management in 2021 and the quartiles of pupils' educational outcomes in the September 2021 Rio Assessments. Each of these subgroups was analysed separately with a regression that included the centred covariate representing the subgroup and its interaction with the treatment dummy. The regressions used clustered robust standard errors with clusters defined as the school pairs formed for the random assignment. The outcome variable was generated by averaging the reading and mathematics standardised scores from the December 2023 Rio Assessments. Scores were standardised to have a mean of zero and a standard deviation of one in each subject, grade, and test application date using the control group as the reference. The pupils' performance quartiles were defined using the average of the reading and mathematics standardised scores from the September 2021 Rio Assessments.

The treatment impact differences were very small and not statistically significant between male and female pupils; black, brown, or white pupils; pupils from families receiving Bolsa Familia or not; pupils from different grades; and pupils from different 2021 educational performance quartiles. The only difference worth noting is that pupils from schools with better management in 2021, based on our

2021 survey, seem to benefit more from the programme. The impact of the treatment on pupils' learning was 0.174 standard deviations (SD) larger for pupils from schools with one SD more management in 2021.

IV.C. Robustness Checks

Table A9 in Appendix K shows the causal effects of the treatment on pupils learning under different specifications; however, we replace the outcome variable representing pupils' learning, Rio Assessments scores based on IRT, with Rio Assessments scores from the Classical Test Theory (CTT). The CTT scores are simply the rate or proportion of correct answers. The results based on IRT scores from the Rio Assessments reported in Table III are similar to the estimates reported in Table A9 that are based on CTT scores from the same Rio Assessments. The ATE estimates from unadjusted regressions are 0.209 standard deviations (SD) for reading and 0.220 for mathematics, and the estimates from adjusted regressions are 0.173 SD for reading and 0.192 SD for mathematics. The IV estimates from unadjusted regressions are 0.630 SD for reading and 0.662 SD for mathematics, and the estimates from adjusted regressions are 0.522 SD for reading and 0.583 SD for mathematics. All the results are statistically significant at the 1% level.

Our primary analysis uses data at the pupil level and adjusts the standard errors to consider clusters formed by the school pairs used for the random assignment (Chaisemartin and Ramirez-Cuellar 2024). The goal is to test our results' robustness with a school-level analysis. Reading and mathematics pupils' December 2023 Rio Assessments scores were standardised by grade and subject using the control group as the reference. Each school's educational outcome is represented by the average of the standardised scores of its pupils in reading and mathematics. We used the number of pupils in the sample as weights in our model since the

schools have varying sizes, and we want to estimate the impact of the programme (ATE) on pupils' educational outcomes (Kahan et al. 2023). The standard errors are robust standard errors. Thus, for this analysis, 80 observations represent the 80 sample schools.

Using the school as the unit of analysis, the estimates of the impact of the treatment on pupils' educational performance (ATE) are 0.224(0.063) standard deviations (SD) for reading and 0.235(0.064) SD for mathematics. The values in parentheses represent the standard errors. The IV estimates of the impact of a change of one score point in the management of the schools due to the treatment are 0.674 SD for reading and 0.708 SD for mathematics. The estimates are significant at the 1% level. Unsurprisingly, the estimates are almost identical to the ATE and IV estimates conducted at the pupil level shown in Table III.

For robustness checks, we also used Anderson (2008) to compute sharpened False Discovery Rate (FDR) q-values to deal with the multiple hypothesis testing issues. The FDR is the expected proportion of rejections that are type I errors (false rejections). Although the method does not account for correlations between p-values, Anderson (2008) shows through simulations that the method works well if the p-values are positively correlated, as in this study. For example, if the treatment improves pupils' reading scores, then we can think that it is likely to improve their mathematics outcomes, or if the treatment improves Target balance (t10) management practice, it is likely to improve Target interconnection (t11).

Table A10 in Appendix K shows that all statistically significant results presented in this study remain statistically significant under sharpened False Discovery Rate (FDR) q-values computation.

As the last robustness check, we tried to falsify the Exclusion Restriction assumption used as the IV estimation basis. The exclusion restriction required for

a causal understanding of the IV estimates claims that random assignment (IV) can affect pupils' educational outcomes only by changing the management of the schools. Although we can not directly test this assumption, we can provide evidence of its validity using a subgroup less affected by the instrument (random assignment) as follows.

According to the heterogeneity analysis conducted in section IV.B, the lower the level of management of schools in 2021, the smaller the impact of the instrument on the management level of schools in 2023. A regression analysis (first stage) of the 2023 school management on the instrument, as discussed in section III.E, but limited to the subgroup of schools that had a management level in 2021 less than 2, reveals that there was no impact of the instrument on the management of these schools in 2023. A null first stage.

A regression analysis of pupils' educational outcomes on the instrument, as discussed in section III.E, but again limited to the subgroup of schools with a management level in 2021 less than 2, shows a null result. Thus, a null reduced form. Consequently, this no-first-stage sample does not provide any signal of violation of the exclusion restriction.

V. DISCUSSION AND CONCLUSION

This paper provides experimental evidence on the power of management to drive pupils' learning. The Science and Management for Education Programme had an average treatment effect (ATE) of 0.226 (0.059) standard deviations (SD) for reading and 0.237 (0.059) SD for mathematics. The impact estimates were consistent across subjects, which is not a common fact in the related academic literature. The estimates were statistically significant at the 1% level.

Since the outcome variable measured pupils from first to ninth grade on the

same scale of competencies and scores, it was possible to translate the standard deviation estimates to years of learning. The impact estimates mean that pupils who received treatment were $\frac{3}{4}$ of an academic year ahead in reading and mathematics compared to those who did not. Two years in a treatment school taught pupils reading and mathematics equivalent to what they would have learned in almost three years ($2\frac{3}{4}$) in a Rio de Janeiro school. Thus, treatment schools were 37.5% more productive in providing reading and mathematics learning to their pupils than non-treated schools.

The treatment also had a statistically significant impact on pupils' absenteeism. Pupils from treatment schools were 19% less absent than pupils from control schools in 2023. In 2022, treatment pupils were 18% less absent than control pupils. The differences between the control and treatment groups in 2021 were almost null. It is also relevant to mention that there were no differences between control and treatment schools regarding pupils' abandonment rate (around 0.37% of pupils abandoned school in 2023 in both groups).

Importantly, the differences in the programme's impact were very small and not statistically significant between male and female pupils; black, brown, or white pupils; pupils from families receiving Bolsa Familia or not; pupils from different grades; and pupils from different 2021 educational performance quartiles. The treatment did not increase inequalities across the mentioned groups

The programme's impact is likely to have affected not only the learning of reading and mathematics but also the other subjects taught to pupils in treatment schools. This is because the programme did not focus on specific subjects such as reading and mathematics but on improving the management of the schools regarding the 23 best management practices. Reading and mathematics were the only subjects with standardised tests available to be used to assess the programme's

effect. Reading and mathematics subjects were used in this study to represent the overall learning of pupils across disciplines. Better-managed schools are likely to be more productive in providing pupils with learning across all subjects.

We also applied two management surveys to the 80 sample schools before and after the experiment to precisely measure the management of the schools regarding 23 best management practices extensively discussed in the literature. The surveys adopted the World Management Survey (WMS) methodology developed by Bloom and Van Reenen (2007, 2010), Bloom et al. (2012b), and Bloom et al. (2015). The estimate of the programme's causal effect (ATE) on the school's general management was 0.916 (0.268) standard deviations (SD). This estimate is statistically significant at the 1% level. Furthermore, a segregated analysis of the practices reveals that the treatment positively impacted most of the 23 'best' managerial practices. This fact confirms the programme's success in achieving its goal of improving treatment school management regarding the 'best' management practices. However, the effects were heterogeneous across the practices.

The impact of treatment (ATE) was 1.530 standard deviations (SD) for the target-setting group of practices that aggregates Target balance (T10), Target time horizon (T12), Target interconnection (T11), Target clarity/comparability (T14) and Target stretch (T13), and 1.069 SD for the leadership group of management practices that is formed by Leadership vision (L21), Leadership accountability (L22) and Clearly defined roles (L23). Both estimates are statistically significant at the 1% level.

The programme's causal effect (ATE) was 0.370 SD for the operation group of management practices, which includes Adopting best practices (O4), Data-driven planning (O3), Instruction personalisation (O2), and Planning standardisation (O1). The impact was 0.324 SD for the monitoring group of management prac-

tices, which is formed by Performance tracking (M6), Performance review (M7), Continuous improvement (M5), Performance dialogue (M8) and Consequence management (M9). Both estimates were not statistically significant.

Although there were no differences between treatment and control schools when we analysed the people management practices as a group, the analyses conducted separately by practice that form this group reveal a slightly different frame. The programme did not impact three of the six practices that form the people management group: Fixing poor performers (P16), Promoting high performers (P17), or Rewarding high performers (P15). This was somewhat expected since public school teachers are civil servants under Brazil's public administration, and it is very challenging under this legislation to remove or dismiss poor performers and promote or reward high performers. The other practices that form the people management group - Managing talent (P18), Attracting employees (P20) and Retaining talent (P19) management practices - were positively impacted by the treatment. However, the estimates were not statistically significant.

The treatment had a much higher impact on the target setting and leadership management practices than on operation and monitoring practices (people management practices were discussed in the previous paragraph). One possible explanation for the instrument's heterogeneous effect across the management practices is as follows. On the one hand, schools had poor levels of target setting and leadership management practices implemented in 2021, as discussed in Section III.D, which facilitated the programme's impact. On the other hand, schools had better implementation levels of operation and monitoring management practices in 2021, which made it more difficult for the treatment to make a larger difference.

We use the random assignment to treatment as the instrumental variable to disentangle the causal effects of management on pupils' educational outcomes.

Management is the causal endogenous variable of the treatment intensity analysis. The causal effect of a change of one score point in the schools' management due to the treatment is 0.680 (0.245) standard deviation (SD) for reading and 0.714 SD for mathematics. If we put management in standard deviations (SD), a one SD change in school management because the treatment had a causal effect of around 0.24 SD in reading and mathematics. The estimates were statistically significant at the 1% level. It is the first study to demonstrate the impact of management, defined as the 23 WMS best management practices, on pupils' learning. The findings have shown a noticeable similarity to the correlation estimates found by Bloom et al. (2015).

Pupils in high implementation schools - schools that improved their management level by one score point due to the treatment - were found to be more than two years of learning ahead of pupils from schools that did not receive the treatment. Pupils in these high implementation schools achieved reading and mathematics proficiency equivalent to what is typically achieved in more than four years at a typical Rio de Janeiro school. High implementation schools were 117% more productive in providing reading learning and 102% more productive in providing mathematics learning to their pupils than non-treated schools. The treatment intensity analysis has shown the larger the management improvement due to the treatment, the larger the impact on pupils' learning.

Let's consider the programme's impact on pupils' educational outcomes in a broader context. According to Kraft (2020), based on the analysis of 1,942 effect sizes from 747 experiments that evaluated educational interventions with standardised test outcomes, an impact of 0.15 or even 0.10 SD should be deemed significant and impressive when such impacts arise from large-scale (> 2,000 pupils) field RCTs. Our findings - IV (High Implementation) and ATE estimates, both

statically significant at the 1% level - show causal effects among the largest in the education intervention literature according to Kraft (2020).

The literature based on causal methodologies provides conflicting and often unclear evidence on the causal effects of management on productivity. On the one hand, Abdulkadiroğlu et al. (2011), Angrist et al. (2010, 2012), Barros et al. (2019, 2021), Beg, Fitzpatrick, and Lucas (2023), Bloom et al. (2020, 2012a), Bruhn, Karlan, and Schoar (2018), Curto and Fryer (2014), Dobbie and Fryer (2011), Fryer (2017), Fryer (2014), Gosnell, List, and Metcalfe (2020), and Tavares (2015) discussed positive effects of management on productivity. We have discussed the ‘pros’ and ‘cons’ of this ‘positive’ literature in Appendix L. On the other hand, Hoyos, Ganimian, and Holland (2019), Muralidharan and Singh (2020), and Romero et al. (2022) presented findings on the null effect of management on productivity. We have also discussed the null results evidence in Appendix L without considering our findings. However, it is essential to explore the null effect literature further in light of our results.

More than bringing strong evidence in favour of the positive causal relationship between management and productivity, our findings clarify the issues with the trials that showed no effect of management on school productivity. One first question is that management information or training *per se* programmes may not change processes within schools and, consequently, the school production (pupils’ learning), as shown by the large experiment conducted by Muralidharan and Singh (2020). Our intervention was successful because the programme made sure that the procedures within the schools were changing through our on-the-job training and one-to-one coaching with managers.

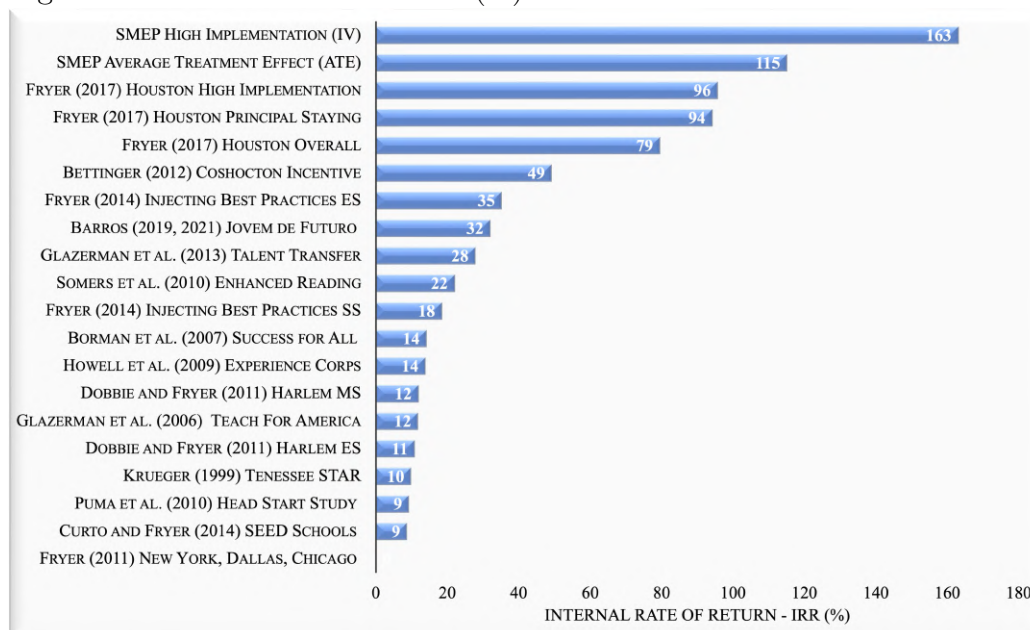
The second issue in the null effects literature is related to the size of the change needed in management to affect productivity. Our study shows that it was neces-

sary to increase management by 0.916 standard deviations (SD) to improve pupils' educational performance in reading and mathematics by 0.226 SD and 0.237 SD, respectively. A large experiment conducted in Mexico by Romero et al. (2022) showed that a change of 0.13 SD on management had no meaningful impact on pupils' test scores. Based on our results, 0.13 SD of impact size on management seems insufficient to affect pupils' test scores. Similarly, Hoyos, Ganimian, and Holland (2019) reported no impact of some support in school management on pupils' educational achievements. Again, the 'management' delivered to schools was one visit per year and two workshops, which seems very little to improve the management of the schools. The third point to explain the null effect literature on the impact of management on productivity is that not all management practices affect productivity, and even a practice with the potential to affect productivity may fail if implemented at a low level. Our programme not only successfully delivered the right managerial practices to affect pupils' learning but also implemented these practices at a level capable of impacting pupils' educational performance.

Including the USD (PPP GDP) 15.22 per year per pupil programme cost in the discussion, we can say that the Science and Management for Education Programme is highly cost-beneficial and cost-effective compared to other important interventions in the literature. Figure IX shows the internal rates of return (IRR) for our treatment, 'SMEP Average Treatment Effect (ATE)', the high implementation of our treatment, 'SMEP High Implementation (IV)', the Brazilian programme *Jovem de Futuro* discussed in Barros et al. (2019, 2021) and 17 important causal studies discussed in Fryer (2017, 2016). The IRR is the discount rate that equates the cost of an intervention with the present value of the expected future earnings inflows that the intervention generates. We calculated the IRRs following the methodology from Krueger (2003). More details are in Appendix F.

Figure IX shows that our experiment’s internal rates of return (IRR) are larger than other relevant interventions: 163% for the ‘SMEP High Implementation (IV)’ and 115 % for the ‘SMEP Average Treatment Effect (ATE)’. The present value of the future Brazilian earnings inflows - that are expected to be generated by the learning increase caused by our treatment (ATE) - is around USD (PPP GDP) 133,086.24 (R\$ 343,362.19 where R\$ means Brazilian Reais) for a typical Rio de Janeiro pupil.

Figure IX: Internal rates of return (%) for relevant interventions in education



Notes: The figure shows the internal rates of return (IRR) for our treatment, ‘SMEP Average Treatment Effect (ATE)’, the high implementation of our treatment, ‘SMEP High Implementation (IV)’, the Brazilian programme Jovem de Futuro discussed in Barros et al. (2019, 2021) and 17 important causal studies discussed in Fryer (2017, 2016). The IRR is the discount rate that equates the cost of an intervention with the present value of the expected future earnings inflows that the intervention generates. Apart from our experiment and the Jovem de Futuro Programme, all the other studies’ IRRs come from Fryer (2017). We calculated the IRR for our Science and Management for Education Programme (SMEP) and the Jovem de Futuro Programme following the methodology from Krueger (2003).

Since the average yearly earnings of a typical Brazilian in 2022 was around

USD (PPP) 13,953.50 (R\$ 36,000.00)²⁴, the expected gains due to the learning increase caused by the programme for a typical pupil from a Rio de Janeiro school affected by the programme are equivalent to almost ten years ($\frac{343,362.19}{36,000.00} = 9.54$) of what would be their future earnings without the programme.

Furthermore, considering the present value of the future earnings expected to be generated by the learning increase caused by the treatment for all 15,626 pupils who benefited from the programme, the programme's social impact is USD (PPP GDP) 2 billion (R\$ 5.36 billion).

The present value of the expected future Brazilian earnings inflows that are expected to be generated by the learning increase caused by the high implementation of the treatment is around USD (PPP GDP) 400,625.58 (R\$ 1,033,794.60, where R\$ means Brazilian Reais) for a typical Rio de Janeiro pupil. Based on the average earnings of Brazilians in a year, the gains expected due to the learning increase caused by the programme for a typical pupil from a high implementation school are equivalent to almost 29 years ($\frac{1,033,794.60}{36,000.00} = 28.7$) of what would be their future earnings without the programme.

Table A7 in Appendix F helps to understand why the Science and Management for Education Programme (SMEP) has the highest internal rate of return among the interventions discussed until now. Our programme has one of the lowest per-pupil costs and the highest impact on pupils' learning (sum of the impact in reading and mathematics). For instance, the SEED Schools intervention had a large impact of 0.42 SD; however, accompanied by a very high cost per pupil per year of USD 51,904.45. The SEED cost per pupil per year is 3,410 ($\frac{51,904.45}{15.22}$) times higher than the SMEP cost per pupil per year of USD (PPP) 15.22.

We also followed the J-PAL approach presented in Dhaliwal et al. (2013) to cal-

²⁴Information from IPEA/PNAD

culate the cost-effectiveness of our treatment. More details on the J-PAL approach are in Appendix G. The high implementation of our programme (IV estimates) achieves a cost-effectiveness of 15.61 ‘Additional SD per \$100’. Based on the average treatment effects (ATE) estimates, the cost-effectiveness of our treatment reaches 1.95 ‘Additional SD per \$100’.

Table IV shows the cost-effectiveness of our programme, ‘SMEP (ATE), Brazil’, the high implementation of our programme, ‘SMEP High Implementation (IV), Brazil’, and 30 randomised interventions on education analysed by J-PAL (2020). Differently from Dhaliwal et al. (2013) and J-PAL (2020) that used a cut off of 10% of significance level to select experiments to have their cost-effectiveness calculated, we only show in Table IV the ‘Additional SD per \$ 100’ for the randomised experiments that are significant at the 5% level. Apart from the non-significant interventions, the statistically significant programmes at the 5% level are organised in descending order of their cost-effectiveness measure, i.e., ‘Additional SD per \$100’. In Table IV, ‘Average Impact (SD)’ represents the average impact of a programme on pupils’ test scores across subjects (if more than one subject was used) in standard deviations (SD). All costs are in USD (PPP GDP) of 2011.

Table IV shows the ‘SMEP High Implementation (IV)’ is the third most cost-effective intervention among the programmes assessed based on the ‘Additional SD per \$100 (PPP)’. ‘SMEP High Implementation (IV)’ represents the programme’s impact on the high implementation schools (schools that changed their management level by one score point on a scale from one, worst management, to five, best management). In this case, the intervention cost-effectiveness achieves a striking 15.61 ‘Additional SD per \$100 (PPP)’, showing the power of management to improve pupils’ learning and school productivity.

