Policy Research Working Paper 9748

Scores, Camera, Action

Social Accountability and Teacher Incentives in Remote Areas

Arya Gaduh
Menno Pradhan
Jan Priebe
Dewi Susanti

WORLD BANK GROUP
Social Sustainability and Inclusion Global Practice
August 2021

Policy Research Working Paper 9748

Abstract

Remote schools in developing countries are costly to supervise, resulting in low teacher accountability and poor education outcomes. This paper reports the results of a randomized evaluation of three treatments that introduced teacher incentives based on community monitoring of teacher effort against locally agreed standards. The Social Accountability Mechanism (SAM) treatment facilitated a joint commitment between schools and community members to improve learning. Teacher performance was rated against it, discussed in monthly public meetings and passed on to authorities. The second and third treatments combined SAM with a performance pay mechanism that would penalize eligible teachers' remote area allowance for poor performance. In the SAM+Camera (SAM+Cam) treatment, the cut was based on absence as recorded by a tamper-proof camera; while in the SAM+Score treatment, it was based on the overall rating. After one year, the findings indicate improvements in learning outcomes across all treatments; however, the strongest impact of 0.20 standard deviation is observed for SAM+Cam. The evaluation also finds a small positive impact on the effort of affected teachers for SAM+Cam and SAM, and significant positive improvements on parental educational investments in all treatments. For SAM and SAM+Cam, additional data were collected in the second year (one year after project facilitators left). The findings show that SAM+Cam's impacts on learning outcomes and parental investments—but not teacher effort—persisted into the second year.

This paper is a product of the Social Sustainability and Inclusion Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dsusanti@worldbank.org.

Acknowledgments *A large number of people contributed to the design, implementation, data collection, data analysis, and policy recommendation of this research project. For design, we are grateful to Amanda Beatty, Christopher Bjork, Jishnu Das, Deon Filmer, Scott Guggenheim, Rema Hanna, Nur Hidayat, Gunawan, Marliyanti, Karthik Muralidharan, Setiawan Cahyo Nugroho, Lant Pritchett, Jurist Tan, Robert Wrobel, Deny Purwo Sambodo, Halsey Rogers, Dewi Sudharta, and Daniel Suryadarma. For excellent research assistantships, we thank Usha Adelina, Emilie Berkhout, Kurniawati, Sharon Kanthy Lumbanraja, Marliyanti, and Indah Ayu Prameswari. Survey data collection was led by Dedy Junaedi, Lulus Kusbudiharjo, Anas Sutisna, and Mulyana. Implementation by BaKTI was led by Muhammad Yusran Laitupa, Setiawan Cahyo Nugroho, Tri Yuni Rinawati, and Caroline Tupamahu. We are also grateful to all of the KIAT Guru district coordinators, facilitators, data management officers, administrative assistants, and enumerators.*

Research and implementation supports were provided by the World Bank under the management of Nina Bhatt, Janmejay Singh, and Kevin Tomlinson. Dewi Susanti led the task team with inputs from Gregorius Kelik Endarso, Tazeen Fasih, Yulia Herawati, Lily Hoo, Megha Kapoor, Camilla Holmemo, Javier Luque, Cristobal Ridao-Cano, Audrey Sacks, Chatarina Ayu Widiarti, Noah Bunce Yarrow, and Fazlania Zain. We are grateful to Andrew Brownback, David Evans, Deon Filmer, Robert Garlick, Jose Antonio Cuesta Leiva, Tobias Linden, Alejandro Ome, Lant Pritchett, Halsey Rogers, Mauricio Romero, Susan Wong, and seminar participants at the 2019 briq/IZA Workshop on Behavioral Economics of Education, 2019 RISE Seminar, 2019 Pacific Development Conference, 2019 Midwest International Economic Development Conference, 2019 DIAL Development Conference, 2019 Annual International Conference of the Research Group on Development, 2019 NEUDC conference, EUDN 2019, 2020 KDIS-3ie-ADB-ADBI Conference on Impact Evaluation, the World Bank's Social Sustainability and Inclusion GP & Data, Analytics, and Digital GSG BBL, and the Hong Kong University Business School for helpful comments and suggestions.