Table IV: Cost-Effectiveness Analysis (CEA)

Programmes	Average Impact SD	95% Lower Bound	95% Upper Bound	Additional SD per \$100 (PPP)
Linking school to local govt, Indonesia	0.165	0.034	0.296	26.10
Streaming by achievement, Kenya	0.176	0.025	0.327	16.44
SMEP High Implementation (IV), Brazil	0.700	0.210	1.190	15.61
Electing school & linking local govt, Indonesia	0.216	0.034	0.398	10.05
SMEP (ATE), Brazil	0.232	0.122	0.342	1.95
Textbooks for top quintile, Kenya	0.218	0.030	0.406	1.68
Remedial education, India	0.138	0.046	0.230	1.29
Village-based schools, Afghanistan	0.588	0.302	0.874	0.98
Extra contract teacher+streaming, Kenya	0.248	0.068	0.428	0.93
Read-a-thon, Philippines	0.130	0.032	0.228	0.68
Individual computer assisted learning, India	0.475	0.342	0.608	0.65
Contract teachers, Kenya	0.228	0.114	0.342	(0.14)
Unconditional cash transfers, Malawi	-0.030	-0.195	0.135	
Minimum cond cash transfers, Malawi	0.202	-0.029	0.433	
Girls' merit scholarships, Kenya	0.270	-0.044	0.584	N
Providing earnings information, Madagascar	0.202	-0.006	0.410	O
Reducing class size, Kenya	0.074	-0.098	0.246	T
Textbooks, Kenya	0.023	-0.148	0.194	
Flipcharts, Kenya	-0.006	-0.101	0.089	S
Reducing class size, India	0.056	-0.077	0.189	I
Building/improving libraries, India	-0.045	-0.168	0.078	G
School committee grants, Indonesia	0.129	-0.055	0.313	N
School committee grants, Gambia	0.030	-0.146	0.206	I
Computers to classrooms, Colombia	0.109	-0.095	0.313	F
One Laptop Per Child, Peru	0.003	-0.105	0.111	I
Diagnostic feedback, India	0.002	-0.086	0.090	C
Teacher incentives (year 1), Kenya	0.048	-0.072	0.168	A
Teacher incentives (year 2), Kenya	0.136	-0.003	0.275	N
Teacher incentives (long-run), Kenya	0.077	-0.062	0.216	T
Camera monitoring, India	0.170	-0.006	0.346	
Training school committees, Indonesia	-0.049	-0.184	0.086	
Grants/training for school cmte, Gambia	-0.080	-0.256	0.096	

Notes: This table shows the cost-effectiveness of our programme, 'SMEP (ATE), Brazil', the high implementation of our programme, 'SMEP High Implementation (IV), Brazil', and 30 randomised interventions on education analysed by J-PAL (2020). We follow the J-PAL approach presented in Dhaliwal et al. (2013) to calculate the cost-effectiveness of our treatment. However, differently from Dhaliwal et al. (2013) that uses a cut off of 10% of significance level to select experiments to have their cost-effectiveness calculated, we only show the 'Additional SD per \$100' for the random experiments that are significant at the 5% level. 'SD' means standard deviation. Apart from the non-significant interventions, the significant programmes at the 5% level are organised in descending order of their cost-effectiveness measure, i.e., 'Additional SD per \$100'. 'Average Impact (SD)' shows the average impact of a programme on pupils' test scores in standard deviations (SD) across subjects (if more than one subject was used). All costs are in USD (PPP GDP) 2011.

Table IV shows that the two interventions performing better than our high implementation treatment regarding ‘Additional SD per \$100 (PPP)’ have their 95 % confidence intervals’ lower bound technically **at zero standard deviations (SD)**. While the high implementation of our programme has a 95% lower bound of 0.210 standard deviations (SD), Pradhan et al. (2014) ‘Linking school to local govt’ in Indonesia, and Duflo, Dupas, and Kremer (2011) ‘Streaming by achievement’ in Kenya have 95% confidence intervals’ lower bounds of 0.034 SD, and 0.025 SD, respectively. They have a much larger chance than the high implementation level of our intervention to have an actual null impact. Based on a public policy view, the high implementation of the Science and Management for Education Programme should rank first since the intervention has a much lower level of uncertainty on non-zero effects.

The ‘SMEP (ATE)’ represents our programme’s estimated average treatment effect. The ‘SMEP (ATE)’ cost-effectiveness reaches 1.95 ‘Additional SD per \$100 (PPP)’. Table IV shows the three interventions (apart from the ‘SMEP, High Implementation (IV)’ ranking higher than our treatment regarding ‘Additional SD per \$100 (PPP)’ have their 95 % confidence intervals’ lower bound technically **at zero standard deviations (SD)**. While our programme has a 95% lower bound of 0.130 SD, Pradhan et al. (2014) ‘Linking school to local govt’ in Indonesia, Duflo, Dupas, and Kremer (2011) ‘Streaming by achievement’ in Kenya, and Pradhan et al. (2014) ‘Electing school & linking to local govt’ in Indonesia have 95% confidence intervals’ lower bounds of 0.034 SD, 0.025 SD and 0.034 SD, respectively. They have a much larger chance than our treatment (ATE) to have an actual impact of zero. Policywise, the ATE estimates of our treatment would rank second (only behind the ‘SMEP High Implementation (IV)’ since our intervention has a much lower uncertainty regarding the zero impact than the three interventions

mentioned.

A crucial factor in understanding the significant results achieved by our intervention in the cost-benefit and cost-effectiveness analyses is its low cost of USD (PPP GDP) 15.22 per pupil per year. The programme's low cost can be explained by the following factors: the intervention was implemented without changing any existing systems or personnel and without providing any financial incentives, and the programme only worked with school managers without any additional work conducted with teachers or pupils. These factors helped keep the programme's implementation team at a few professionals.²⁵ These professionals were already civil servants of the city of Rio de Janeiro's public administration. Their salaries are the main cost of the programme.

The programme's implementation over two years and two deep management surveys clarified our understanding of the challenges faced by school managers. School managers' routines are often dragged down by bureaucracy and daily crises. Moreover, while there are many incentives provided by civil society, media and watchdog organisations, such as the courts of accounts, for school managers to focus on building facilities, repairs, pupils' food, security, etc., what is obviously important, there is little incentive provided to focus on managing to improve pupils' learning. School managers also lack information on the best management practices to improve pupils' education and the necessary skills to implement these practices. Lack of information - managers not knowing that they are performing poorly and not knowing what they need to do to improve — and lack of motivation — managers not being incentivised to improve or accountable for improvements - are also identified by the management theory as reasons for the persistence of poor

²⁵Since civil servants worked part-time in the implementation, we calculated the number of professionals by the time spent implementing the programme. The equivalent of 6 full-time civil servants worked for two years to implement the programme in 40 public schools.

management practices in organisations Gibbons and Henderson (2012).

Based on our public school's management diagnosis, the reasons behind the success of the Science and Management for Education Programme become clear. Firstly, the programme provided school managers with detailed information on the best management practices tailored to their specific needs. This helped fill the gap in their knowledge of the best management practices to improve pupils' learning. Secondly, the programme not only equipped school managers with the necessary skills to implement these practices but also provided them with simple tools to do so. We made sure to keep things easy for school managers, following a large body of evidence on the impact of "make it easy" on programme participation (Thaler 2021). Thirdly, the programme successfully drew the attention of school managers to ways of improving pupils' learning through better management practices.

From an external validity viewpoint, it is straightforward to generalise the results for all public schools in the city since our experimental sample was randomly drawn from the Rio de Janeiro public school population. Figure A6 in Appendix M shows that Rio de Janeiro would have the best performance in mathematics across all capital cities of the federated states if the city implemented the programme according to the estimated average treatment effect (ATE). If the programme had been implemented to improve the management of the schools by one score point (high-level implementation), the city would have achieved one of the best performances among all Brazilian cities.

School management is likely to matter not only for Rio de Janeiro schools but for all Brazilian public schools. Despite culture, wealth, public investment in education, and other differences across Brazilian cities, public schools have similar organisational structures to be managed. For example, each school has a principal, supervisor (s) or pedagogical coordinator (s) and teachers. Furthermore, public

schools have to follow the same set of Brazil's public administration laws. The organisation of schools are standardised even across the world according to Dobbie and Fryer (2013), Fryer (2017), and Fryer (2014).

Since management matters to improving Brazil's education, this research has important implications for Brazil's education policy. The country faces significant challenges in providing good education to pupils in public schools. Let us conduct a similar generalisation exercise, considering expanding the programme to all Brazilian cities.

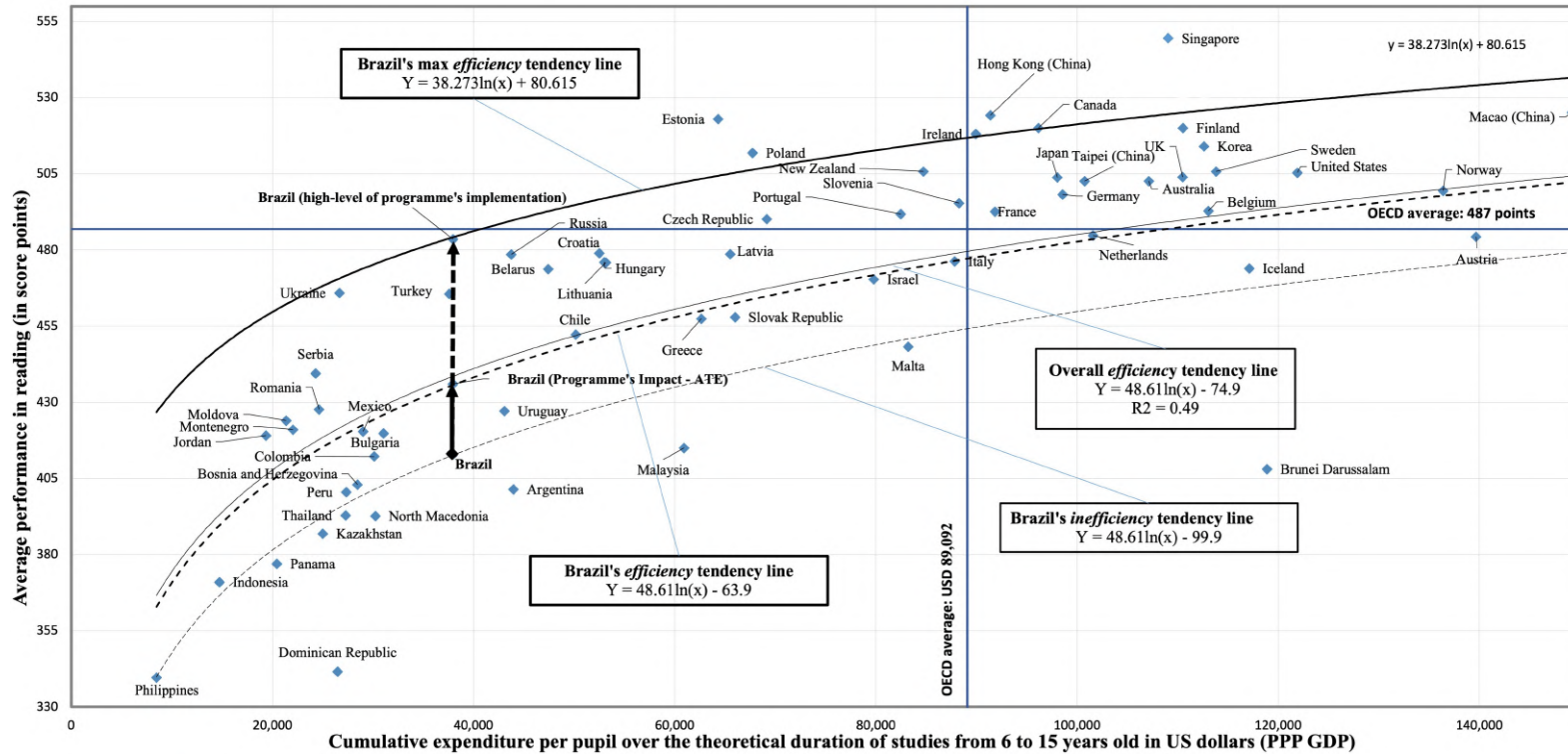
Suppose the Brazilian cities had adopted the Science and Management for Education Programme as a joint national effort to improve education in Brazil. Figure X displays by country²⁶ the average reading performance (in score points) of pupils in OECD's Programme for International Student Assessment (PISA) 2018²⁷ versus the cumulative expenditure per pupil enrolled in primary or secondary education between the ages of 6 and 15 (in equivalent USD converted using PPPs for GDP).

The 'Overall *efficiency* tendency line' represents the PISA 2018 countries' efficiency in using educational resources (performance vs spending), and 'Brazil's *inefficiency* tendency line' represents 2018 Brazil's inefficiency (comparatively with the other countries that participated in PISA 2018) in using its educational resources. Brazil's efficiency gap is the distance (difference) between Brazil's *inefficiency* line and the Overall *efficiency* line. This distance shows how less efficient Brazil was compared to the average efficiency (education spending per pupil vs educational performance) among all countries participating in PISA 2018.

²⁶Although Luxembourg and Qatar are used as data points for the tendency lines regressions, they are not shown in the figure only because they have such large expenditures that would make the x-axis much longer unnecessarily.

²⁷PISA measures 15-year-olds' ability to use their reading, mathematics and science knowledge and skills to meet real-life challenges.

Figure X: Programme's Impact in Brazil Put in a Global Context



82

Notes: The figure displays by country the average reading performance (in score points) of pupils in OECD's Programme for International Student Assessment (PISA) 2018 versus the cumulative expenditure per pupil enrolled in primary or secondary education between the ages of 6 and 15 (in equivalent USD converted using PPPs for GDP). The 'Overall efficiency tendency line' represents the average PISA 2018 countries' efficiency in education (spending vs learning). 'Brazil's inefficiency tendency line' represents Brazil's efficiency (or inefficiency) in education. 'Brazil's efficiency tendency line' and 'Brazil's max efficiency tendency line' show how the programme would change Brazil's performance and efficiency. Data source: PISA 2018 Results (Volume I) - © OECD 2019

Brazil's *efficiency* line shows how Brazil's efficiency would have improved if Brazil had adopted the Science and Management for Education Programme across its cities. Brazil's *efficiency* line considers the PISA score performance that Brazil would have achieved if the programme had been implemented nationwide. Brazil's public per-pupil expenditure in education would not have changed significantly with the programme's presence due to the very low intervention cost of USD (PPP GDP) 15.22 per pupil per year.

Since one standard deviation (SD) is approximately a hundred score points in PISA, the programme's impact (ATE) of 0.23 SD in reading can be translated to 23 PISA score points. Thus, if Brazil had implemented the SMEP, it would have not only largely improved its pupils' learning based on its expected PISA performance, but it would have also closed the educational efficiency gap between the country and all other countries participating in PISA 2018 if it had adopted the SMEP.

It is important also highlight that even spending double (100% more) per pupil per year on education (from USD 37,954 to USD 75,908), Brazil would be below Chile's PISA performance (Chile's cumulative per pupil expenditure on education is USD 50,149,00) as it is possible to see from 'Brazil's *inefficiency* line' in Figure X. However, adopting the programme (ATE) would have made Brazil as efficient as Chile in using educational public resources since both countries would be on very close efficiency lines.

The last line to be discussed in Figure X is the 'Brazil max *efficiency* tendency line'. This line represents how Brazil would have improved its performance and efficiency trajectory if it had implemented the high level of the programme to improve the schools' management by one score point on a scale that goes from one, worst management, to five, best management. Brazil's performance in PISA

could have grown by 68 score points, equivalent to the high implementation of our treatment's impact (IV estimates) on pupils' educational outcomes, 0.68 SD.

The most impressive fact regarding this generalisation exercise is that the intervention discussed in this study simply through improving specific management practices could have made Brazil's efficiency in education similar to developed countries such as Italy, Netherlands or Norway, as we can see from following the 'Brazil's *efficiency* tendency line' from left to right. Moreover, if Brazil had the programme implemented at a high level to change the schools' management level by one score point, Brazil would have closed the entire educational gap between the country and OECD countries' average (72 score points for reading in PISA 2018), keeping costs almost unchanged, as it is possible to see from the vertical dotted arrow in Figure X.

Although representative, this experiment is limited to one context and tests a subset of possible management interventions. In the future, it would be beneficial to conduct similar trials in different contexts to determine how well this approach can be adapted to other low- and middle-income countries and even high-income countries. Furthermore, since the results were collected just after the programme's second year, exploring long-run effects on the management of schools and pupils' learning would be valuable.

Further research may also go deeper into incentives to increase the implementation of the best management practices since we have shown the greater the increase in management, the larger the effects on pupils' learning. Turning the programme into compulsory seems to be the most intuitive path to increase the level of participation. Another solution can be to link higher implementation of the practices with some moral incentives, such as an award for the school management effort to implement the best practices. More research can also be done to tackle the ques-

tion of which of the 23 practices is more important in driving pupils' educational outcomes.

UNIVERSITY OF CAMBRIDGE, UNITED KINGDOM and COURT OF ACCOUNTS OF RIO DE JANEIRO (TCMRio), BRAZIL

UNIVERSITY OF CAMBRIDGE, UNITED KINGDOM and GETULIO VARGAS FOUNDATION (FGV), BRAZIL

Appendix

CAN SCHOOL MANAGEMENT IMPROVE PRODUCTIVITY? EXPERIMENTAL EVIDENCE FROM EDUCATION

Felipe Galvão Puccioni

Tiago Cavalcanti

.A. Efficiency Gap on Education

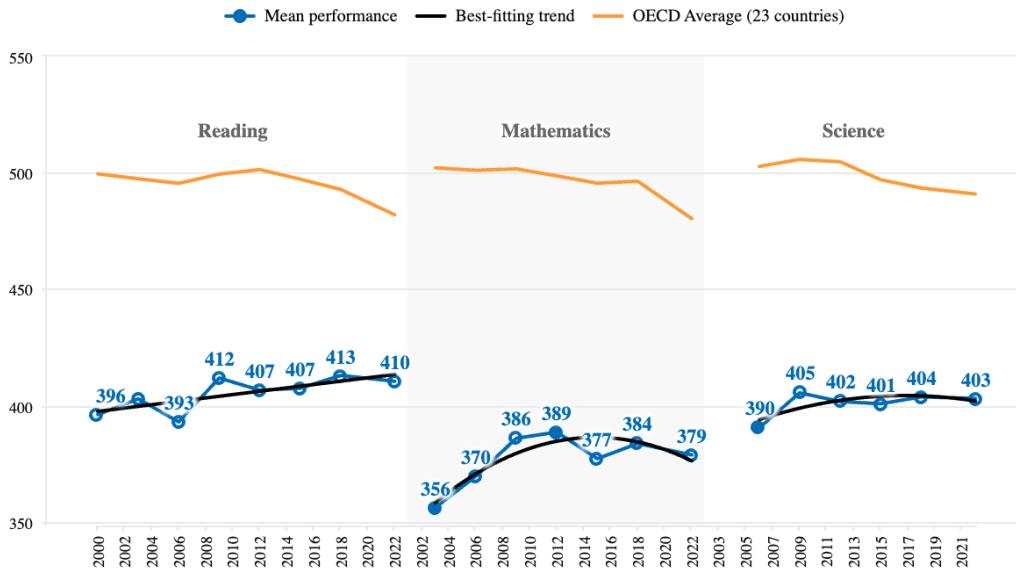
OECD's Programme for International Student Assessment (PISA) evaluates the aptitude of 15-year-old students to use their knowledge and skills in reading, mathematics, and science to tackle real-life problems. Figure A1 compares Brazil's and developed countries' average performance in PISA from 2000 to 2022. The 'OECD Average (23 countries)' yellow lines represent the arithmetic mean performance across the following 23 OECD Member countries: Australia, Belgium, Canada, Czechia, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Latvia, Mexico, New Zealand, Norway, Poland, Portugal, Sweden, and Switzerland.

Figure A1 illustrates the significant educational gap between Brazil and the 23 OECD Member countries (developed countries) for similar years of schooling. A typical 15-year-old Brazilian pupil scored approximately 72 score points lower in reading (482 vs 410), 88 score points lower in science (491 vs 403), and 103 score points lower in mathematics (482 vs 379), compared to the average scores of 15-year-olds from developed countries in PISA 2022 (OECD 2023).

It is estimated that Brazilian pupils' performance in mathematics, reading, and science improve by an average of 12 PISA score points over one year of schooling and age Avvisati and Givord (2021, 2023). This allows for measuring how many

years of schooling a typical 15-year-old Brazilian pupil would need to achieve the PISA performance level of a typical 15-year-old pupil from a developed country. For instance, dividing the difference of 103 score points between Brazil's and developed countries' performance in mathematics in PISA 2022 by Brazil's rate of increase, 12 PISA score points over each year of schooling, gives the years at a Brazilian school a typical 15-year-old Brazilian pupil would need to reach the PISA performance of a typical 15-year-old pupil from developed countries.

Figure A1: Brazil's and Developed Countries' Performance in PISA



Notes: The figure compares Brazil's and 23 developed countries' performance in PISA from 2000 to 2022. 'OECD Average (23 countries)' yellow lines represent the arithmetic mean performance across the following OECD Member countries: Australia, Belgium, Canada, Czechia, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Latvia, Mexico, New Zealand, Norway, Poland, Portugal, Sweden, and Switzerland. Source: OECD, PISA 2022 Database, Tables I.B1.5.4, I.B1.5.5 and I.B1.5.6. Can be accessed here.

Thus, a typical 15-year-old Brazilian pupil would need eight and a half more years in a Brazilian school ($8.58 = \frac{103}{12}$) to reach the mathematics performance of a typical 15-year-old pupil from developed countries. We can also say that, on average, 15-year-old Brazilian pupils are eight and a half years behind pupils from

developed countries in mathematics despite similar levels of schooling (or years at a school). Brazil is approximately six years behind in reading ($6 = \frac{72}{12}$) and seven in sciences ($7.3 = \frac{0.88SD}{0.12SD}$).

In PISA 2022, 75% of 15-year-old Brazilian pupils did not achieve a proficiency level of 2 in mathematics. The proficiency scale in PISA ranges from 1 to 6. This means that these pupils did not meet the most basic level of proficiency expected by the end of their first year of high school, i.e., level 2. After their first year of high school, most 15-year-old Brazilian pupils cannot interpret simple mathematical situations, such as comparing distances or converting local prices to dollars. Interestingly, in PISA 2003, Brazil had the same proportion of 15-year-old students, 75%, performing below the basic PISA level of proficiency in mathematics (level 2).

The large gap between Brazil's and developed country's educational performance is not isolated. Nearly all developing countries are many years of learning behind developed countries for similar years of schooling (Angrist et al. 2021; Bank 2018; Pritchett 2013). Learning inequality also exists within developed countries. The PISA index of economic, social, and cultural status (ESCS) reveals that underprivileged pupils in developed countries scored on average over 100 points lower than their more privileged counterparts with similar years of schooling (OECD 2023). This indicates that underprivileged pupils are lagging years behind their more privileged peers, even though they have received the same amount of schooling in developed countries.

Figure A2 displays the average reading performance (in score points) of pupils in PISA 2018 by country versus the cumulative expenditure per pupil enrolled in primary or secondary education between the ages of 6 and 15 (in equivalent

USD converted using PPPs for GDP).²⁸ Straight lines represent the averages of the OECD Member countries in PISA 2018.

The ‘Overall *efficiency* tendency line’ in Figure A2 is the logarithmic tendency line that represents the average relationship between reading performance in PISA 2018 and the cumulative per-pupil spending on education, considering all countries participating in PISA 2018. We use the word *efficiency* because the Overall tendency line can also be understood as the average PISA 2018 countries’ efficiency in using educational resources (overall efficiency), i.e., resources for education vs educational performance.

‘Brazil’s *inefficiency* tendency line’ in Figure A2 is a logarithmic tendency line showing Brazil’s expected reading performance for different cumulative per pupil spending in education. The difference between Brazil’s inefficiency tendency line and the Overall line is the ‘Brazil’s *efficiency* gap’. This difference represents how less efficient Brazil’s spending on education is compared to the average efficiency across all countries participating in PISA 2018.

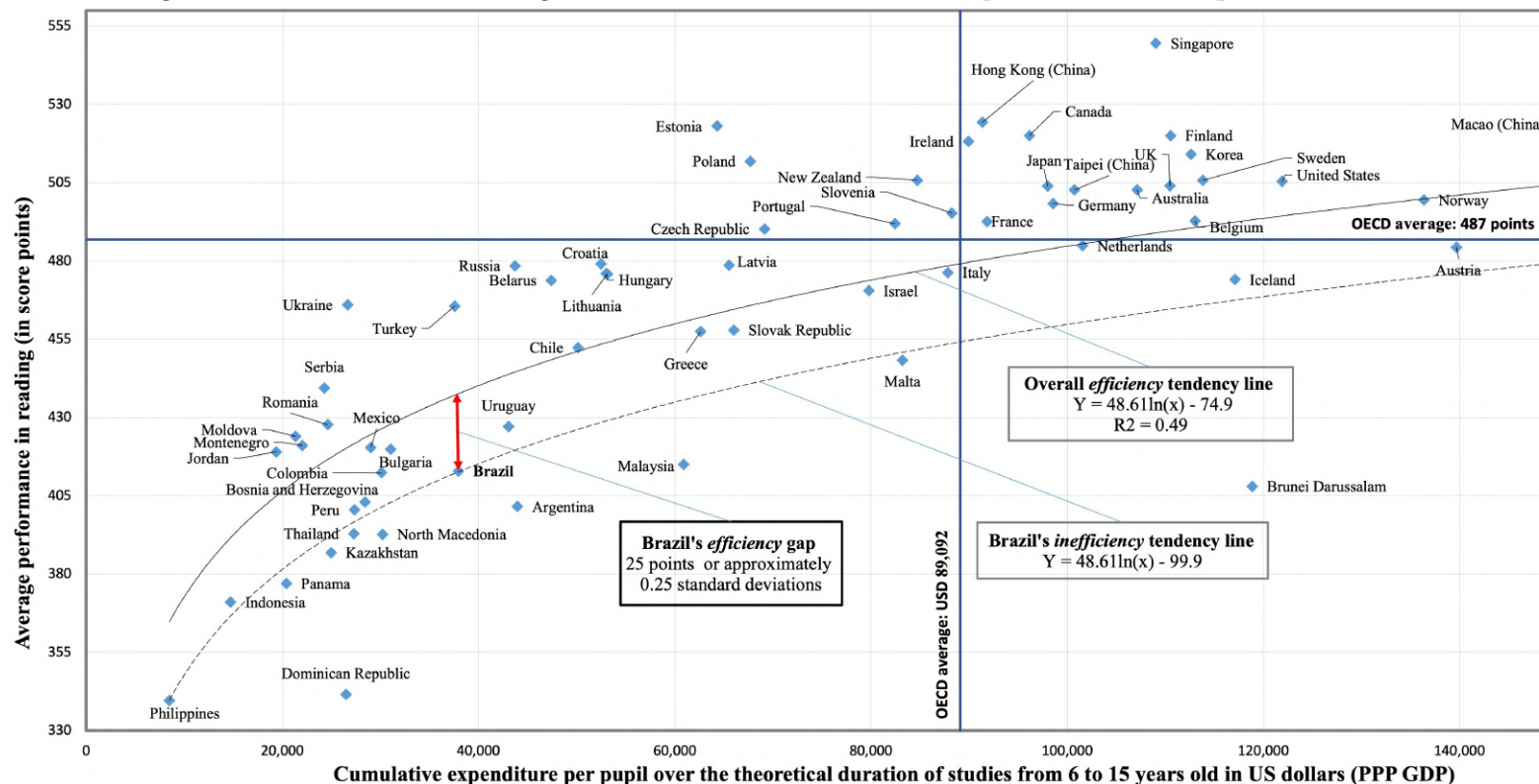
Brazil reached, on average, a reading performance of 413 points for a cumulative expenditure of USD (PPP GDP) 37,954.00, which is 25 points²⁹ lower than expected for its per pupil expenditure. For instance, Russia (479 points), Belarus (474 points), Ukraine (466 points), Turkey (466 points), Chile (452 points), Serbia (439), Romania (428), Uruguay (427), Moldova (424), Mexico (420), Bulgaria (420), and Jordan (419) reached higher score points than Brazil in PISA 2018 reading test despite similar cumulative per-pupil spending on education.

²⁸Although Luxembourg and Qatar are used as data points for the tendency lines (regressions), they are not shown in the figure only because they have such large expenditures that would make the x-axis much longer, worsening the visualisation.

²⁹A hundred PISA score points is approximately one standard deviation. Thus, 25 score points are approximately 0.25 standard deviations (SD).

Figure A2: PISA 2018 Reading Performance vs Cumulative Expenditure Per Pupil on Education

06



Notes: The figure displays the average reading performance (in score points) by country in PISA 2018 versus the cumulative expenditure per pupil enrolled in primary or secondary education between the ages of 6 and 15 (in equivalent USD converted using PPPs for GDP).
 Data source: PISA 2018 Results (Volume I) - © OECD 2019

To better understand, consider that Brazil and Turkey spent around USD (PPP GDP) 37,954.00 on education per pupil from 6 to 15 years old. However, the difference in their proficiency levels was quite significant - 53 points. Turkey seems to be much more efficient than Brazil in allocating educational resources.

As discussed before, Brazilian pupils' test scores in PISA increase by about 12 score points over a year of schooling (and age) (Avvisati and Givord 2021, 2023). Thus, the 53 score points of difference between Brazil and Turkey in PISA means that a typical 15-year-old Brazilian pupil would need more than four years ($\frac{0.53}{0.12}$) at a Brazilian school to reach the educational performance level of a typical 15-year-old Turkish pupil.

Brazil's *inefficiency* tendency line also shows that Brazil is one of the most ineffective countries in allocating educational resources among the countries in PISA 2018. Brazil's inefficiency in education is also discussed by Barros et al. (2019, 2021). Even based on the optimistic assumption that Brazil would follow the same efficiency pattern shown by the Overall line, Brazil would not achieve Chile's performance, even doubling its expenditure per pupil on education. That is, even spending 50% more than Chile, Brazil would be below the country in PISA.

Figure A3 shows the performance of OECD countries, Brazil, Turkey and Vietnam, within each quintile of the PISA index of economic, social and cultural status (ESCS). ESCS allows pupils with similar characteristics, such as parent education and family income, to be comparable across countries. Quintiles are defined at the international level to include 20% of PISA participants in each quintile. The proportion can, therefore, differ from 20% within each national sample. The size of markers is proportional to the share of the pupil population within each quintile of socio-economic status (as determined by the ESCS). Vertical bars extending

beyond the markers represent the 95% confidence interval associated with each estimate. Horizontal, dashed lines represent the 95% confidence interval associated with the mean score of Brazil's largest group of pupils (as defined by international quintiles).

Figure A3 highlights that Brazil's performance in mathematics is much lower than that of OECD Member countries, even comparing pupils within the same quintile of the socio-economic status. The figure also compares Brazil with Turkey and Vietnam, which have similar or lower GDP per capita and educational expenditure per pupil.

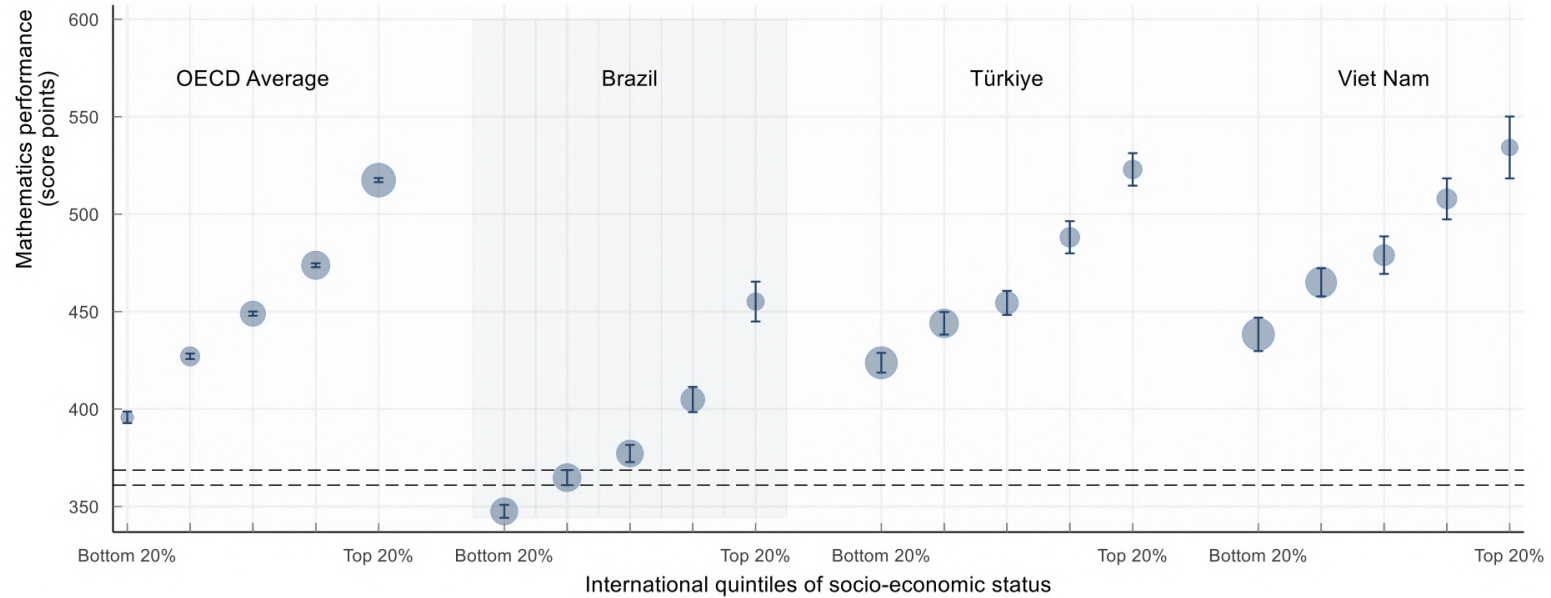
The comparison within each quintile of the socio-economic index shows that Brazil (GDP per capita - USD 8,917.70 in 2022) scored much lower than Turkey, which has a similar GDP per capita (GDP per capita - USD 10,661.20 in 2022).³⁰ Turkey has even the same cumulative per pupil expenditure in education as Brazil, as shown in Figure A2. Even Vietnam, which had a GDP per capita of only USD 4,163.50 in 2022 (half of the Brazilian GDP per capita), scored much higher than Brazil within every quintile of the socio-economic index.

Brazil allocated 6.2% of its GDP to public investment in education in 2018 (Estudos e Pesquisas Anísio Teixeira (INEP) 2021). This is higher than the average of 4.9 % among OECD countries in the same year (OECD 2021). The amount invested by Brazil aligns with the educational spending benchmark set by 160 countries in the 2015 Incheon Declaration³¹, which requires at least 4-6% of GDP to be spent on public education.

³⁰Countries' GDP per capita in 2022 can be accessed here.

³¹The document can be accessed at <https://unesdoc.unesco.org/ark:/48223/pf0000233137>.

Figure A3: Mean performance in mathematics, by international quintiles of socio-economic status



96

Note: The size of markers is proportional to the share of the pupil population within each quintile of socio-economic status (as determined by the PISA index of economic, social and cultural status, ESCS). Quintiles are defined at the international level to include 20% of PISA participants in each quintile; within each national sample, the proportion can therefore differ from 20%. Vertical bars extending beyond the markers represent the 95% confidence interval associated with each estimate. Horizontal, dashed lines represent the 95% confidence interval associated with the mean score of Brazil's largest group of pupils (as defined by international quintiles). *Source:* OECD, PISA 2022 Database, Tables I.B1.4.6 and I.B1.4.8

Brazil's public expenditure on primary, secondary and post-secondary non-tertiary education amounted to 4% of its GDP in 2018, surpassing the OECD average of 3.4% for the same year (Estudos e Pesquisas Anísio Teixeira (INEP) 2021; OECD 2021).

Brazil's fiscal efforts towards education are not lower than OECD countries. However, Figures A1, A3 and A2 show that Brazil's commitment to education is not enough, both from the point of view of the necessity of more investment - low per pupil expenditure - and the efficiency of this investment - lower than expected educational performance for the corresponding per pupil spending. Even though many factors may be responsible for driving pupils' learning, such as family background, culture, country wealth, investment in education, etc., our focus here is to understand how a country, a city or a school can provide more education from each dollar available to education.

A. Additional Information on Background

Table A1 compares the municipality of Rio de Janeiro and other Brazilian municipalities from 2007 to 2019. The public expenditure per pupil considers not only the municipality spending on grades 1-9 public schools but also on nurseries, special education and young adults education.³² Table A1 was based on data from the Brazilian Education Census³³ and 2020 Court of Accounts of Rio de Janeiro (TCMRio) Special Report (p.113).³⁴ All monetary values are updated to December 2019 and are in Brazilian Reais (R\$).

³²The municipality's public expenditure per pupil is calculated by dividing the municipality's total spending on public education by the number of pupils in public nurseries, grades 1-9 schools, and special education, plus the number of young adults in state-run grades 1-9 schools.

³³It can be accessed here.

³⁴It can be accessed here.

Table A1: Education in the Municipality of Rio de Janeiro - 2007 and 2019

Expenditure per pupil on education			
	2007	2019	Variation
Brazil's cities average	R\$ 4,841.72	R\$ 7,885.57	+63%
Rio de Janeiro City	R\$ 4,805.56	R\$ 9,707.65	+102%

Rio de Janeiro City position - 5th grade Ideb			
	2007	2019	Variation
In RJ Federated State	24th	28th	- 4 positions
In Brazil	1423rd	2529th	- 1106 positions

Rio de Janeiro City position - 9th grade Ideb			
	2007	2019	Variation
In RJ Federated State	7th	28th	- 21 positions
In Brazil	366th	952nd	- 586 positions

Notes: Table based on data from the Brazilian Education Census (can be accessed here) and the 2020 Court of Accounts of Rio de Janeiro (TCMRio) Special Report (p.113) (can be accessed here). All monetary values are updated to December 2019 and expressed in Brazilian Reais (R\$). The expenditure per pupil is calculated by dividing the city's total spending on education by the number of pupils in nurseries, grades 1-9 schools and special education plus the number of young adults in special grades 1-9 schools. The Basic Education Development Index (Ideb) brings together, in a single indicator, the results of two equally important concepts for the quality of education: the rate of pupils' approval and average performance in reading and mathematics assessments (Saeb). We compare the city of Rio de Janeiro's performance in Ideb with the cities within the Rio de Janeiro Federated State and Brazil. RJ Federated State means Rio de Janeiro Federated State.

B. Additional Information about the Science and Management for Education Programme (SMEP) and Its Implementation

First, it is important to highlight that our study complies with all critical research protocols. It received approval from the University of Cambridge's Ethical Committee and is registered on the AEA RCT registry (AEARCTR-0007669 on 15 May 2021; <https://www.socialscienceregistry.org/trials/7669>). Rio de Janeiro authorities and institutions responsible for education in the city also approved the experiment.

Our intervention aimed to provide the 23 best management practices for schools described and discussed by Bloom et al. (2015). These practices can be grouped into five groups: operations, monitoring, target-setting, people management and leadership as follows.

- Operations

- Standardisation of instructional planning processes: school uses meaningful processes that allow pupils to learn over time (o1).
- Personalisation of instruction and learning: school incorporates teaching methods that ensure all pupils can master the learning objectives (o2).
- Data-driven planning and pupil transitions: school uses assessment and easily available data to verify learning outcomes at critical stages (o3).
- Adopting educational best practices: school incorporates and shares teaching best practices and pupil strategies across classrooms accordingly (o4).

- Monitoring

- Continuous improvement: school implements processes towards continuous improvement and encourages lessons to be captured and documented (m5).
- Performance tracking: school performance is regularly tracked with useful metrics (m6).
- Performance review: school performance is reviewed with appropriate metrics (m7).






- Performance dialogue: school performance is discussed with appropriate content, depth and communicated to teachers (m8).
 - Consequence management: mechanisms exist to follow up on performance issues (m9).
- Target setting
 - Target balance: school covers a sufficiently broad set of targets at the school, department and individual levels (t10).
 - Target interconnection: school establishes well-aligned targets across all levels (t11).
 - Time horizon of targets: there is a rational approach to planning and setting targets (t12).
 - Target stretch: school sets targets with the appropriate difficulty level (t13).
 - Clarity and comparability of targets: school sets understandable targets and openly communicates and compares school, department and individual performance (t14).
- People management
 - Rewarding high performers: school implements a systematic approach to identifying good and bad performance, rewarding teachers proportionately (p15).
 - Fixing poor performers: school deals with under-performers promptly (p16).
 - Promoting high performers: school promotes employees based on job performance (p17).

- Managing talent: school nurtures and develops teaching and leadership talent (p18).
 - Retaining talent: school attempts to retain high-performing employees (p19).
 - Creating a distinctive employee value proposition: school has a thought-through approach to attract employees (p20).
- Leadership
 - Leadership vision - School leaders have an understanding of the broader set of challenges that the school, system and key actors face (l21).
 - Clearly defined accountability for school leaders - School leaders are accountable for delivery of student outcomes the right mindset to address them (l22).
 - Clearly defined leadership and teacher roles - How clearly the roles, responsibilities and required attributes of teachers, students and staff are defined within the school (l23).

The code between parenthesis after each practice description aimed to help future citations of the management practices in this paper.

Figure A4 presents the 23 best management practices for schools and the description of the best school scenario (score five) for each practice. This table was provided for each school manager from treatment schools (principals, deputy principals and pedagogical coordinators). The goal was to make it easy for school managers to identify the school's goal regarding each one of the management practices.