The research would not be possible without the supports from the Indonesian Ministry of Education and Culture (MoEC), the National Team for Acceleration of Poverty Reduction under the Office of the Vice President of Indonesia (TNP2K), and the five district governments of Ketapang, Landak, Sintang, Manggarai Barat and Manggarai Timur. We are especially grateful for advice provided by TNP2K team, under the leaderships of Bambang Widianto, Suahazil Nazara, Elan Satriawan, and Sudarno Sumarto, and by MoEC team, under the leaderships of Sumarna Surapranata, Supriano, Iwan Syahril, Nurzaman, Dian Wahyuni, Praptono, Rahmadi Widdharto, Suharti, Temu Ismail, and Budi Kusumawati. We acknowledge financial support from the Government of Australia's Department of Foreign Affairs and Trade and USAID through Trust Funds managed by the World Bank. RISE Study in Indonesia, managed by SMERU Research Institute, also co-financed the second round of surveys. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/ World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Scores, Camera, Action: Social Accountability and Teacher Incentives in Remote Areas

Arya Gaduh*
University of Arkansas

Menno Pradhan †
University of Amsterdam, Vrije Universiteit Amsterdam, AIGHD, and Tinbergen Institute

Jan Priebe ‡
GIGA Institute Hamburg, University of Göttingen

Dewi Susanti §
World Bank, Jakarta

JEL Classifications: H52, I21, I25, I28, O15

Keywords: social accountability, community-based monitoring, teacher incentives, performance pay, remote-area policy

*Sam M. Walton College of Business. Department of Economics. Business Building 402, Fayetteville, AR 72701-1201. Email: agaduh@walton.uark.edu.

†Department of Development Economics, University of Amsterdam and Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. Email: m.p.pradhan@uva.nl.

‡GIGA Institute of Asian Studies, Rothenbaumchaussee 32, 20148 Hamburg, Germany. Email: jpriebe@uni-goettingen.de.

§The World Bank, Indonesia Stock Exchange (IDX) Tower 2 L12, Jalan Jend Sudirman, Senayan, DKI Jakarta 12190, Indonesia. Email: dsusanti@worldbank.org.

1 Introduction

Weak accountability is a primary reason behind the low quality of public service provision in many low- and middle-income countries (LMICs) (World Bank, 2003, 2018). It is responsible for high rates of absenteeism and given the centralized structure of public services in most LMICs, accountability is weakest in remote areas where top-down supervision is costliest. Social accountability initiatives offer a promising solution to this problem (Grandvoinnet et al., 2015). By providing citizens with information, voice, and influence, they aim at enabling communities to hold service providers accountable and provide better-quality services (Brixi et al., 2015; Bruns et al., 2011; Ringold et al., 2012). To what extent social accountability initiatives can live up to their promise is still debatable, with mixed evidence on its effectiveness (Joshi, 2013; Mansuri and Rao, 2012; Molina et al., 2017).

Proponents of social accountability initiatives have argued that many of these interventions fail to show a meaningful impact because citizens are not sufficiently empowered (Fox, 2015a; Kosec and Wantchekon, 2020). Initiatives that focus on increasing information alone often do not provide the opportunities to transform information into action (Afridi et al., 2020; Lieberman et al., 2014). Moreover, even if such opportunities exist — e.g., through formal grievance redress mechanisms — other challenges remain. Staff of public service providers (e.g. doctors, principals, and teacher) are often in a position of power because they are better educated, have better connections to policy-makers, and are hard to replace. This makes it easier for service providers to bargain for leniency even if service delivery falls short (Banerjee et al., 2008).

Endowing citizens with some control over provider incentives can add "teeth" to social accountability interventions. It allows for "the short route of accountability" (World Bank, 2003) by improving citizens' ability to directly hold service providers accountable. Combining incentive payment with locally generated information about performance can be an effective measure to improve service quality. For example, Cilliers et al. (2018) show that disseminating principal report of teacher presence to communities was only effective to reduce absenteeism when linked with teacher bonuses. A well-designed incentive payment around a social accountability intervention holds promise since the latter can not only generate performance information, but also raise community awareness and engagement. Nonetheless, we lack evidence on its effectiveness.

This paper fills this gap. We use a randomized controlled trial (RCT) to study how combinations of performance pay and social accountability interventions can improve learning in Indonesia's remote area schools. We work with national and district governments to implement three treatments where we varied the presence and type of performance pay mechanism added to the social accountability mechanism. We test their effectiveness in a large sample of mostly public primary schools. The performance pay component incentivizes a government-financed supplemental allowance for teachers assigned to a remote area. By working with an existing government allowance, our study offers a unique insight on the potential impact of a public sector pay reform in LMICs.