Figure A4: The 23 Best Management Practices for Schools Chart

 Operations	 Monitoring	 Target Setting	 People Management	 Leadership
<p>1. Standardisation of Instructional Processes School has implemented a clearly defined instructional planning process designed to align instructional strategies and materials with learning expectations and incorporate flexibility to meet student needs; these are followed up on through comprehensive monitoring or oversight.</p> <p>2. Personalization of Instruction and Learning Emphasis is placed on personalization of instruction based on student needs; school encourages student involvement and participation in classrooms; school provides information to and connects students and parents with sufficient resources to support student learning.</p> <p>3. Data-Driven Planning and Student Transitions Student transitions are managed in an integrated and proactive manner, supported by formative assessments tightly linked to learning expectations; data is widely available and easy to use.</p> <p>4. Adopting Educational Best Practices School provides staff with opportunities to collaborate and share best practice techniques and learnings with multiple methods to support their monitored implementation in the classroom.</p>	<p>5. Continuous Improvement Texposing and solving problems (for the school, individual students, teachers, and staff) in a structured way is integral to individual's responsibilities, and resolution involves all appropriate individuals and staff groups; resolution of problems is performed as part of regular management processes.</p> <p>6. Performance Tracking Performance is continuously tracked and communicated, both formally and informally, to all staff using a range of visual management tools.</p> <p>7. Performance Review Performance is continually reviewed, based on indicators; all aspects are followed up to ensure continuous improvement; results are communicated to all staff.</p> <p>8. Performance Dialogue Regular review/ performance conversations focus on problem solving and addressing root causes; purpose, agenda and follow-up steps are clear to all; meetings are an opportunity for constructive feedback and coaching.</p> <p>9. Consequence Management A failure to achieve agreed targets drives retraining in identified areas of weakness, moving individuals to where their skills are more appropriate.</p>	<p>10. Target Balance Performance metrics and targets are defined for the school and individuals (leaders, teachers, staff) that include both absolute and value-added measures of student outcomes and other metrics linked to key drivers of student outcomes.</p> <p>11. Target Inter-Connection Goals are aligned and linked at system level and increase in specificity as they cascade, ultimately defining individual expectations for all staff groups.</p> <p>12. Time Horizon of Targets Long-term goals are translated into specific short-term targets so that short-term targets become a 'staircase' to reach long-term goals.</p> <p>13. Target Stretch Goals are genuinely demanding for all parts of the organization and developed in consultation with senior staff (e.g. to adjust external benchmarks appropriately).</p> <p>14. Clarity and Comparability of Targets Performance measures are well defined, strongly communicated and reinforced at all reviews; school performance data includes both quantitative and qualitative measures and are made public.</p>	<p>15. Rewarding High Performers There is an evaluation system which rewards individuals based on performance; the system includes both personal financial and non-financial awards; rewards are awarded as a consequence of well-defined and monitored individual achievements.</p> <p>16. Removing Poor Performers Repeated poor performance is addressed, beginning with targeted interventions; poor performers are moved out of the school when weaknesses cannot be overcome.</p> <p>17. Promoting High Performers School actively identifies, develops and promotes its top performing staff members.</p> <p>18. Managing Talent School proactively controls the number and types of teachers, staff and leadership needed to meet goals; school defines hiring criteria and processes based on understanding of what drives student achievement.</p> <p>19. Retaining Talent We do whatever it takes to retain our talent.</p> <p>20. Attracting Talent We provide a unique value proposition to encourage talented people join our school above our competitors.</p>	<p>21. Leadership Vision School leaders define and broadly communicate a shared vision and purpose for the school that focuses on improving student learning and outcomes (often beyond those required by law); vision and purpose is built upon a keen understanding of student and community needs, and defined collaboratively with a wide range of stakeholders; school leader proactively builds environment conducive to learning.</p> <p>22. Clearly Defined Accountability for School Leaders School leaders are held accountable for quality, equity and cost- effectiveness of student outcomes within the school, with school-level and individual consequences for good and poor performance; leaders are provided sufficient autonomy to impact the areas of accountability.</p> <p>23. Clearly Defined Leadership and Teacher Roles School defines clear roles, responsibilities and desired competencies of teachers and staff across the school, built upon an understanding of what drives student performance and outcomes; leadership responsibilities are distributed across school.</p>

Source: The 23 best management practices for schools (Bloom et al. 2015). Since the WMS defines for each practice the worst management scenario (score one), the midday management scenario (score three) and the best management scenario (score five), we include below each practice in the table the description of the best school scenario (score five) for that practice. This table was provided for treatment school managers (principals, deputy principals and pedagogical coordinators).

C. Additional Information on the Management Level of Schools in 2021 and 2023

Table A2 presents the distribution of ‘subjects’ by rater 1 and 2 and category for the management surveys conducted in 2021 and 2023. Subjects are the 1,840 (23 practices x 80 schools) management practices surveyed in 2021 and 2023 since the same management practices in different schools are different survey ‘subjects’. In the 2021 survey, 77 subjects were not scored by rater two because the interview recording had problems.

The table shows that raters agreed most of the time when the scores were low. The percentage of disagreement rises for higher scores.

More systematically, Table A3 shows inter-rater reliability coefficients for our 2021 and 2023 management surveys and a benchmark scale based on Landis and Koch (1977). These coefficients measure the agreement between raters when the ratings are ordinal (Gwet 2021).

The Percent Agreement shown in Table A3 is the only coefficient not corrected by chance agreement, i.e., the Percent Agreement includes agreements between raters that are not due to chance and agreements due to chance. All other coefficients have paths to exclude rater agreement due to chance from their calculations, i.e., they measure the systematic agreement that is not due to chance.

As seen from Table A2, the survey scores are more concentrated in some categories than others. This phenomenon is called high trait prevalence, and it is known to have a dramatic effect on many inter-rater reliability coefficients, such as Cohen/Conger’s Kappa, Scott/Fleiss’s Pi, and Krippendorff’s Alpha (Feinstein and Cicchetti 1990; Gwet 2021). Consequently, we focus our analysis on the Brennan-Prediger and Gwet’s AC2 coefficients because Gwet (2021) has shown they are more resistant to high trait prevalence paradoxes.

Table A2: Contingency Tables for the 2021 and 2023 Management Surveys
2021 management survey

Rater 1	Rater 2					Total
	1	2	3	4	5	
1	409	128	51	11	2	601
2	71	148	113	28	5	365
3	38	106	219	111	12	486
4	5	33	92	99	24	253
5	0	3	22	21	12	58
Total	523	418	497	270	55	1,763

2023 management survey

Rater 1	Rater 2					Total
	1	2	3	4	5	
1	281	80	6	2	1	370
2	87	195	98	32	2	414
3	13	87	215	153	4	472
4	5	26	148	297	36	512
5	1	1	10	51	9	72
Total	387	389	477	535	52	1,840

Notes: This tables shows the 2021 and 2023 distribution of ‘subjects’ by rater and category. Subjects are the 1,840 (23 dimensions x 80 schools) management practices surveyed in the 80 sample schools since the same management practice in different schools are different ‘subjects’. In the 2021 survey, 77 subjects were not scored by rater two because the interview recording had problems. Raters are the survey team members who scored the management practices across the sample schools. Raters scored each of the 23 WMS management practices in each school considering the WMS scale of categories that goes from category one, worst management level, to category 5, best management level.

Table A3 shows a Brennan-Prediger coefficient of 0.720 and a Gwet’s AC2 of 0.760 for the 2021 survey. The analysis of the 2023 survey presents higher agreement coefficients, such as a Brennan-Prediger of 0.788 and a Gwet’s AC2 of 0.811. These values are considered by Landis and Koch (1977) to be relatively large and close to perfect agreement between raters. These high agreement coefficients in both surveys indicate that the scores assigned to each of the 23 management practices for each school are independent of the specific survey team member (rater

Table A3: Agreement Coefficients for the 2021 and 2023 Management Surveys

	2021	2023
Percent Agreement	0.922	0.941
Brennan and Prediger	0.720	0.788
Cohen/Conger's Kappa	0.608	0.709
Scott/Fleiss' Pi	0.608	0.709
Gwet's AC2	0.760	0.811
Krippendorff's Alpha	0.608	0.709

Benchmark scale

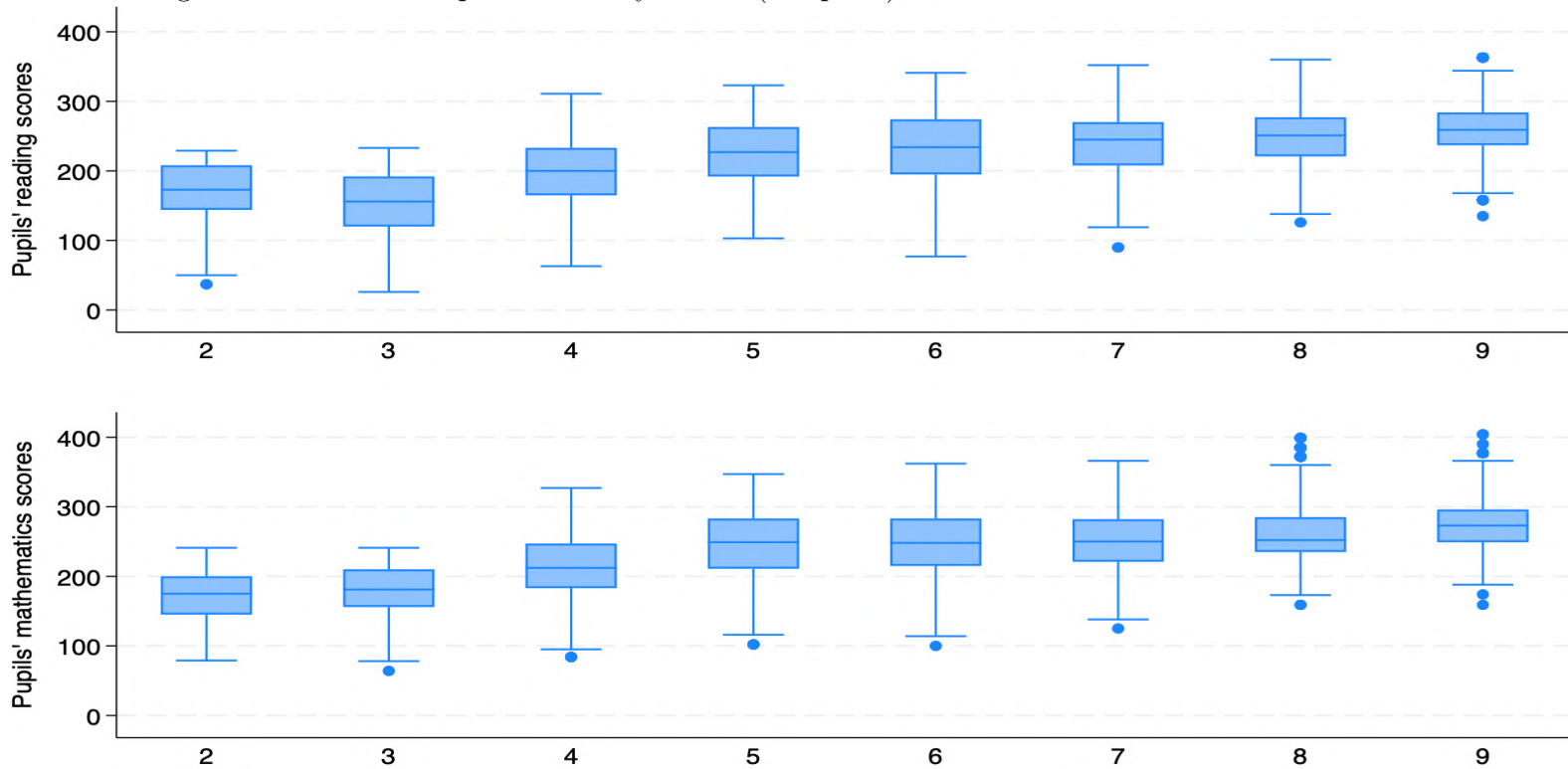
<0.000	Poor
0.000 - 0.200	Slight
0.200 - 0.400	Fair
0.400 - 0.600	Moderate
0.600 - 0.800	Substantial
0.800 - 1.000	Almost Perfect

Notes: This table shows different inter-rater reliability coefficients for the 2021 and 2023 management surveys. They measure the agreement between raters when the ratings are ordinal (Gwet 2021). The WMS management score goes from 1, worst management, to 5, best management. Two raters scored each of the 23 management practices for each school in both surveys. Since there were 80 schools and 23 dimensions, 1,840 'subjects' were scored by two raters in each survey. The Percent Agreement is the only coefficient not corrected by chance agreement. The Brennan-Prediger and Gwet's AC2 coefficients are in bold because they are robust to high trait prevalence paradoxes such as encountered in the scores from our surveys (Gwet 2021). The benchmark scale was developed by Landis and Koch (1977).

or coder) who evaluated the school. The evidence suggests that the management scores given to schools are not subjective views of raters but actual school characteristics.

D. Additional Information on Other Relevant Variables

Figure A5: Control Pupils' Scores by Grade (Boxplots) in the December 2023 Rio Assessments



Notes: The figure shows the control pupils' scores distribution by grade for reading and mathematics in the December 2023 Rio Assessments. Boxplots represent the distribution of scores for each grade.

Table A4: Pupils' Educational Performance Missing Data in 2022 and 2023

Pupils performance	Before adjustment		After adjustment	
	Missing	Percent	Missing	Percent
Reading - 04/22	2,650	8.34 %	2,650	8.34 %
Mathematics - 04/22	2,689	8.47 %	2,689	8.47 %
Reading - 06/22	2,950	9.29 %	844	2.66 %
Mathematics - 06/22	3,111	9.80 %	881	2.77 %
Reading - 04/22	3,008	9.47 %	511	1.61 %
Mathematics - 09/22	3,060	9.63 %	519	1.63 %
Reading - 09/22	3,001	9.45 %	399	1.26 %
Mathematics - 12/22	3,049	9.60 %	398	1.25 %
Reading - 04/23	5,289	16.65 %	368	1.16 %
Mathematics - 04/23	5,311	16.72 %	375	1.18 %
Reading - 06/23	3,849	12.12 %	351	1.11 %
Mathematics - 06/23	3,973	12.51 %	362	1.14 %
Reading - 09/23	4,006	12.61 %	350	1.10 %
Mathematics -09/23	4,067	12.81 %	356	1.12 %
Reading - 12/23	3,878	12.21 %	348	1.10 %
Mathematics - 12/23	3,985	12.55 %	355	1.12 %

Notes: The table shows the pupils' education performance missing data in reading and mathematics across eight different dates of application of the Rio Assessments. Our sample has 31,760 fixed pupils. The columns below 'Before adjustment' show information without any change to fill in missing scores. The columns under 'After adjustment' present information after we fill in missing scores with standardised scores from previous Rio Assessments tests. For instance, if a pupil did not have the standardised reading score from the December 2023 Rio Assessments, we used the standardised reading score from the test held in September 2023 to replace the December 2023 missing score. If the September 2023 reading score was also missing, we used the standardised score from the Rio Assessments held in June 2023. We continue this process until April 2022 if necessary. This is a conservative approach since we use data from when a pupil was less exposed to the programme.

E. Additional information on the Pre-Treatment Summary Statistics

Data in Table A5 comes from the Rio de Janeiro Municipal Secretariat of Education (SMERJ). This table presents information on the characteristics and educational outcomes of grade 1-8 pupils enrolled (during November/December 2021) in the population of 991 schools under the SMERJ for the 2022 school year. During the 2021 school year, there were 992 grade 1-9 schools under the responsibility of the city of Rio de Janeiro. However, one of the schools was deactivated at the end of 2021. Thus, there were 991 schools under the city's responsibility for the 2022 school year. The data was collected at the end of 2021.

The 'Sample' column from Table A5 presents the difference between the sample and population averages with standard errors of the differences in parenthesis. Standard errors are clustered robust standard errors with clusters defined as the schools. Each average, difference between averages, and standard error comes from a regression analysis where the variables in the table are regressed on a population-sample dummy. This dummy switches off for a population's pupil and on for a sample's pupil. Sample pupils have two observations in the data, each with a different identification - one for the population and one for the sample. The educational outcomes come from standardised reading and mathematics tests, which are meant to be taken by all students across two different dates: September and December 2021. We call these tests Rio Assessments. Pupils' scores from the 2021 Rio Assessments were standardised by subject, grade, and test application date using the population as the reference.

Table A5 reveals that the Rio de Janeiro schools have 34.6%, 53.3% and 11.9% of white, brown and black pupils, respectively. Also, 29.7% of pupils' families receive support from the Brazilian conditional cash transfer programme, Bolsa Família.

Table A5: Pre-Treatment Statistics: Pupils Population

	Valid obs	Population (mean)	Sample (difference/SE)
A. Characteristics			
White (%)	357,623	0.346	-0.003 (0.011)
Brown (%)	357,623	0.533	0.002 (0.008)
Black (%)	357,623	0.119	0.000 (0.005)
Female (%)	390,018	0.486	0.000 (0.003)
Bolsa família (% receiving)	390,018	0.297	0.013 (0.021)
B. Educational outcomes			
Reading - 09/2021	281,002	0.000	0.053 (0.027)
Mathematics - 09/2021	280,546	0.000	0.065 (0.026)
Reading - 12/2021	291,466	0.000	0.035 (0.028)
Mathematics - 12/2021	290,602	0.000	0.040 (0.029)
Pupils total		393,871	31,760

Notes: This table presents information on the characteristics and educational outcomes of pupils enrolled (during November/December 2021) in the Rio de Janeiro population of 991 grade 1-8 schools for the 2022 school year. The Municipal Secretary of Education of Rio de Janeiro (SMERJ) collected the data at the end of 2021. The column ‘Valid obs’ shows the number of observations or pupils with data available. The ‘Population’ column presents the pupils’ population averages. The ‘Sample’ column shows the difference between the sample and population averages with standard errors of the differences in parenthesis. Standard errors are robust standard errors. Each average, difference between averages, and standard error comes from a regression analysis where relevant variables in the table are regressed on a population-sample dummy. This dummy switches off for a population’s pupil and on for a sample’s pupil. Sample pupils have two observations in the data, each with a different identification - one for the population and one for the sample. Pupils’ scores were standardised by subject, grade, and test application date using the entire population as the reference.

The differences between the population and the experimental sample pupils’ averages regarding race, gender, and Bolsa Família are almost null in Table A5.

The differences between the population and sample averages regarding pupils' mathematics and reading scores are also minimal. The population average scores were zero because the population was the reference group for the standardisation procedure.

Table A6 reports the average characteristics of principals, pedagogical coordinators, teachers and schools for the 2022 school year. Data was collected by the Municipal Secretariat of Education of Rio de Janeiro at the end of 2021. The column 'Valid obs' shows the number of observations or schools with data available. The 'Sample' column shows the difference between the sample averages and the population. 'SE' means robust standard errors and is reported in parenthesis. Each average, difference between averages, and standard error comes from a regression (at the school level) in which we regress each relevant variable in the table on a population-sample dummy that switches off for a population's school and on for a sample's school (each sample school has two observations in the data with different identifications, one for the population and one for the sample).

The data presented in Table A6 indicates that a much higher percentage of women work as principals (84.3%), pedagogical coordinators (89.3%), and teachers (80.1%) in the Rio de Janeiro schools compared to men. Also, more than 80% of the professionals in these three careers have a college degree. On average, principals are approximately six years older than pedagogical coordinators. The average number of grade 1-8 pupils in Rio de Janeiro schools is 397. Segment I schools (grades 1-5) are 70.9% of the schools' population and are 70% of the schools' sample. Segment II schools are 29.1% in the population and are 30% in the sample.

The differences between the averages of the population of schools and the averages of the experimental sample of schools are almost null.

Table A6: Pre-Treatment Statistics - Schools Population

	Valid obs	Population (mean)	Sample (difference/SE)
Principals			
Female (%)	986	0.843	0.007 (0.042)
Age (mean)	986	50.6	0.2 (1.0)
College degree (%)	985	0.812	-0.075 (0.051)
Pedagogical Coordinators			
Female (%)	940	0.893	0.041 (0.031)
Age (mean)	940	44.1	-0.5 (0.9)
College degree (%)	933	0.874	-0.049 (0.046)
Teachers (avg by school)			
Female (%)	991	0.801	0.012 (0.017)
College degree (%)	991	0.842	-0.008 (0.017)
School characteristics			
Pupils per school (mean)	991	397	0 (23)
Segment I schools (%)	991	0.709	-0.009 (0.053)
Segment II schools (%)	991	0.291	0.009 (0.053)

Notes: This table reports the average characteristics of principals, pedagogical coordinators, teachers and schools for the 2022 school year. The Municipal Secretary of Education of Rio de Janeiro (SMERJ) collected the data at the end of 2021. The column ‘Valid obs’ shows the number of observations or schools with data available. The ‘Population’ column shows the averages of the Rio de Janeiro population of schools. The ‘Sample’ column shows the difference between the averages of the sample and the population with standard errors of the difference in parenthesis. Standard errors are robust standard errors. Each average, difference between averages, and standard error comes from a regression (at the school level) in which we regress each relevant variable in the table on a population-sample dummy that switches off for a population’s school and on for a sample’s school (sample schools have two observations in the data with different identifications, one for the population and one for the sample).

As expected from a random sampling procedure, Tables A5 and A6 reveal that our experimental sample is very similar, on average, to the Rio de Janeiro population of grade 1-8 pupils and schools.

F. Cost-Benefit Analysis - Krueger Approach

We followed Krueger (2003) to calculate the internal rates of return (IRRs) of the Science and Management for Education Programme (SMEP) based on the expected income benefits from the increased pupils' educational performance. Consider pupils entering the 8th grade at the beginning of 2022 without loss of generality. Suppose that the earnings of the current labour force in Brazil represent the profile of earnings by age that the average pupil who entered 8th grade (13 years old) in 2022 will experience when they start in the labour market.

Denote the per pupil cost of the programme in year t as C_t . We used the ingredients method to calculate the Science and Management for Education Programme cost (Dhaliwal et al. 2013). The programme lasted two years. Let E_t be the individual's annual earnings in Brazil each year t from age 18 until the individual retires at 65. δ represents the increase in earnings associated with one standard deviation (SD) increase in either mathematics or reading. According to Chetty, Friedman, and Rockoff (2014), Krueger (2003), and Neal and Johnson (1996), the increase in earnings associated with an increase of one standard deviation in reading or mathematics is between 8 and 20%. We used $\delta = 0.12$ to make the comparison with (Fryer 2017) possible.

Real earnings are likely to grow largely between 2022 and when a typical Brazilian pupil who started the 8th grade in 2022 retires decades later. Per capita earnings and productivity have historically grown around 3.71% in Brazil (Tombolo and Sampaio 2013). Let g be the real earnings growth rate by year. Lastly, let

β_{ATE_m} and β_{ATE_r} be the programme’s impact on pupils’ mathematics and reading test scores due to being assigned to a treatment school. The internal rate of return (IRR), r , can be calculated by determining the discount rate at which the present value of future cash flows (benefits) equals the initial investment (programme cost).

$$\sum_{t=1}^2 C_t / (1+r)^t = \sum_{t=18}^{65} E_t \times (1+g)^{t-13} \times \delta \times (\beta_{ATE_m} + \beta_{ATE_r}) / (1+r)^{t-13} \quad (4)$$

The superscripts on $(1+g)$ and $(1+r)$ are $t-13$ since the pupils starting the 8th grade are 13 years old and are five years from being 18 years old (and start in the labour market). We also calculate the IRR considering the treatment intensity analysis (IV) by replacing the impact parameters from the equation above with the IV impact estimates.

Table A7 reports the cost per pupil per year considering treatment and control arms, pupils’ gains in reading and mathematics, and the IRR for the interventions discussed before. Apart from our experiment and the Jovem de Futuro Programme, all the other studies’ values in Table A7 come from Fryer (2017, 2016). We updated all cost values from Fryer (2017) to January 2022. The Science and Management for Education Programme and the Jovem de Futuro costs are in USD PPP GDP (2022). The column ‘Gains’ shows the sum of the causal effect of each intervention on pupils’ reading and mathematics in SD.

The different interventions in education presented in Figure IX are related to different areas.

- **Management:** our ‘SMEP Average Treatment Effect (ATE)’, and our ‘SMEP High Implementation (IV)’; Fryer (2017) ‘Houston High Implementation (HI)’, ‘Houston Principal Staying (PS)’, and ‘Houston Overall’; Barros et al. (2019, 2021) ‘Jovem de Futuro’.

Table A7: Cost-Benefit Analysis (CBA)

	Cost per pupil per year (treatment)	Cost per pupil per year (control)	Gains Read+Math (SD)	Internal rates of return (IRR)
Management				
SMEP, High Implementation (IV)	\$15.22	\$0.00	1.39	163%
SMEP (ATE)	\$15.22	\$0.00	0.46	115%
Houston Overall	\$11.55	\$0.42	0.06	79%
Houston HI	\$11.55	\$0.42	0.12	96%
Houston PS	\$11.55	\$0.42	0.11	94%
Jovem de Futuro	\$216.79	\$0.00	0.20	32%
Early Childhood				
Head Start Impact Study	\$12,002.31	\$3,782.33	0.32	9 %
Charter Schools				
Injecting Best Practices ES	\$433.60	\$ 0.00	0.26	35 %
Injecting Best Practices SS	\$2,243.74	\$ 0.00	0.13	18 %
Harlem Children's Zone ES	\$25,560.18	\$16,502.98	0.31	11 %
Harlem Children's Zone MS	\$25,560.19	\$16,502.98	0.28	12 %
SEED Schools	\$51,904.45	\$27,122.53	0.42	9 %
Teacher Incentives				
Talent Transfer Initiative	\$788.67	\$0.00	0.24	28 %
Teacher Certification				
Teach For America	\$4,455.20	\$0.00	0.18	12 %
Class Size				
Tennessee STAR	\$6,111.44	\$0.00	0.24	10 %
Professional Dev				
Success for All	\$1,024.49	\$0.00	0.09	14 %
Tutoring				
Experience Corps	\$1,045.16	\$0.00	0.08	14 %
Curriculum				
Enhanced Reading	\$2,519.73	\$0.00	0.18	22 %
Financial Incentives				
Coshocton Incentive Program	\$91.62	\$0.00	0.12	49 %
New York, Dallas, Chicago	\$429.01	\$0.00	0.00	-

Notes: This table presents the per pupil costs per year, the impact and the internal rates of return for our experiment, 17 interventions presented in Fryer (2017) and the Jovem de Futuro Programme analysed by Barros et al. (2019, 2021). Apart from our experiment and the Jovem de Futuro Programme, all other studies' values come from Fryer (2017, 2016). We updated all the cost values reported by Fryer (2017) to January 2022. The IRR is the discount rate that equates the cost of an intervention with the present value of the expected future earnings inflows that the intervention generates. We calculated the IRR for our Science and Management for Education Programme (SMEP) and the Jovem de Futuro Programme following the methodology developed by Krueger (2003). The Science and Management for Education Programme and the Jovem de Futuro costs are in USD PPP GDP (2022).

- **Early Childhood:** Puma et al. (2010) ‘Head Start Impact Study’.
- **Charter schools:** Fryer (2014) ‘Injecting Best Practices Elementary Schools (ES)’, and ‘Injecting Best Practices Secondary Schools (SE)’; Dobbie and Fryer (2011) ‘Harlem Children’s Zone Elementary Schools (ES)’, and ‘Harlem Children’s Zone Middle Schools (ME)’; Curto and Fryer (2014) ‘SEED Schools’.
- **Teacher Incentives:** Glazerman et al. (2013) ‘Talent Transfer Initiative’.
- **Teacher Certification:** Glazerman, Mayer, and Decker (2006) ‘Teach For America’.
- **Class Size:** Krueger (1999) ‘Tennessee STAR’.
- **Managed Professional Development:** Borman, Slavin, and Cheung (2007) ‘Success for All’.
- **Tutoring:** Morrow-Howell et al. (2009) ‘Experience Corps’.
- **Curriculum:** Somers et al. (2010) ‘Enhanced Reading Opportunities’.
- **Financial Incentives:** Bettinger (2012) ‘Coshocton Incentive Programme’; and Fryer (2011) ‘New York, Dallas, Chicago’.

G. Cost-Effectiveness Analysis (J-PAL Approach)

We also used the J-PAL approach to measure the cost-effectiveness of our programme and make comparisons with other interventions (Dhaliwal et al. 2013; J-PAL 2020). We followed the sequence recommended by J-PAL to address inflation, exchange rates, and the present value of the programme costs. The programme’s local cost data in Brazilian Reals (R\$) is exchanged into US dollars using the exchange rate from the year the costs were incurred, i.e., 2022 and 2023. The 2023

costs are deflated back to the actual value in 2022 prices using the average annual US inflation rate. Then, the present value of the costs incurred in 2023 deflated to 2022 are calculated using a ten per cent discount rate. Following this, we used the average US inflation rate to deflate the total costs from 2022 to 2011. Lastly, the costs in 2011 US dollars were transformed into 2011 USD (PPP GDP).

We ran the same ATE and treatment intensity regressions (without covariates) shown in Section III.E; however, using the average of the reading and mathematics pupils' standardised scores as the outcome variable. The ATE and treatment intensity (IV) estimates using the subjects' average as the outcome are called the 'average impact' of the treatment for this cost-effectiveness analysis.

As the average impact was measured two years after the beginning of the treatment, a ten per cent discount rate should be applied as recommended by Dhaliwal et al. (2013). The total impact of the treatment is the average impact in standard deviations times the number of pupils in treated schools for the ATE. Regarding the intensity treatment estimates, we used the number of pupils in schools that changed their management by one score point due to the treatment.³⁵ The total cost in 2011 US dollars (PPP GDP) divided by the total impact in standard deviations (SD) is the cost per additional standard deviation (SD). Thus, if we divide 100 by the cost per additional SD, we have the additional SD per \$100.00, which is the cost-effectiveness indicator recommended by the J-PAL.

The 30 interventions by country are as follows presented in Table IV:

- **Kenya:** Duflo, Dupas, and Kremer (2011) 'Streaming by achievement', 'Extra contract teacher + streaming', and 'Contract teachers'; Glewwe, Kre-

³⁵It is not possible to identify which are the high implementation schools based on our causal analyses. However, it is possible to use the difference between the management in 2021 and 2023 to approximate the high implementation schools. We used the treatment schools that changed their management from 2021 to 2023 by 0.6 score points or more because, on average, these schools improved their management level by one score point.

mer, and Moulin (2009) ‘Textbooks for the top quintile’, ‘Textbooks’, and ‘Flipcharts’; Kremer, Miguel, and Thornton (2009) ‘Girls’ merit scholarships’; Duflo, Dupas, and Kremer (2015) ‘Reducing class size’; Glewwe, Ilias, and Kremer (2010) ‘Teacher incentives (year 1)’, ‘Teacher incentives (year 2)’, and ‘Teacher incentives (long-run)’.

- **Indonesia:** Pradhan et al. (2014) ‘Linking school to local govt’, ‘Electing school & linking to local govt’, ‘School committee grants’, and ‘Training school committees’.
- **India:** Banerjee et al. (2007) ‘Remedial education’; Banerjee et al. (2007) ‘Individually-paced computer assisted learning’; Banerjee et al. (2007) ‘Reducing class size’; Borkum, He, and Linden (2012) ‘Building/improving libraries’; Muralidharan and Sundararaman (2010) ‘Diagnostic feedback’; and Duflo, Hanna, and Ryan (2012) ‘Camera monitoring’.
- **Afghanistan:** Burde and Linden (2013) ‘Village-based schools’.
- **Philippines:** Abeberese, Kumler, and Linden (2014) ‘Read-a-thon’.
- **Malawi:** Baird, McIntosh, and Özler (2011) ‘Unconditional cash transfers’, and ‘Minimum cond cash transfers’.
- **Madagascar:** Lassibille et al. (2010) ‘Providing earnings information’.
- **Gambia:** Blimpo and Evans (2013) ‘School committee grants’, and ‘Grants & training for school committee’.
- **Colombia:** Barrera-Osorio and Linden (2009) ‘Adding computers to classrooms’.
- **Peru:** Cristia et al. (2017) ‘One Laptop Per Child’.

H. WMS questionnaire

2009 Education Survey Instrument

Interview Details	School and Manager's Information
<p>School ID: _____</p> <p>School Name: _____</p> <p>Interviewer Name: _____</p> <p>Date (DD/MM/YY): _____</p> <p>Time (24 hour clock): _____</p> <p>Running interview <input type="checkbox"/> Listening to interview <input type="checkbox"/></p>	<p>a) Position: _____</p> <p>b) Specialty: English <input type="checkbox"/> Maths <input type="checkbox"/> Reading <input type="checkbox"/> Science <input type="checkbox"/> Social Studies <input type="checkbox"/> None <input type="checkbox"/> Other <input type="checkbox"/></p> <p>c) If "Other", what is his/her specialty? _____</p> <p>d) Tenure in post (<i>number of years</i>): _____</p> <p>e) Tenure in school (<i>number of years</i>): _____</p> <p>f) How old is your school (<i>number of years</i>)? _____</p> <p>g) Country: _____</p> <p>h) Region: _____</p> <p>i) Number of other secondary schools within 30 minutes drive: _____</p>

Management Questions*

<p><u>1) Standardisation of Instructional Processes</u></p> <p style="text-align: center;"><i>Tests how well materials and practices are standardised and aligned in order to be capable of moving students through learning pathways over time</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) How structured or standardised are the instructional planning processes across the school?</p> <p>b) What tools and resources are provided to teachers (e.g. standards-based lesson plans and textbooks) to ensure consistent level of quality in delivery across classrooms?</p> <p>c) What are the expectations for the use of these resources and techniques?</p> <p>d) How does the school leader monitor and ensure consistency in quality across classrooms?</p>			
	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 33%; padding: 5px;"> <p>Score 1: No clear or institutionalized instructional planning processes or protocols exist; little verification or follow-up is done to ensure consistency across classrooms</p> </td> <td style="width: 33%; padding: 5px;"> <p>Score 3: School has defined instructional planning processes or protocols to support instructional strategies and materials and incorporate some flexibility to meet students needs; monitoring is only adequate</p> </td> <td style="width: 33%; padding: 5px;"> <p>Score 5: School has implemented a clearly defined instructional planning process designed to align instructional strategies and materials with learning expectations and incorporate flexibility to meet student needs; these are followed up on through comprehensive monitoring or oversight</p> </td> </tr> </table>	<p>Score 1: No clear or institutionalized instructional planning processes or protocols exist; little verification or follow-up is done to ensure consistency across classrooms</p>	<p>Score 3: School has defined instructional planning processes or protocols to support instructional strategies and materials and incorporate some flexibility to meet students needs; monitoring is only adequate</p>	<p>Score 5: School has implemented a clearly defined instructional planning process designed to align instructional strategies and materials with learning expectations and incorporate flexibility to meet student needs; these are followed up on through comprehensive monitoring or oversight</p>
<p>Score 1: No clear or institutionalized instructional planning processes or protocols exist; little verification or follow-up is done to ensure consistency across classrooms</p>	<p>Score 3: School has defined instructional planning processes or protocols to support instructional strategies and materials and incorporate some flexibility to meet students needs; monitoring is only adequate</p>	<p>Score 5: School has implemented a clearly defined instructional planning process designed to align instructional strategies and materials with learning expectations and incorporate flexibility to meet student needs; these are followed up on through comprehensive monitoring or oversight</p>		

<p>2) Personalization of Instruction and Learning</p> <p><i>Tests for flexibility in teaching methods and student involvement ensuring all individuals can master the learning objectives</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) How much does the school attempt to identify individual student needs? How are these needs accommodated for within the classroom?</p> <p>b) How do you as a school leader ensure that teachers are effective in personalising instruction in each classroom across the school?</p> <p>c) What about students, how does the school ensure they are engaged in their own learning? How are parents incorporated in this process?</p>		
<p>3) Data-Driven Planning and Student Transitions</p> <p><i>Tests if the school uses assessment to verify learning outcomes at critical stages, make data easily available and adapt student strategies accordingly</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) Is data used to inform planning and strategies? If so how is it used – especially in regards to student transitions through grades/ levels?</p> <p>b) What drove the move towards more data-driven planning/ tracking?</p>		
<p>4) Adopting Educational Best Practices</p> <p><i>Tests how well the school incorporates teaching best practices and the sharing of these resources into the classroom</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) How does the school encourage incorporating new teaching practices into the classroom?</p> <p>b) How are these learning or new teaching practices shared across teachers? What about across grades or subjects? How does sharing happen across schools (community, state-wide etc), if at all?</p> <p>c) How does the school ensure that teachers are utilising these new practices in the classroom? How often does this happen?</p>		
	<p>Score 1: Teachers lead learning with very low involvement of students; there is little or no identification of diverse student needs</p>	<p>Score 3: Teachers lead students through learning with students having some influence over their own learning</p>	<p>Score 5: Emphasis is placed on personalization of instruction based on student needs; school encourages student involvement and participation in classrooms; school provides information to and connects students and parents with sufficient resources to support student learning</p>
	<p>Score 1: School may be aware of critical transitions for students, but little or no effort is made to match support services to students; data is often unavailable or difficult to use</p>	<p>Score 3: School may understand the critical transitions points for students, although these are not identified in a consistent manner; some data is available, although not necessarily in an integrated or easy to use manner</p>	<p>Score 5: Student transitions are managed in an integrated and proactive manner, supported by formative assessments tightly linked to learning expectations; data is widely available and easy to use</p>

2009 Education Survey Instrument

<p>5) Continuous Improvement</p> <p><i>Tests attitudes towards continuous improvement</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) When problems (e.g. within school/ teaching tactics/ etc.) do occur, how do they typically get exposed and fixed?</p> <p>b) Can you talk me through the process for a recent problem that you faced?</p> <p>c) Who within the school gets involved in changing or improving process? How do the different staff groups get involved in this?</p> <p>d) Does the staff ever suggest process improvements?</p>	<p>Score 1: Exposing and solving problems (for the school, individual students, teachers, and staff) is unstructured; no process improvements are made when problems occur, or there is only one staff group involved in determining the solution</p>	<p>Score 3: Exposing and solving problems (for the school, individual students, teachers, and staff) is approached in an ad-hoc way; resolution of the problems involves most of the appropriate staff groups</p>	<p>Score 5: Exposing and solving problems (for the school, individual students, teachers, and staff) in a structured way is integral to individual's responsibilities, and resolution involves all appropriate individuals and staff groups; resolution of problems is performed as part of regular management processes</p>
<p>6) Performance Tracking</p> <p><i>Tests whether school performance is measured with the right methods and frequency</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) What kind of main indicators do you use to track school performance? What sources of information are used to inform this tracking?</p> <p>b) How frequently are these measured? Who gets to see this performance data?</p> <p>c) If I were to walk through your school, how could I tell how it was doing against these main indicators?</p>	<p>Score 1: Measures tracked do not indicate directly if overall objectives are being met; tracking is an ad-hoc process (certain processes are not tracked at all)</p>	<p>Score 3: Most performance indicators are tracked formally; tracking is overseen by the school leadership only</p>	<p>Score 5: Performance is continuously tracked and communicated, both formally and informally, to all staff using a range of visual management tools</p>
<p>7) Performance Review</p> <p><i>Tests whether performance is reviewed with appropriate frequency and follow-up</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) How often do you review (school) performance --formally or informally-- with teachers and staff?</p> <p>b) Could you walk me through the steps you go through in a process review?</p> <p>c) Who is involved in these meetings? Who gets to see the results of this review?</p> <p>d) What sort of follow-up plan would you leave these meetings with? Is there an individual performance plan?</p>	<p>Score 1: Performance is reviewed infrequently or in an un-meaningful way (e.g. only success or failure is noted)</p>	<p>Score 3: Performance is reviewed periodically with successes and failures identified; results are only communicated to senior staff members (e.g. department heads); no clear follow up/ action plan is adopted</p>	<p>Score 5: Performance is continually reviewed, based on indicators; all aspects are followed up to ensure continuous improvement; results are communicated to all staff</p>
<p>8) Performance Dialogue</p> <p><i>Tests the quality of review conversations</i></p>	<p>a) How are these review meetings structured?</p> <p>b) Do you generally feel that you do have enough data for a fact-based review?</p> <p>c) What type of feedback occurs during these meetings?</p>			

2009 Education Survey Instrument

<p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>Score 1: The right data or information for a constructive discussion is often not present or conversations overly focus on data that is not meaningful; clear agenda is not known and purpose is not stated explicitly</p>	<p>Score 3: Review conversations are held with appropriate data and information present; objectives of meetings are clear to all participating and a clear agenda is present; conversations do not, as a matter of course, drive to the root cause of the problems</p>	<p>Score 5: Regular review/ performance conversations focus on problem solving and addressing root causes; purpose, agenda and follow-up steps are clear to all; meetings are an opportunity for constructive feedback and coaching</p>
<p>9) Consequence Management</p> <p><i>Tests whether differing levels of school performance (NOT only individual teacher performance) lead to different consequences</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) Let's say you've agreed to a follow-up plan at one of your meetings, what would happen if the plan was not enacted?</p> <p>b) How long does it typically go between when a problem is identified to when it is solved? Can you give me a recent example?</p> <p>c) How do you deal with repeated failures in a specific department or area of process?</p>		
<p>10) Target Balance</p> <p><i>Tests whether the system tracks meaningful targets tied to student outcomes</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>Score 1: Failure to achieve agreed objectives does not carry any consequences</p>	<p>Score 3: Failure to achieve agreed results is tolerated for a period before action is taken</p>	<p>Score 5: A failure to achieve agreed targets drives retraining in identified areas of weakness, moving individuals to where their skills are more appropriate</p>
<p>11) Target Inter-Connection</p> <p><i>Tests whether the school and individual targets are aligned with each other and the overall system goals</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) How are these goals cascaded down to the different staff groups or to individual staff members?</p> <p>b) How are your targets linked to the overall school-system performance and its goals?</p>		
	<p>Score 1: Goals do not cascade down the throughout the school or school system</p>	<p>Score 3: Goals do cascade, but only to some staff and/ or departmental heads</p>	<p>Score 5: Goals are aligned and linked at system level and increase in specificity as they cascade, ultimately defining individual expectations for all staff groups</p>

2009 Education Survey Instrument

<p><u>12) Time Horizon of Targets</u></p> <p><i>Tests whether the school has a rational approach to planning and setting targets</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) What kind of time scale are you looking at with your targets? b) Which goals receive the most emphasis? c) Are the long-term and short-term goals set independently? d) Could you meet all your short-run goals but miss your long-run goals?</p>		
<p><u>13) Target Stretch</u></p> <p><i>Tests whether targets are appropriately difficult to achieve</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) How tough are your targets? How pushed are you by the targets? b) On average, how often would you say that you and your school meet its targets? How are your targets benchmarked? c) Do you feel that on targets all departments/ areas receive the same degree of difficulty? Do some departments/ areas get easier targets?</p>		
<p><u>14) Clarity and Comparability of Targets</u></p> <p><i>Tests how easily understandable performance measures are and whether performance is openly communicated</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>Score 1: Performance measures are complex and not clearly understood; school performance data is not made public unless mandated</p>	<p>Score 3: Performance measures are well defined and communicated; school performance data is purely quantitative but goes beyond government requirements and is made public</p>	<p>Score 5: Performance measures are well defined, strongly communicated and reinforced at all reviews; school performance data includes both quantitative and qualitative measures and are made public</p>
<p><u>15) Rewarding High Performers</u></p> <p><i>Tests whether good teacher performance is rewarded proportionately</i></p>	<p>a) How does your evaluation system work? What proportion of your employees' pay is related to the results of this review? b) Are there any non-financial or financial bonuses/ rewards for the best performers across all staff groups? How does the bonus system work (for staff and teachers)? c) How does your reward system compare to that of other schools?</p>		

2009 Education Survey Instrument

<p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>Score 1: People are rewarded in the same way irrespective of performance level</p>	<p>Score 3: There is an evaluation system which awards good performance; the system may include individual financial and non-financial awards, but these are always or never awarded</p>	<p>Score 5: There is an evaluation system which rewards individuals based on performance; the system includes both personal financial and non-financial awards; rewards are awarded as a consequence of well-defined and monitored individual achievements</p>
<p>Manager's Bonus:</p> <p>What is your bonus as a percentage of salary? _____</p>	<p>% of the bonus based on individual performance _____</p> <p>% of the bonus based on school performance _____</p> <p>% of the bonus based on district performance _____</p> <p>Refused to answer Yes <input type="checkbox"/> No <input type="checkbox"/></p> <p>Bonus on individual, school, and district performance MUST add up to 100</p>		
<p><u>16) Removing Poor Performers</u></p> <p><i>Tests whether the school is able to deal with underperformers</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) If you had a teacher who was struggling or who could not do his/ her job, what would you do? Can you give me a recent example?</p> <p>b) How long is under-performance tolerated? How difficult is it to terminate a teacher?</p> <p>c) Do you find staff members/ teachers who lead a sort of charmed life? Do some individuals always just manage to avoid being fired?</p>		
	<p>Score 1: Poor performance is not addressed or inconsistently addressed; poor performers are rarely removed from their positions</p>	<p>Score 3: Poor performance is addressed, but typically through a limited range of methods (e.g. coaching); the process of terminating an employee often takes more than a year to complete and is therefore infrequent, even under conditions of repeated poor performance</p>	<p>Score 5: Repeated poor performance is addressed, beginning with targeted interventions; poor performers are moved out of the school when weaknesses cannot be overcome</p>
<p><u>17) Promoting High Performers</u></p> <p><i>Tests whether promotions and career progression are based on performance</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) Can you tell me about your career progression/ promotion system?</p> <p>b) How do you identify and develop your star performers?</p> <p>c) What types of professional development opportunities are provided? How are these opportunities personalised to meet individual teacher needs?</p> <p>d) How do you make decisions about promotion/ progression and additional opportunities within the school, such as performance, tenure, other? Are better performers likely to be promoted faster, or are promotions given on the basis of tenure/ seniority?</p>		
	<p>Score 1: Staff members are promoted primarily upon the basis of tenure (e.g. years of service)</p>	<p>Score 3: Staff members are promoted upon the basis of performance; school provides career opportunities but usually based on non-performance related factors</p>	<p>Score 5: School actively identifies, develops and promotes its top performing staff members</p>
<p><u>18) Managing Talent</u></p> <p><i>Tests how well the school identifies and targets needed teaching, leadership and other capacity in the school</i></p>	<p>a) How do school leaders show that attracting talented individuals and developing their skills is a top priority?</p> <p>b) How do you ensure you have enough teachers of the right type in the school?</p> <p>c) Where do you seek out and source teachers?</p> <p>d) What hiring criteria do you use?</p>		

2009 Education Survey Instrument

Score: 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/>	Score 1: School has very limited or no control over the number and types of teachers, staff and leadership needed to meet goals	Score 3: School reactively controls the number and types of teachers, staff and leadership needed to meet goals; school may define hiring criteria and processes, but they are not linked with key drivers of student outcomes	Score 5: School proactively controls the number and types of teachers, staff and leadership needed to meet goals; school defines hiring criteria and processes based on understanding of what drives student achievement
<u>19) Retaining Talent</u> <i>Tests whether the school will go out of its way to keep its top talent</i>	a) If you had a top performing teacher who wanted to leave, what would the school do? b) Could you give me an example of a star performer being persuaded to stay after wanting to leave? c) Could you give me an example of a star performer who left the school without anyone trying to keep him?		
Score: 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/>	Score 1: We do little to try and keep our top talent	Score 3: We usually work hard to keep our top talent	Score 5: We do whatever it takes to retain our talent
<u>20) Attracting Talent/ Creating a Distinctive Employee Value Proposition</u> <i>Tests how strong the teacher value proposition is to work in the individual school</i>	a) What makes it distinctive to teach at your school, as opposed to other similar schools? If you were to ask the last three candidates would they agree? Why? b) How do you monitor how effectively you communicate your value proposition and the following recruitment process?		
Score: 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/>	Score 1: Other schools offer stronger reasons for talented people to join	Score 3: Our value proposition to those joining our school is comparable to those offered by other schools	Score 5: We provide a unique value proposition to encourage talented people join our school above our competitors
Leadership Questions*			
<u>21) Leadership Vision</u> <i>Tests whether school leaders have an understanding of the broader set of challenges that the school, system and key actors face and the right mindset to address them</i>	a) What is the school's vision for the next five years? Do teachers/ staff know and understand the vision? b) Who does your school consider to be your key stakeholders? How is this vision communicated to the overall school community? c) Who is involved in setting this vision/ strategy? When there is disagreement, how does the school leader build alignment?		
Score: 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/>	Score 1: School either has no clear vision, or one defined without substantial stakeholder collaboration and which focuses primarily on meeting state/ national mandates; school leader does not or cannot articulate a clear focus on building an environment conducive to learning	Score 3: School has defined a vision that focuses on improvement in student outcomes, but largely focused on meeting state/ national mandates, and usually defined with limited stakeholder collaboration; school leaders may focus on the quality of the overall school environment, but often in response to specific issues	Score 5: School leaders define and broadly communicate a shared vision and purpose for the school that focuses on improving student learning and outcomes (often beyond those required by law); vision and purpose is built upon a keen understanding of student and community needs, and defined collaboratively with a wide range of stakeholders; school leader proactively builds environment conducive to learning

<p>22) Clearly Defined Accountability for School Leaders</p> <p><i>Tests whether school leaders are accountable for delivery of student outcomes</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) Who is accountable for delivering on school targets? b) How are individual school leaders held responsible for the delivery of targets? Does this apply to equity and cost targets as well as quality targets? c) What authority do you have to impact factors that would allow them to meet those targets (e.g. budgetary authority, hiring & firing)? Is this sufficient?</p>		
<p>23) Clearly Defined Leadership and Teacher Roles</p> <p><i>Tests how clearly the roles, responsibilities and required attributes of teachers, students and staff are defined within the school</i></p> <p>Score:</p> <p>1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> -99 <input type="checkbox"/></p>	<p>a) How are the roles and responsibilities of the school leader defined? How are they linked to student outcomes/ performance? b) How are leadership responsibilities distributed across individuals and teams within the school? c) How are the roles and responsibilities of the teachers defined? How clearly are required teaching competences defined and communicated? d) How are these linked to student outcomes/ performance?</p> <p>Score 1: School does not define clear roles, responsibilities and desired competencies of school leaders and teachers</p>	<p>Score 3: School defines clear roles, responsibilities and desired competencies of school leaders and teachers, but not necessarily linked with the drivers of student performance and outcomes; concentrated leadership amongst senior staff</p>	<p>Score 5: School defines clear roles, responsibilities and desired competencies of teachers and staff across the school, built upon an understanding of what drives student performance and outcomes; leadership responsibilities are distributed across the school</p>
<p>Organization Questions</p>			
<p>a) How many students are in the school? _____</p> <p>b) How many teachers are in the school? _____</p> <p>c) How many people (including support staff) work in the school? _____</p> <p><i>Please say "Can you walk me through the school's hierarchy?". Then iteratively ask "Who does a teacher report to?", "Who would [his/her boss] report to"...., Keep asking until you reach the School Head.</i></p> <p>d) Number of levels in the school BETWEEN the teacher and the School Head: _____</p> <p>e) How many people directly report to the head of the school (i.e. the number of people directly in the hierarchical layer below him/her)? _____</p>			

f) To hire a FULL-TIME TEACHER what agreement would your school head need?

Score:

1 2 3 4 5 -99

Score 1: The school has no authority

Score 3: Requires sign-off from above the school head based on the individual case. Typically agreed (i.e. about 80 or 90% of the time)

Score 5: Complete authority of the school head

g) To add a new class - for example, introducing a new language such as Mandarin - what agreement would the school head need?

Score:

1 2 3 4 5 -99

Score 1: The school has no authority

Score 3: Requires sign-off from above the school head based on the individual case. Typically agreed (i.e. about 80 or 90% of the time)

Score 5: Complete authority of the school head

h) To expand the school size - for example admitting 5% more students - what agreement would the school head need?

Score:

1 2 3 4 5 -99

Score 1: The school has no authority

Score 3: Requires sign-off from above the school head based on the individual case. Typically agreed (i.e. about 80 or 90% of the time)

Score 5: Complete authority of the school head

i) Do you use admissions criteria to select students?

Yes No -99

j) Can you take me through the criteria you use to select students?

Academics

Geographical

Siblings

Other

If other, what? _____

k) Who determines these criteria?

Score:

1 2 3 4 5 -99

Score 1: School or school board has NO authority to set the admission criteria (mandated by external authorities)

Score 3: School or school board has shared authority with external authorities to set the admissions criteria

Score 5: School or school board has complete authority to set the admissions criteria

l) What is the largest CAPITAL INVESTMENT the school leader can make without PRIOR authorization from outside? (ignore form filling) [PLEASE CROSS CHECK ANY ZERO RESPONSE BY ASKING "what about buying a new computer - would that be possible?", and then probe further. _____

UK only:

m) Approximately, how many other 'competing' schools provide teaching for a similar age group (public and private schools) within your catchment area? _____

Ownership

a) What type of school is it? _____

b) Is the school state owned or non-state owned?

State owned Non-state owned Other -99

If other, who? _____

c) Is the school for-profit or not-for-profit?

For profit Not for profit -99

If other, who? _____

d) Does the school have a religious affiliation – if so with what religion?

Not religious Anglican Catholic Hindu

Jesuit Jewish Mormon Muslim

Protestant Other

If other, who? _____

Human Resources

a) Percent of teachers who are union members _____

If the question above is equal to 100, then the question below is also equal to 100.

Anywhere in between, ensure answer is provided

b) Percent of teachers whose pay is set by union negotiations _____

c) Average classroom teaching hours per week by teachers _____

d) Average actual hours worked per week by teachers (including time at home) _____

e) Percent of teachers who have left in the past 12 months _____

f) Roughly how many times bigger is the school leader's salary than a starting teacher's salary. That is, does the school head earn twice as much, ten times as much, or 100 times as much?

Refused to answer: Yes No

g) Ignoring yourself, how well managed do you think the rest of the school is on scale: 1 to 10, where 1 is worst practice, 10 is best practice and 5 is average

Overall _____

Operations _____
(teaching practices, student transitions)

Talent _____
(people, promotions, incentives, etc.)

Would you like me to send you a copy of this report when it is

written? Yes No

Post - Interview

a) Interview duration (minutes) _____

b) Interviewee knowledge of management practices

Score:

1 2 3 4 5

Score 1: Some knowledge his school, and no knowledge of its daily operations

Score 3: Expert knowledge of his school, and some knowledge of its daily operations

Score 5: Expert knowledge about his school and its daily operations

c) Interviewee willingness to reveal information

Score:

1 2 3 4 5

Score 1: Very reluctant to provide more than basic information

Score 3: Provides all basic information and some more confidential information

Score 5: Totally willing to provide any information about the school

d) Interviewee patience

Score:

1 2 3 4 5

Score 1: Little patience - wants to run the interview as quickly as possible. I felt heavy time pressure

Score 3: Some patience - willing to provide richness to answers but also time constrained. I felt moderate time pressure

Score 5: Lot of patience - willing to talk for as long as required. I felt no time pressure.

e) Attitude on the government (if mentioned)

Score:

1 2 3 4 5

Score 1: Government seen entirely as a hindrance - bad for the school

Score 3: Government helps the school in some ways but also a constraint in other ways - mixed for the school

Score 5: Government helps the school - good for the school

f) Number of times mentioned overriding economic factors (e.g. recession)? _____

g) Number of times rescheduled (0=never rescheduled) _____

h) Seniority of interviewee

1 - Superintendent/Governor/Director/ Father 2 - Principal/ Head Teacher/ Head Master

3 - Assistant Principal/ Vice Principal/ Deputy Head/ Curriculum Coordinator

4 - Department Head/ Subject Coordinator 5 - Teacher

i) Age of interviewee (don't ask) - guess if not told _____

j) Gender of interviewee Male Female

k) Did the interviewee have a degree - guess if not told _____

l) Interview language _____

*The Management and Leadership questions were asked in the following order during the interview: 21,1,2,3,4,5,6,7,8,9,10,11,12,13,22,23,14,15,16,17,18,19,20.

I. Regression Adjustment

For robustness check, we also identify the causal effect of the treatment on the school management using a full regression adjustment model with a centred covariate and its interaction with the treatment dummy (Imbens and Rubin 2015; Lin 2013; Negi and Wooldridge 2021; Sloczynski 2022). Let X_s be the school management average score from the management survey conducted in 2021. Thus, the ATE, λ_{ATE} , can be identified for each subject using the following weighted full regression adjustment causal equation:

$$M_s\sqrt{n_s} = \iota\sqrt{n_s} + \lambda_{ATE}Z_s\sqrt{n_s} + (X_s - \bar{X})\gamma\sqrt{n_s} + Z_s(X_s - \bar{X})\delta\sqrt{n_s} + \varepsilon_s\sqrt{n_s}. \quad (5)$$

As in the unadjusted Equation 1, the adjusted Equation 5 identifies the average treatment effect (ATE) of the programme on the management of the schools. Standard errors are robust standard errors (Chaisemartin and Ramirez-Cuellar 2024). Equation 5 can be used to identify the ATE of the programme on a specific management practice or even in a group of management practices.

Also, for robustness check, we identify the ATE of the programme on pupils' learning using a full regression adjustment model with centred pre-treatment covariates and their interactions with the treatment dummy (Imbens and Rubin 2015; Lin 2013; Negi and Wooldridge 2021; Sloczynski 2022). Let \mathbf{X}_{isp} be the pupils' performance in three previous Rio Assessments applied in April, June and September 2021. Let $\bar{\mathbf{X}}$ be the average of \mathbf{X}_{isp} . Therefore, the ATE, β_{ATE} , can be identified for each subject using the following causal equation:

$$Y_{isp} = \alpha + \beta_{ATE}Z_{isp} + (\mathbf{X}_{isp} - \bar{\mathbf{X}}) \cdot \boldsymbol{\gamma} + Z_{isp} \cdot (\mathbf{X}_{isp} - \bar{\mathbf{X}}) \cdot \boldsymbol{\delta} + \epsilon_{isp}. \quad (6)$$

As in the unadjusted Equation 2, the ATE identification and estimation in the adjusted Equation 6 relies on using the clustered robust standard errors with clusters defined at the school pair level plus the non-inclusion of pair-fixed effects in the regression (Chaisemartin and Ramirez-Cuellar 2024).

Yet, for robustness check, we also identify the causal effect of the intensity of management provided by the treatment using a full regression adjustment model with centred pre-treatment covariates and their interactions with the treatment dummy (Imbens and Rubin 2015; Lin 2013; Negi and Wooldridge 2021; Sloczynski 2022). Let \mathbf{X}_{isp} be the set of pre-treatment covariates: the pupils' performance in three previous Rio Assessments applied in April, June and September 2021, and the school's average score regarding the 23 management practices surveyed by our pre-treatment survey. Let $\bar{\mathbf{X}}$ be the average of \mathbf{X}_{isp} . Thus, the impact of the treatment intensity (school management changes driven by the random assignment/the programme) using a full interaction adjustment model can be identified from the following Two-Stage Least Square (2SLS) procedure that uses the random assignment to treatment as the instrumental variable (IV):

$$Y_{isp} = \alpha + \beta_{IV} \hat{M}_{isp} + (\mathbf{X}_{isp} - \bar{\mathbf{X}}) \cdot \boldsymbol{\gamma} + Z_{isp} \cdot (\mathbf{X}_{isp} - \bar{\mathbf{X}}) \cdot \boldsymbol{\delta} + \epsilon_{isp}, \quad (7)$$

where the \hat{M}_{isp} is the fitted M_{isp} from the following first-stage equation

$$M_{isp} = \omega + \nu Z_{isp} + (\mathbf{X}_{isp} - \bar{\mathbf{X}}) \cdot \boldsymbol{\nu} + Z_{isp} \cdot (\mathbf{X}_{isp} - \bar{\mathbf{X}}) \cdot \boldsymbol{\theta} + \xi_{isp}.$$

Standard errors are clustered robust standard errors, with clusters set to be the school pairs used for the random assignment and on the non-inclusion of pair-fixed effects in the regression (Chaisemartin and Ramirez-Cuellar 2024).

J. Additional Information about the Management Groups of Practices' Impact

Table A8 gives more information on the treatment's impact across the five management practice groups. Column (1) reports average treatment effects (ATE) estimates from weighted regressions of the 2023 group of practices' average scores on a treatment dummy without baseline covariates. Column (2) shows ATE estimates generated from weighted regressions of the 2023 group of practices' average scores on a treatment dummy, the centred average score reached by the schools in the same group of practices in 2021, and the interaction between the centred baseline covariate and the treatment dummy. The weights used are the number of grades 1-8 pupils in each school, as discussed in Section III.E. Standard errors are reported in parentheses. They are robust standard errors. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

In the unadjusted regression, Table A8 reports the treatment's impact was 1.530 standard deviations (SD) for the target-setting group of practices and 1.069 SD for the leadership group of managerial practices. Both estimates are statistically significant at the 1% level. The programme's causal effect on the operation group of management practices was 0.370 SD in the unadjusted regression, but it is not statistically significant. However, the ATE estimate of the impact of the treatment on the operation group of management practices reaches 0.400 SD and is statistically significant at the 10% level in the adjusted regression.

Table A8 reports that the ATE estimate regarding the monitoring group of practices was 0.324 SD, but it is not statistically significant. People management did not change due to the treatment.

Table A8: The Impact of the Treatment (ATE) on the Management Groups

	No covariate	With covariate
	(1)	(2)
	ATE	ATE
	OLS	RA
Target setting (SD)	1.530*** (0.364)	1.528*** (0.368)
Leadership (SD)	1.069*** (0.278)	1.079*** (0.272)
Operation (SD)	0.370 (0.223)	0.400* (0.217)
Monitoring (SD)	0.324 (0.232)	0.332 (0.232)
People management (SD)	0.054 (0.227)	0.000 (0.000)

Note: The table shows the causal effects of the treatment across five different groups of management practices: target setting, leadership, operation, monitoring, and people management. These groups are formed by the 23 best management practices discussed in Bloom et al. (2015). The management level of each of the 80 experimental schools regarding the 23 best management practices was measured by a survey conducted in December 2023 that followed the WMS methodology. Each practice was scored against a scoring grid that goes from one, worst management, to five, the best management. A school's average score in a group of management practices is calculated by averaging the scores received regarding the practices that belong to that specific group. The school's average score was standardised using the control group as the reference. Column (1) reports average treatment effects (ATE) estimates from regressions of the 2023 group of practices' average scores on a treatment dummy without baseline covariates. Column (2) shows ATE estimates generated from regressions of the 2023 group of practices' average scores on a treatment dummy, the centred average score reached by the schools in the same group of practices in 2021, and the interaction between the centred baseline covariate and the treatment dummy. Standard errors are reported in parentheses. They are robust standard errors. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

K. Additional Information on Robustness Checks

Table A9 shows the causal effects of the treatment on pupils learning under different specifications. All the outcome variables were generated in December of 2023, i.e., two years after the beginning of the programme implementation. The sample includes all pupils (31,760) enrolled in grades 1 to 8 at the beginning of 2022 in one of the 80 sample schools randomly selected from the Rio de Janeiro population of schools. These pupils registered to a sample school in the regular period of enrolment, November and December 2021, before the experiment began. The pupils' reading and mathematics scores from the December 2023 Rio Assessment represent their learning. The scores were standardised to have a mean of zero and a standard deviation of one in each grade and subject regarding the control group. 'Mgmt ([1,5])' represents the schools' average management score from one, worst management, to five, the best management. 'Mgmt ([1,5])' comes from our survey conducted in December 2023 that adopted the WMS methodology.

From 31,760 sample pupils, Table A9 shows results from analyses with 31,710 and 31,706 valid pupils' scores for reading and mathematics, respectively. Columns (1) and (3) report average treatment effects (ATE) estimates from regressions without and with baseline covariates, respectively. Columns (2) and (4) show IV estimates generated from regressions without and with baseline covariates, respectively. IV estimates can be interpreted as the causal effects of a one-score point change (on a management scale that goes from one, worst management, to five, best management) in school management, driven by the random assignment/treatment, on pupils' educational outcomes.

Table A9: The Treatment Causal Effects - Alternative Pupils' Scores (CTT)

	No covariates		With covariates	
	(1)	(2)	(3)	(4)
	ATE	Mgmt ([1,5])	ATE	Mgmt ([1,5])
	OLS	IV	RA	IV/RA
Reading (SD)	0.209*** (0.055)	0.630*** (0.252)	0.174*** (0.042)	0.522*** (0.192)
Mathematics (SD)	0.220*** (0.053)	0.662*** (0.247)	0.194*** (0.044)	0.583*** (0.201)

Note: The table shows the impact of the treatment, SMEP, under different specifications. All the outcome variables were generated in December 2023, two years after the beginning of the treatment. The sample includes all pupils (31,760) enrolled in grades 1 to 8 at the beginning of 2022 in one of the 80 sample schools randomly selected from the Rio de Janeiro population of schools. These pupils registered to a sample school in the regular period of enrolment, November and December 2021, before the experiment began. The pupils' reading and mathematics scores from the December 2023 Rio Assessment represent their learning. These scores are based on the Classical Test Theory (CTT). The scores were standardised to have a mean of zero and a standard deviation of one in each grade and subject regarding the control group. 'Mgmt ([1,5])' represents the schools' average management score from one, worst management, to five, the best management. 'Mgmt ([1,5])' comes from our survey conducted in December 2023 that adopted the World Management Survey (WMS) methodology. From 31,760 sample pupils, there are 31,710 and 31,706 valid pupils' scores for reading and mathematics, respectively. Columns (1) and (3) report average treatment effects (ATE) estimates from regressions without and with baseline covariates, respectively. Columns (2) and (4) show IV estimates generated from regressions without and with baseline covariates, respectively. IV estimates can be interpreted as the causal effects of a one-score point change (on a management scale that goes from one, worst management, to five, best management) in school management, driven by the random assignment/treatment, on pupils' educational outcomes. Columns (3) and (4) report ATE and IV estimates from regressions with the following centred baseline covariates (and their interaction with treatment): pupils' reading and mathematics scores from two different Rio Assessments that were applied in September and December 2021, and schools' average management score from our 2021 management survey. Standard errors are reported in parentheses. They are clustered robust standard errors. The clusters are the school pairs used for the random assignment procedure. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Columns (3) and (4) from Table A9 report ATE and IV estimates from regressions with the following centred baseline covariates (and their interaction with treatment): pupils' reading and mathematics scores from three different Rio Assessments that were applied in April, June and September 2021, and schools' average management score from our 2021 management survey. Standard errors are reported in parentheses. They are clustered robust standard errors. The clusters are the school pairs used for the random assignment procedure. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

The results based on IRT scores from the Rio Assessments reported in Table III are similar to the estimates reported in Table A9 that are based on CTT scores from the same Rio Assessments. The ATE estimates from unadjusted regressions are 0.209 standard deviations (SD) for reading and 0.220 for mathematics, and the estimates from adjusted regressions are 0.173 SD for reading and 0.192 SD for mathematics. The IV estimates from unadjusted regressions are 0.630 SD for reading and 0.662 SD for mathematics, and the estimates from adjusted regressions are 0.522 SD for reading and 0.583 SD for mathematics. All the results are statistically significant at the 1% level.

Table A10: Sharpened False Discovery Rate (FDR) q-values

Location	Identifier	p-values	bky06_qval
Table IX - Unadjusted	Reading (ATE/OLS)	0.000	0.002
	Maths (ATE/OLS)	0.000	0.002
	Management (ATE/OLS)	0.001	0.003
	Reading (IV)	0.005	0.009
	Maths (IV)	0.007	0.011
Table IX - Adjusted	Reading (ATE/RA)	0.000	0.002
	Maths (ATE/RA)	0.000	0.002
	Management (ATE/RA)	0.001	0.002
	Reading (IV/RA)	0.002	0.005
	Maths (IV/RA)	0.003	0.007
IV.A Absence	Absence 22	0.061	0.047
	Absence 23	0.000	0.002
Table X - Unadjusted	Target 23 (ATE/OLS)	0.000	0.002
	Leadership 23 (ATE/OLS)	0.000	0.002
Table X - Adjusted	Target 23 (ATE/RA)	0.000	0.002
	Leadership 22 (ATE/RA)	0.000	0.002
Figure XIV	Planning standardisation (o1)	0.828	0.43
	Instruction personalisation (o2)	0.362	0.248
	Data-driven planning (o3)	0.153	0.109
	Adopting best practices (o4)	0.030	0.027
	Continuous improvement (m5)	0.588	0.356
	Performance tracking (m6)	0.028	0.026
	Performance review (m7)	0.189	0.127
	Performance dialogue (m8)	0.717	0.407
	Consequence management (m9)	0.928	0.46
	Target balance (t10)	0.000	0.001
	Target interconnection (t11)	0.000	0.002
	Target time horizon (t12)	0.001	0.003
	Target stretch (t13)	0.022	0.021

Table A10: Sharpened False Discovery Rate (FDR) q-values

Location	Identifier	p-values	bky06_qval
	Target clarity/comparability (t14)	0.008	0.012
	Rewarding high performers (p15)	0.933	0.46
	Fixing poor performers (p16)	0.520	0.327
	Promoting high performers (p17)	1.000	0.491
	Managing talent (p18)	0.201	0.134
	Retaining talent (p19)	0.409	0.28
	Attracting employees (p20)	0.228	0.151
	Leadership vision (l21)	0.000	0.002
	Leadership accountability (l22)	0.034	0.029
	Clearly defined roles (l23)	0.150	0.109
Figure XV	Apr-22	0.013	0.016
	Jun-22	0.017	0.019
	Sep-22	0.021	0.021
	Dec-22	0.015	0.017
	Apr-23	0.001	0.004
	Jun-23	0.001	0.002
	Sep-23	0.000	0.002
	Dec-23	0.000	0.002
IV.C Heterogeneous	Management 21	0.004	0.007
	Segments (I or II)	0.768	0.421
	Grade 2	0.743	0.419
	Grade 3	0.756	0.42
	Grade 4	0.433	0.289
	Grade 5	0.973	0.48
	Grade 6	0.572	0.351
	Grade 7	0.845	0.431
	Grade 8	0.923	0.46
	School size	0.425	0.288
	Female	0.865	0.437
	Brown	0.304	0.207
	Black	0.569	0.351

Table A10: Sharpened False Discovery Rate (FDR) q-values

Location	Identifier	p-values	bky06_qval
	Asian/Indigenous	0.526	0.327
	Bolsa Familia	0.832	0.43
	Quartil 2	0.365	0.248
	Quartil 3	0.179	0.122
	Quartil 4	0.657	0.387
Table XI - Unadjusted	Reading (ATE/OLS)	0.001	0.002
	Maths (ATE/OLS)	0.000	0.002
	Reading (IV)	0.012	0.016
	Maths (IV)	0.007	0.011
Table XI - Adjusted	Reading (ATE/RA)	0.000	0.002
	Maths (ATE/RA)	0.000	0.002
	Reading (IV/RA)	0.007	0.011
	Maths (IV/RA)	0.004	0.007
IV.E Robustness II	Reading (ATE/RA)	0.001	0.002
	Maths (ATE/RA)	0.000	0.002
	Reading (IV/RA)	0.008	0.012
	Maths (IV/RA)	0.009	0.012
V - CBA/CEA	Learning (ATE/OLS)	0.000	0.002
	Learning (IV)	0.005	0.009

Notes: This table presents information on the p-values generated by this study's analyses and their corresponding sharpened False Discovery Rate (FDR) q-values. The FDR is the expected proportion of type I errors (false rejections) (Anderson 2008).

L. Management and Education

Economists have been sceptical about the importance of management due to the difficulty of measuring management and the belief that profit-oriented firms try to minimise costs and any variations in management practices reflect optimal

responses to different market conditions (Bloom et al. 2012a). Despite scepticism, a growing literature in Economics has discussed the relationship, from non-causal to causal methodology studies, between specific management practices and productivity, including the connection between management practices and educational outcomes (Angrist et al. 2010; Angrist, Pathak, and Walters 2013; Barros et al. 2019, 2021; Beg, Fitzpatrick, and Lucas 2023; Bloom, Sadun, and Van Reenen 2016; Bloom and Van Reenen 2007, 2010; Bloom et al. 2020, 2012a, 2014, 2012b, 2019; Bloom et al. 2015; Bruhn, Karlan, and Schoar 2018; Dobbie and Fryer 2013; Fryer 2017; Gosnell, List, and Metcalfe 2020; Hoyos, Ganimian, and Holland 2019; Muralidharan and Singh 2020; Romero et al. 2022; Tavares 2015).

However, the causal literature on the impact of management on productivity provides some unclear and conflicting evidence. On one hand, Abdulkadiroğlu et al. (2011), Angrist et al. (2010, 2012), Barros et al. (2019, 2021), Beg, Fitzpatrick, and Lucas (2023), Bloom et al. (2020, 2012a), Bruhn, Karlan, and Schoar (2018), Curto and Fryer (2014), Dobbie and Fryer (2011), Fryer (2017), Fryer (2014), Gosnell, List, and Metcalfe (2020), and Tavares (2015) reported positive connections between management and productivity. On the other hand, Hoyos, Ganimian, and Holland (2019), Muralidharan and Singh (2020), and Romero et al. (2022) have shown that management may not affect productivity.

Abdulkadiroğlu et al. (2011), Angrist et al. (2010, 2012), Curto and Fryer (2014), Dobbie and Fryer (2011), and Fryer (2014) have investigated not exactly the impact of management but the impact of US charter schools on pupils' performance. These studies are not focused on analysing the impact of management on pupils' educational results. They have a focus on the effects of charter schools on educational outcomes. These studies provide evidence of charter schools as a package of features where one cannot extract separate causal evidence about

‘management’.

For instance, Fryer (2014) conducted a randomised field experiment in which the treatment was to transform 20 of the lowest-performing traditional schools in Houston (Texas) into charter schools. Among many changes conducted in the treatment schools, almost all principals and half of the teachers were changed before the experiment started, and pupils were exposed to at least 20% more teaching time. It is impossible to disentangle, for instance, the effect of management practices from the impact of the changes in the school staff. One can not either extract the causal effect of management from the increase of more than 20% of school time.

The main focus of the mentioned studies discussing charter schools was not on analysing the causal relationship between management and educational productivity but on investigating the impact of the charter school models on pupils’ achievement. Logically, one of the many factors differentiating charter schools from other schools may be their management practices. However, these studies do not provide causal evidence about the effects of specific management practices on pupils’ educational results.

The first significant contribution to the discussion on the positive effect of management on productivity is the randomised field experiment conducted by Bloom et al. (2012a). Management consulting was provided to randomly selected plants within large multi-plant Indian textile firms. The consulting goal was to introduce 38 standard management practices in productive manufacturing firms. The consulting involved diagnosing areas with potential for improvement and supporting firms as they implemented the new procedures. The treatment increased the plants’ productivity by 17% compared to the control plants in the first year. Also, the treatment plants opened more production plans in three years than the

control plants. Moreover, even after nine years, there were still differences in managerial practices between the treatment and control groups (Bloom et al. 2020). Management last!

A randomised field experiment conducted by Bruhn, Karlan, and Schoar (2018) involving 432 small and medium enterprises in Mexico showed that treatment companies increased productivity, return on assets, and "entrepreneurial spirit" among the owners, which measures entrepreneurial confidence and goal setting. Moreover, using Mexican social security data, the researchers observed a persistent 50% increase in the number of employees and total wage bills even 5 years after the intervention.

Another relevant study on the positive causal effects of management on productivity is the field experiment conducted by Gosnell, List, and Metcalfe (2020) that showed that specific management practices such as monitoring and target-setting significantly increased aeroplane captains' productivity on the targeted fuel-saving, reducing the CO2 emissions. Moreover, treatment captains reported higher job satisfaction.

Regarding education, only a few studies discussed the positive effects of management on pupils' educational achievement (Barros et al. 2019, 2021; Beg, Fitzpatrick, and Lucas 2023; Fryer 2017; Tavares 2015). One is the study developed by Fryer (2017) to investigate the causal effect of principals' management training on pupils' educational achievement through a randomised field experiment involving 58 public schools with the worst educational performance from Houston, Texas. The experiment started with principals from 28 schools being randomly assigned to receive 300 hours of training on lesson planning, data-driven instruction, and teacher observation and coaching. This management training was based on a book written by Bambrick-Santoyo (2018) and the best management practices discussed

by the World Management Survey (WMS). There was no staff change or increased teaching time.

The results reported in the main tables of Fryer (2017) are positive and statistically significant. However, the estimates from the appendix tables are close to zero. The issue is that the main tables are based on analyses that rely on an outcome variable that sums the subjects (i.e., mathematics + reading), while in the appendix, the analyses were conducted separately by subject. Using the outcome as the sum of the scores of different subjects makes the estimates difficult to interpret and compare. Thus, based on the analysis in the appendix tables, the ITT estimates from a ‘pooled’ analysis were around 0.03 SD for mathematics and 0.046 SD for reading regarding high-stakes test scores. ITT estimates also from a ‘pooled’ analysis for low-stakes test scores were around 0.04 across subjects (mathematics, science, reading and social studies). The effects are slightly better for principals predicted to be high implementers, around 0.06 SD by subject, and for principals staying for two years in treatment schools, around 0.05 SD also by subject.

A randomised field experiment conducted by Beg, Fitzpatrick, and Lucas (2023) claims to have shown the positive impact of management on pupils’ learning. The working paper shows the impact of two treatments: a) teacher training on the Differentiated Instruction (DI) programme plus a monitoring task called ‘management effort’, and b) the same teacher training on DI plus the same ‘management effort’ plus ‘people management’. The problem here is the same as encountered with the charter schools studies: it is impossible to disentangle the causal effects of ‘management’ on pupils’ learning since this ‘management effort’ is mixed with the teacher training on Differentiated Instruction (DI) in both treatment arms. The impact on pupils’ learning can be due to teacher training in the Differentiated

Instruction programme. Therefore, this study also does not provide convincing evidence in favour of the positive effects of management on pupils' educational results.

Three large-scale randomised field experiments conducted by Barros et al. (2019, 2021) with 1,400 state-run high schools (upper secondary schools) across nine Brazilian states analysed the impact of the Jovem de Futuro Programme on schools' educational performance. The programme was delivered in three sequential steps. Each programme step lasted three years and reached different federated states. Also, the programme had substantial changes at each step (Barros et al. 2019, 2021). The first step of the Jovem de Futuro Programme was focused on changing the management mindset and culture at the school level. The programme was directly delivered by the nongovernmental (NGO) Unibanco Institute. Financial incentives were provided to treatment schools.³⁶ No management protocols were part of the programme.

From the second step onward, the programme was implemented by the participating education secretariats with support from the Unibanco Institute. In the second step, management protocols and training were provided for school managers. The third step was characterised by an expansion of the programme to reach higher levels of coordination: regional and central coordination of the education secretaries. Management training, protocols, and support were provided for schools and regional and central coordination of the education secretaries. From the third step, no financial incentives were provided to treatment schools.

Despite the changes to improve the Jovem de Futuro Programme at each step of three years of implementation, the intervention's impact was statistically significant and the same across the steps, i.e., approximately 0.1 SD for reading and

³⁶<https://www.institutounibanco.org.br/>

mathematics. On one side, one can say that the impact of the programme is robust across different populations and settings. On the other hand, since each programme step reached the same impact, it is not clear which of the practices implemented are the cause of the effects: financial incentives (1st and 2nd steps), focus on changing the management mindset and culture without management protocols (1st step), management protocols and training to schools (2nd and 3rd steps), management protocols and training reaching regional and central coordination (2nd and 3rd steps), focus on schools only (1st and 2nd steps) or focus on three levels: schools, regional and central coordination (3rd step). Perhaps improving the regional or central coordination management level did not affect pupils' educational outcomes. Maybe the management protocols implemented did not work. Maybe the programme impact was caused by the effect of being observed.

A relevant study based on a fuzzy regression discontinuity design (half fuzzy RDD) conducted by Tavares (2015) showed the positive impact of specific management practices on pupils' learning. The research shows that specific management practices delivered by the Results-based School Management Programme, a management policy developed by the São Paulo Federated State, Brazil, improved between 0.14 and 0.22 SD 8th-grade pupils' educational performance in mathematics. However, there was no significant impact on pupils' reading performance. The management practices delivered by the programme involve management training, strategic planning, goal setting and goal management. Cost information was not available.

A large-scale experimental evaluation conducted by Muralidharan and Singh (2020) shows no effect of randomly providing 1,774 schools with comprehensive assessments, detailed school ratings, and customised school improvement plans. Notably, the programme did not focus on directly changing the schools' manage-

ment through support or accountability measures. As a result, the authors report there was no increase in oversight or accountability in schools due to the treatment. The study also reports no changes in school management and pupils' achievement. This experiment is a remarkable example of managers not necessarily applying the relevant information provided to them. Management programmes that focus only on providing relevant information or training to school managers may not succeed in changing organisations' management practices.

An experiment conducted by Hoyos, Ganimian, and Holland (2019) in La Rioja, Argentina, reports a large effect of providing schools with pupils' educational diagnostic feedback reports and some support (treatment one). However, the study shows no effect (compared to treatment one) in providing schools with professional development workshops (on school management) and school visits (treatment two) together with diagnostic feedback and some support. The authors suggest that the lack of statistically significant differences between more support on school management (treatment two) and less support (treatment one) would represent a null effect of more 'management' on pupils' learning. However, the issue with this experiment is that the difference between treatments one and two is very weak regarding management. The support given to schools from treatment two, not delivered to schools from treatment one, was only one visit per year, two workshops in the first year and three in the second year. Based on Bloom et al. (2015) and Bloom et al. (2012a), we suppose that only large changes in the general management of an organisation can affect productivity consistently.

A large-scale experiment conducted in Mexico by Romero et al. (2022) provided schools with two random treatments: direct management training and indirect (cascade-style: 'train the trainer' model) training. The study shows that direct training, compared to indirect, improved by 0.13 SD in managerial capacity but

did not affect pupils' educational achievement. Again, we suppose that the change in management was not enough to affect pupils' educational outcomes. Bloom et al. (2015) showed that an increase of one standard deviation (SD) change in management is correlated with an improvement of 0.24 SD in pupils' educational results.

Researchers from the World Management Survey (WMS)³⁷ have successfully identified and discussed a set of specific management practices with the potential to impact productivity in different sectors (Bloom, Sadun, and Van Reenen 2016; Bloom and Van Reenen 2007, 2010; Bloom et al. 2020, 2012a, 2014, 2012b, 2019; Bloom et al. 2015). WMS has called these practices 'the best management practices'.

The WMS also developed a powerful survey tool to measure each of these practices in different sectors, as described in Bloom and Van Reenen (2007, 2010), Bloom et al. (2014, 2012b, 2019), and Bloom et al. (2015). For instance, the WMS survey tool for schools allows one to measure the management of a school by scoring each of the 23 management practices against a scoring grid that goes from one, worst management, to five, best management. The simple average of the scores given to each of the 23 best management practices represents the general management level of the surveyed school.

Although this study focuses on the causal relationship between management and pupils' educational outcomes, an important non-causal study conducted by Bloom et al. (2015) needs to be highlighted because it identified and measured the 23 best management practices specifically for schools through a large-scale survey of 1,800 schools across eight countries. As reported before, Bloom et al. (2015) found that a one-standard deviation (SD) change in the management of the schools

³⁷<https://worldmanagementsurvey.org/>

is strongly associated with 0.24 SD in pupils' educational outcomes.

The 23 best school management practices are very similar to those discussed by the WMS across different sectors (Bloom et al. 2015). Furthermore, these practices are very similar to the practices discussed across other relevant studies such as Barros et al. (2019, 2021), Bloom et al. (2020, 2012a), Fryer (2017), Gerber (2012), Gosnell, List, and Metcalfe (2020), and Tavares (2015). This set of 23 best management practices for schools is the benchmark to represent school management in this study.

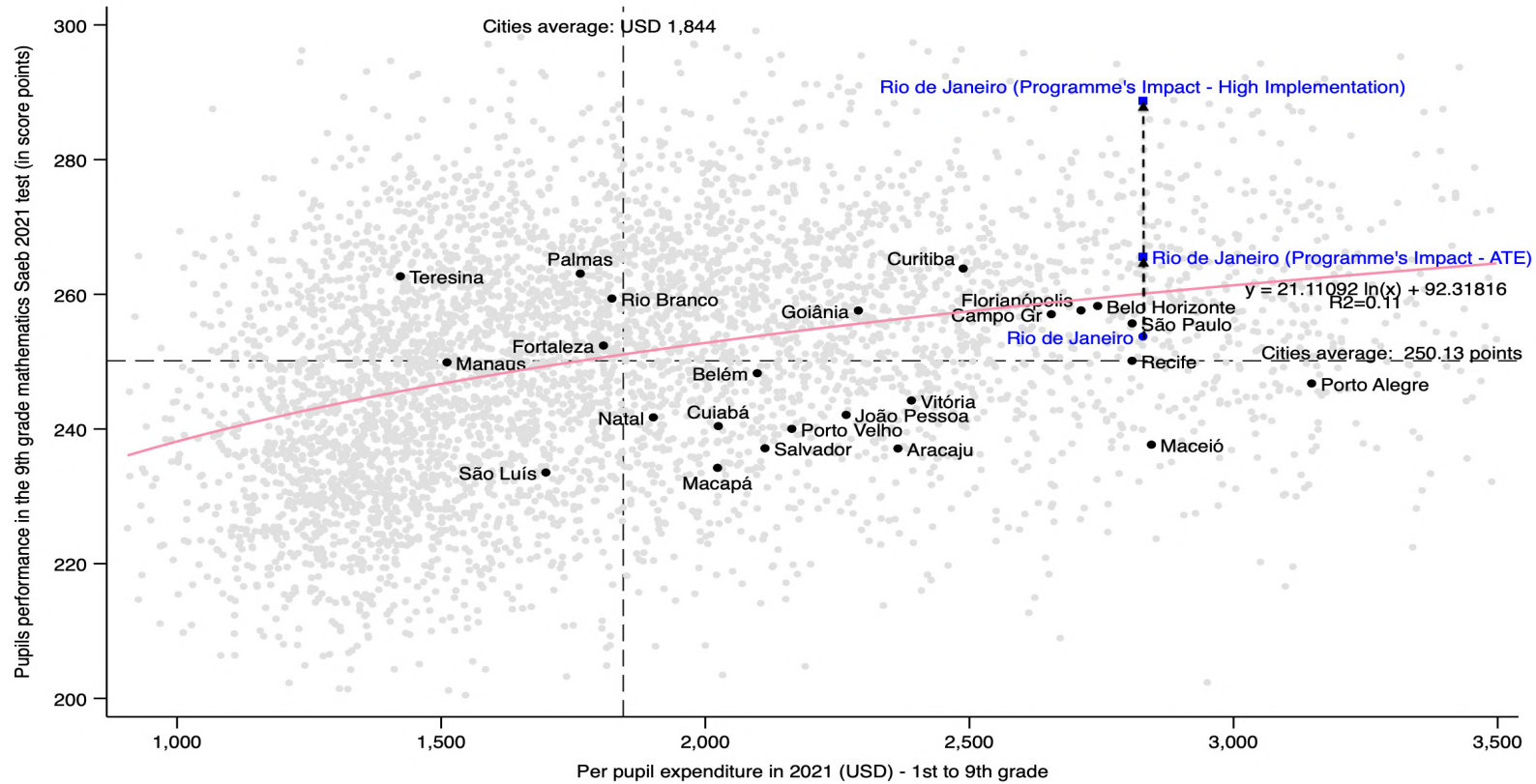
M. Additional Information on External Validity

Figure A6 shows the Brazilian cities' average performance in 9th-grade mathematics in the Saeb 2021 versus their public per pupil expenditure (grades 1-9) in 2021 in US dollars.³⁸ The capital cities of the federated states are highlighted. Data from Siope and Inep.

Since the Rio Assessments have the same scale and items (TRI) used by Saeb (one SD is equivalent to 50 score points), it is straightforward to translate our impact estimate in mathematics (ATE), 0.237 SD, as 11.85 score points in the Saeb scale. The same is done for the high implementation of the programme (IV estimates), i.e., 0.714 SD of impact, due to one score point change in management, can be translated to approximately 35 score points in the Saeb scale. The Science and Management for Education Programme should strongly affect the learning of Rio de Janeiro pupils.

³⁸USD 1.00 = R\$ 5.06 in 19/10/2023, where R\$ means Brazilian Reais.

Figure A6: Programme's Impact in the City of Rio de Janeiro Put in Brazil's Context



146

Notes: The figure shows Brazilian cities' performance in the ninth-grade mathematics test of the 2021 Basic Education Assessment System (Saeb), compared to the 2021 cities' public expenditure per pupil from grades 1-9 in USD (USD 1.00 = R\$ 5.06 in 19/10/2023, where R\$ means Brazilian Reais). The capital cities of the federated states are highlighted. Data from Siope and Inep.

REFERENCES

- Abadie, Alberto and Guido W. Imbens. “Bias-Corrected Matching Estimators for Average Treatment Effects”. *Journal of Business & Economic Statistics* 29.1 (2011), 1–11. DOI: 10.1198/jbes.2009.07333.
- Abdulkadiroğlu, Atila et al. “Accountability and Flexibility in Public Schools: Evidence from Boston’s Charters And Pilots”. *The Quarterly Journal of Economics* 126.2 (May 2011), 699–748. DOI: 10.1093/qje/qjr017.
- Abeberese, Ama Baafr, Todd J. Kumler, and Leigh L. Linden. “Improving Reading Skills by Encouraging Children to Read in School:” *Journal of Human Resources* 49.3 (2014), 611–633. DOI: 10.3368/jhr.49.3.611.
- Anderson, Michael L. “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects”. *Journal of the American Statistical Association* 103.484 (2008), 1481–1495. DOI: 10.1198/016214508000000841.
- Angrist, Joshua D et al. “Inputs and impacts in charter schools: KIPP Lynn”. *The American Economic Review* 100.2 (2010), 239–243. DOI: 10.1257/aer.100.2.239.
- Angrist, Joshua D. and Guido W. Imbens. “Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity”. *Journal of the American Statistical Association* 90.430 (1995), 431–442.
- Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters. “Explaining Charter School Effectiveness”. *American Economic Journal. Applied Economics* 5.4 (2013), 1–27. DOI: 10.1257/app.5.4.1.

- Angrist, Joshua D. et al. “Who Benefits from KIPP?” *Journal of Policy Analysis and Management* 31.4 (2012), 837–860. DOI: <https://doi.org/10.1002/pam.21647>.
- Angrist, Noam et al. *How to Improve Education Outcomes Most Efficiently? A Comparison of 150 Interventions using the New Learning-Adjusted Years of Schooling Metric*. The World Bank, 2020. DOI: 10.1596/1813-9450-9450.
- Angrist, Noam et al. “Measuring human capital using global learning data”. *Nature* 592.7854 (2021), 403–408. DOI: <https://doi.org/10.1038/s41586-021-03323-7>.
- Avvisati, Francesco and Pauline Givord. “The learning gain over one school year among 15-year-olds: An analysis of PISA data for Austria and Scotland (United Kingdom)”. *OECD Education Working Papers* (2021). DOI: <https://doi.org/https://doi.org/10.1787/d99e8c0a-en>.
- “The learning gain over one school year among 15-year-olds: An international comparison based on PISA”. *Labour Economics* 84 (2023). DOI: <https://doi.org/10.1016/j.labeco.2023.102365>.
- Baird, Sarah, Craig McIntosh, and Berk Özler. “Cash or Condition? Evidence from a Cash Transfer Experiment”. *The Quarterly Journal of Economics* 126.4 (2011), 1709–1753. DOI: 10.1093/qje/qjr032.
- Bambrick-Santoyo, Paul. *A Principal Manager’s Guide to Leverage Leadership 2.0: How to Build Exceptional Schools Across Your District*. John Wiley & Sons, 2018.

- Banerjee, Abhijit V. et al. “Remedying Education: Evidence from Two Randomized Experiments in India”. *The Quarterly Journal of Economics* 122.3 (2007), 1235–1264. DOI: 10.1162/qjec.122.3.1235.
- Bank, World. *World Development Report 2018: Learning to Realize Education’s Promise*. 2018. URL: <https://hdl.handle.net/10986/28340>.
- Barrera-Osorio, Felipe and Leigh L. Linden. *The Use And Misuse Of Computers In Education: Evidence From A Randomized Experiment In Colombia*. The World Bank, 2009. DOI: 10.1596/1813-9450-4836.
- Barros, Ricardo Paes et al. “Assessment of the Impact of the Jovem de Futuro Program on Learning”. In: World Bank Group., 2019.
- Barros, Ricardo Paes et al. “Promovendo o Desempenho Educacional Via Melhorias na Gestão Escolar: O Caso do Programa Jovem de Futuro”. *Pesquisa e Planejamento Econômico* 51.3 (2021), 9–44.
- Beg, Sabrin A, Anne E Fitzpatrick, and Adrienne Lucas. “Managing to Learn”. *NBER Working Paper Series* (2023). DOI: 10.3386/w31757.
- Bettinger, Eric. “Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores”. *The Review of Economics and Statistics* 94.3 (2012), 686–698.
- Blimpo, Moussa P. and David K. Evans. “School-Based Management and Educational Outcomes: Lessons From a Randomized Field Experiment”. In: *enGender Impact: The World Bank’s Gender Impact Evaluation Database*, 2013.
- Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. *Management as a Technology?* Working Paper 22327. National Bureau of Economic Research, June 2016. DOI: 10.3386/w22327.

- Bloom, Nicholas and John Van Reenen. “Measuring and Explaining Management Practices Across Firms and Countries”. *The Quarterly Journal of Economics* 122.4 (2007), 1351–1408. DOI: <https://doi.org/10.1162/qjec.2007.122.4.1351>.
- “New Approaches to Surveying Organizations”. *The American Economic Review* 100.2 (2010), 105–109. DOI: [10.1257/aer.100.2.105](https://doi.org/10.1257/aer.100.2.105).
- Bloom, Nicholas et al. “Do Management Interventions Last? Evidence from India”. *American Economic Journal: Applied Economics* 12.2 (Apr. 2020), 198–219. DOI: [10.1257/app.20180369](https://doi.org/10.1257/app.20180369).
- Bloom, Nicholas et al. “Does Management Matter? Evidence from India”. *The Quarterly Journal of Economics* 128.1 (Nov. 2012), 1–51. DOI: [10.1093/qje/qjs044](https://doi.org/10.1093/qje/qjs044).
- Bloom, Nicholas et al. “JEEA-FBBVA Lecture 2013: The New Empirical Economics of Management”. *Journal of the European Economic Association* 12.4 (2014), 835–876. DOI: <https://doi.org/10.1111/jeea.12094>.
- Bloom, Nicholas et al. “Management Practices Across Firms and Countries”. *Academy of Management Perspectives* 26.1 (2012), 12–33. DOI: <https://doi.org/10.5465/amp.2011.0077>.
- Bloom, Nicholas et al. “What Drives Differences in Management Practices?” *American Economic Review* 109.5 (May 2019), 1648–83. DOI: [10.1257/aer.20170491](https://doi.org/10.1257/aer.20170491).
- Bloom, Nick et al. “Does Management Matter in Schools?” *The Economic Journal* 125.584 (2015), 647–674. DOI: <https://doi.org/10.1111/econj.12267>.

- Borkum, Evan, Fang He, and Leigh L Linden. “The Effects of School Libraries on Language Skills: Evidence from a Randomized Controlled Trial in India”. *NBER Working Paper Series* (2012), 18183. DOI: 10.3386/w18183.
- Borman, Geoffrey D, Robert E Slavin, and Alan C. K Cheung. “Final Reading Outcomes of the National Randomized Field Trial of Success For All”. *American Educational Research Journal* 44.3 (2007), 701–731. DOI: <https://doi.org/10.3102/0002831207306743>.
- Bruhn, Miriam, Dean Karlan, and Antoinette Schoar. “The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico”. *Journal of Political Economy* 126.2 (2018), 635–687. DOI: 10.1086/696154.
- Burde, Dana and Leigh L. Linden. “Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools”. *American Economic Journal: Applied Economics* 5.3 (2013), 27–40. DOI: 10.1257/app.5.3.27.
- Card, David. “Chapter 30 - The Causal Effect of Education on Earnings”. In: *Handbook of Labor Economics*. Vol. 3. Elsevier, 1999, 1801–1863. DOI: [https://doi.org/10.1016/S1573-4463\(99\)03011-4](https://doi.org/10.1016/S1573-4463(99)03011-4).
- Chaisemartin, Clément de and Jaime Ramirez-Cuellar. “At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?” *American Economic Journal: Applied Economics* 16.1 (2024), 193–212. DOI: 10.1257/app.20210252.
- Chetty, Raj, John N Friedman, and Jonah E Rockoff. “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adult-

- hood”. *The American Economic Review* 104.9 (2014), 2633–2679. DOI: 10.1257/aer.104.9.2633.
- Cristia, Julian et al. “Technology and Child Development: Evidence from the One Laptop per Child Program”. *American Economic Journal. Applied Economics* 9.3 (2017), 295–320. DOI: 10.1257/app.20150385.
- Curto, Vilsa E. and Roland G. Fryer. “The Potential of Urban Boarding Schools for the Poor: Evidence from SEED”. *Journal of Labor Economics* 32.1 (2014), 65–93. DOI: <https://doi.org/10.1086/671798>.
- Davies, Neil M et al. “The causal effects of education on health outcomes in the UK Biobank”. *Nature Human Behaviour* 2.2 (2018), 117–125. DOI: <https://doi.org/10.1038/s41562-017-0279-y>.
- Dhaliwal, Iqbal et al. “Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education”. In: *Education Policy in Developing Countries*. 1st ed. University of Chicago Press, 2013. DOI: <https://doi.org/10.7208/9780226078854-008>.
- Dobbie, Will and Roland G Fryer. “Are High-Quality Schools Enough to Increase Achievement Among the Poor?: Evidence from the Harlem Children’s Zone”. *American Economic Journal. Applied Economics* 3.3 (2011), 158–187. DOI: 10.1257/app.3.3.158.
- “Getting Beneath the Veil of Effective Schools: Evidence From New York City”. *American Economic Journal. Applied Economics* 5.4 (2013), 28–60. DOI: 10.1257/app.5.4.28.
- Doran, George T et al. “There’s a SMART way to write management’s goals and objectives”. *Management review* 70.11 (1981), 35–36.

- Duflo, Esther, Pascaline Dupas, and Michael Kremer. “Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya”. *The American Economic Review* 101.5 (2011), 1739–1774. DOI: [10.1257/aer.101.5.1739](https://doi.org/10.1257/aer.101.5.1739).
- “School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools”. *Journal of Public Economics* 123 (2015), 92–110. DOI: <https://doi.org/10.1016/j.jpubeco.2014.11.008>.
- Duflo, Esther, Rema Hanna, and Stephen Ryan. “Incentives Work: Getting Teachers to Come to School”. *The American Economic Review* 102.4 (2012), 1241–1278. DOI: [10.1257/aer.102.4.1241](https://doi.org/10.1257/aer.102.4.1241).
- Estudos e Pesquisas Anísio Teixeira (INEP), Instituto Nacional de. *Educational Financial Indicators*. 2021. URL: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-financeiros-educacionais>. (accessed: 16.10.2023).
- Feinstein, Alvan R. and Domenic V. Cicchetti. “High Agreement but Low Kappa: I. The Problems Of Two Paradoxes”. *Journal of Clinical Epidemiology* 43.6 (1990), 543–549. DOI: [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L).
- Fryer, Roland G. “Financial Incentives and Student Achievement: Evidence From Randomized Trials”. *The Quarterly Journal of Economics* 126.4 (2011), 1755–1798. DOI: <https://doi.org/10.1093/qje/qjr045>.
- Fryer, Roland G Jr. “Management and Student Achievement: Evidence from a Randomized Field Experiment”. *NBER Working Paper Series* (2017), 23437. DOI: [10.3386/w23437](https://doi.org/10.3386/w23437).

- Fryer, Roland G Jr. “The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments”. *NBER Working Paper Series* (2016), 22130. DOI: [10.3386/w22130](https://doi.org/10.3386/w22130).
- Fryer, Roland G. “Injecting Charter School Best Practices Into Traditional Public Schools: Evidence From Field Experiments”. *The Quarterly Journal of Economics* 129.3 (2014), 1355–1408. DOI: <https://doi.org/10.1093/qje/qju011>.
- Gerber, Alan S. *Field Experiments: Design, Analysis, and Interpretation*. New York, N.Y.: W. W. Norton & Company, 2012.
- Gibbons, Robert and Rebecca Henderson. “Relational Contracts and Organizational Capabilities”. *Organization Science* 23.5 (2012), 1350–1364. DOI: [10.1287/orsc.1110.0715](https://doi.org/10.1287/orsc.1110.0715).
- Glazerman, Steven, Daniel Mayer, and Paul Decker. “Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes”. *Journal of Policy Analysis and Management* 25.1 (2006), 75–96. DOI: <https://doi.org/10.1002/pam.20157>.
- Glazerman, Steven et al. “Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment.” *National Center for Education Evaluation and Regional Assistance* (2013).
- Glewwe, Paul, Nauman Ilias, and Michael Kremer. “Teacher Incentives”. *American Economic Journal. Applied Economics* 2.3 (2010), 205–227. DOI: [10.1257/app.2.3.205](https://doi.org/10.1257/app.2.3.205).
- Glewwe, Paul, Michael Kremer, and Sylvie Moulin. “Many Children Left Behind? Textbooks and Test Scores in Kenya”. *American Economic Journal: Applied Economics* 1.1 (2009), 112–35. DOI: [10.1257/app.1.1.112](https://doi.org/10.1257/app.1.1.112).

- Gosnell, Greer K., John A. List, and Robert D. Metcalfe. “The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains”. *The Journal of Political Economy* 128.4 (2020), 1195–1233.
- Gwet, K.L. *Handbook of Inter-Rater Reliability, 5th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2021.
- Hoyos, Rafael de, Alejandro J Ganimian, and Peter A Holland. “Teaching with the Test: Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina”. *The World Bank Economic Review* 35.2 (Nov. 2019), 499–520. DOI: 10.1093/wber/lhz026.
- Imbens, Guido W. and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. DOI: <https://doi.org/10.1017/CB09781139025751>.
- Institute, Project Management. *A Guide to the Project Management Body of Knowledge (PMBOK® Guide) – Seventh Edition and The Standard for Project Management*. Chicago: Project Management Institute, 2021.
- *Agile Practice Guide*. Newtown Square, PA: Project Management Institute, Inc. (PMI), 2017.
- J-PAL. *Conducting cost-effectiveness analysis (CEA)*. 2020. URL: <https://www.povertyactionlab.org/resource/conducting-cost-effectiveness-analysis-cea>. (accessed: 09.01.2024).
- Kahan, Brennan C et al. “Estimands in Cluster-Randomized Trials: Choosing Analyses that Answer the Right Question”. *International Journal of*

- Epidemiology* 52.1 (2023), 107–118. DOI: <https://doi.org/10.1093/ije/dyac131>.
- Kayembe, Mutamba T. et al. “Imputation of Missing Covariates in Randomized Controlled Trials with Continuous Outcomes: Simple, Unbiased and Efficient Methods”. *Journal of Biopharmaceutical Statistics* 32.5 (2022), 717–739. DOI: <https://doi.org/10.1080/10543406.2021.2011898>.
- Kraft, Matthew A. “Interpreting Effect Sizes of Education Interventions”. *Educational Researcher* 49.4 (2020), 241–253. DOI: [10.3102/0013189X20912798](https://doi.org/10.3102/0013189X20912798).
- Kraft, Matthew A., David Blazar, and Dylan Hogan. “The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence”. *Review of Educational Research* 88.4 (2018), 547–588. DOI: <https://doi.org/10.3102/0034654318759268>.
- Kremer, Michael, Edward Miguel, and Rebecca Thornton. “Incentives to Learn”. *The Review of Economics and Statistics* 91.3 (2009), 437–456. DOI: [10.1162/rest.91.3.437](https://doi.org/10.1162/rest.91.3.437).
- Krueger, Alan B. “Economic considerations and class size”. *The Economic Journal (London)* 113.485 (2003), F34–F63. DOI: <https://doi.org/10.1111/1468-0297.00098>.
- “Experimental Estimates of Education Production Functions”. *The Quarterly Journal of Economics* 114.2 (1999), 497–532. DOI: <https://doi.org/10.1162/003355399556052>.
- Landis, J. Richard and Gary G. Koch. “The Measurement of Observer Agreement for Categorical Data”. *Biometrics* 33.1 (1977), 159–174. DOI: <https://doi.org/10.2307/2529310>.

- Lassibille, Gérard et al. “Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions”. *The World Bank Economic Review* 24.2 (2010), 303–329. DOI: <https://doi.org/10.1093/wber/lhq009>.
- Lin, Winston. “Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman’s Critique”. *The Annals of Applied Statistics* 7.1 (2013), 295–318. DOI: 10.1214/12-AOAS583.
- Morrow-Howell, Nancy et al. “Evaluation of Experience Corps: Student Reading Outcomes.” (2009).
- Muralidharan, Karthik and Paul Niehaus. “Experimentation at Scale”. *The Journal of Economic Perspectives* 31.4 (2017), 103–124. DOI: 10.1257/jep.31.4.103.
- Muralidharan, Karthik and Abhijeet Singh. *Improving Public Sector Management at Scale? Experimental Evidence on School Governance India*. Working Paper 28129. National Bureau of Economic Research, 2020. DOI: 10.3386/w28129.
- Muralidharan, Karthik and Venkatesh Sundararaman. “The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India”. *The Economic Journal* 120.546 (2010), F187–F203. DOI: <https://doi.org/10.1111/j.1468-0297.2010.02373.x>.
- Neal, Derek A. and William R. Johnson. “The Role of Premarket Factors in Black-White Wage Differences”. *The Journal of Political Economy* 104.5 (1996), 869–895. DOI: <https://doi.org/10.1086/262045>.
- Negi, Akanksha and Jeffrey M. Wooldridge. “Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects”. *Economet-*

- ric Reviews* 40.5 (2021), 504–534. DOI: <https://doi.org/10.1080/07474938.2020.1824732>.
- Nguyen, Trang. “Information, role models and perceived returns to education: Experimental evidence from Madagascar”. *Unpublished manuscript* 6 (2008).
- OECD. *Education at a Glance 2021*. 2021, 474. DOI: <https://doi.org/https://doi.org/10.1787/b35a14e5-en>.
- *PISA 2022 Results (Volume I)*. 2023, 491. DOI: <https://doi.org/https://doi.org/10.1787/53f23881-en>.
- Oreopoulos, Philip and Kjell G Salvanes. “Priceless: The Nonpecuniary Benefits of Schooling”. *The Journal of Economic Perspectives* 25.1 (2011), 159–184. DOI: 10.1257/jep.25.1.159.
- Pradhan, Menno et al. “Improving educational quality through enhancing community participation: Results from a randomized field experiment in Indonesia”. *American Economic Journal. Applied Economics* 6.2 (2014), 105–126. DOI: 10.1257/app.6.2.105.
- Pritchett, Lant. *The Rebirth of Education: Schooling Ain’t Learning*. Brookings Institution Press, 2013.
- Puma, Michael et al. “Head Start Impact Study. Final Report.” *Administration for Children & Families* (2010).
- Romero, Mauricio et al. “Direct vs indirect management training: Experimental evidence from schools in Mexico”. *Journal of Development Economics* 154 (2022), 102779. DOI: <https://doi.org/10.1016/j.jdeveco.2021.102779>.

- Sloczynski, Tymon. “Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights”. *The Review of Economics and Statistics* 104.3 (2022), 501–509. DOI: https://doi.org/10.1162/rest_a_00953.
- Somers, Marie-Andree et al. “The Enhanced Reading Opportunities Study Final Report: The Impact of Supplemental Literacy Courses for Struggling Ninth-Grade Readers. NCEE 2010-4021.” *National Center for Education Evaluation and Regional Assistance* (2010).
- Tavares, Priscilla Albuquerque. “The impact of school management practices on educational performance: Evidence from public schools in São Paulo”. *Economics of Education Review* 48 (2015), 1–15. DOI: <https://doi.org/10.1016/j.econedurev.2015.05.002>.
- Thaler, Richard H. *Nudge / Richard H. Thaler and Cass R. Sunstein*. The final edition [Allen Lane paperback edition]. 2021.
- Tombolo, Guilherme Alexandre and Armando Vaz Sampaio. “O PIB brasileiro nos séculos XIX e XX: duzentos anos de flutuações econômicas”. *Revista de Economia* 39.3 (2013).
- Zhao, Anqi and Peng Ding. “To Adjust or not to Adjust? Estimating the Average Treatment Effect in Randomized Experiments with Missing Covariates”. *Journal of the American Statistical Association* 0.0 (2022), 1–11. DOI: [10.1080/01621459.2022.2123814](https://doi.org/10.1080/01621459.2022.2123814).