The common component of these treatments is the social accountability intervention, which formulated a locally agreed standard of performance for teachers, and established an instrument and institution to evaluate its implementation. To establish the standard, parents were first informed of the level of learning in their school. Facilitators supported them to formulate a joint agreement with teachers that

specified the responsibilities of teachers and parents to improve learning in the school. To monitor and evaluate progress on the teacher's side, a teacher-specific scorecard was developed and a user committee (UC) established. The UC was responsible for the monthly performance evaluation of each teacher based on the scorecard indicators. Their evaluations were discussed in monthly public meetings and upon completion, sent to the district education office.

The Social Accountability Mechanism (SAM) treatment only has the social accountability component. The other two treatments, SAM+Cam and SAM+Score, augmented social accountability with a performance pay component that targeted the Teacher Special Allowance (TSA). TSA is a government hardship allowance (in the amount of up to the base salary) for select teachers working in remote areas. SAM+Cam and SAM+Score penalize TSA-receiving teachers' poor performance by cutting their TSA. In SAM+Cam, the penalty was based solely on the teacher presence indicator, which was verified every month on the basis of selfies the teacher had taken at the start and end of every workday with a specially provided smartphone.[1] Meanwhile, in SAM+Score, the penalty was based on the overall score on the scorecard.

Our three treatments thus vary in the way communities can use these scorecards to hold teachers accountable. SAM relies on UC nudges during the monthly evaluation meetings to remind teachers of their commitments to improve service delivery. It improves the flow of information on teacher performance to communities and provides UC a direct channel to report to school administrators. As such, SAM might improve teacher management and induce further teacher efforts. By adding incentive payment, SAM+Cam and SAM+Score strengthen community voice with an ability to penalize poor performance. The incentive in SAM+Cam is based on an objective, but incomplete measure of service delivery (presence). The SAM+Score incentive incorporates all agreed indicators but is more prone to subjectivity.

The study was done in 270 schools in 2 districts in East Nusa Tenggara and 3 districts in West Kalimantan provinces of Indonesia from October 2016 to May 2019. Districts were drawn from the central government's list of disadvantaged regions. Within each district, eligible schools need to satisfy our remoteness criteria and have a minimum number of teachers receiving the TSA. We then use stratified-random assignment to assign schools to the control and treatment groups. Implementation by project facilitators began in October 2016 and ended in December 2017, and endline data were collected in early 2018. A year after project facilitators left, we collected a follow-up survey for schools in the control, SAM, and SAM+Cam groups in early 2019.

Our results show that all treatments increased learning outcomes, measured by assessments of Indonesian and mathematics. However, while we find that all treatments led to positive improvements in learning, SAM+Cam shows the most promise. SAM and SAM+Score led to similar improvements in learning when project implementation completed. In comparison, adding the *Cam* aspect to SAM approximately doubled the effect sizes: SAM+Cam led to improvements in learning by 0.20 standard deviation (s.d.).[2] Overall, these impacts do not differ by gender, but are stronger for students in earlier

[1] The intervention is similar to the camera monitoring of teachers studied in (Duflo et al., 2012) The main difference is that in this study, the cameras were read out in the village meetings whereas in their paper, the pictures were sent directly a central administrator. Also the formula used to calculate teacher pay differs.

[2] It is important to emphasize that the relative success of SAM+Cam over SAM is *not* evidence of the effectiveness of a pure *Cam* treatment. The difference between SAM+Cam and SAM cannot be interpreted as the positive impact of a pure camera-based performance-pay treatment without assuming no interaction effects between SAM and the attendance-based, camera-supported performance-pay interventions.

grades and those who were better performing at baseline. For SAM+Cam, the effect persisted into the second year. Importantly, our decomposition exercise shows that these effects were not merely knock-on impacts from the first-year implementation.

The overall effects of the treatments on teacher presence and self-reported work hours were negligible. However, we do observe that teachers shift effort towards learning enhancing activities in SAM and SAM+CAM, and parents report meeting with teachers more often. Looking at the effects on TSA and non-TSA teachers separately, we find that the salary incentive treatments reduced presence of non-TSA teachers relative to TSA teachers. No such effect was found for the SAM treatment.[3] The impacts on time spent on learning enhancing activities did not vary by TSA status. None of the impacts on teacher effort persisted into the second year.

The treatments led to generally positive effects on parental engagement in children's' education, reductions in child labor, parental satisfaction and aspirations. Most effects, with the exception to those related to child employment, persisted into the second year. Parents report meeting with teacher more often, although these effects, in line with the fading of teacher efforts, declined in the second year. Interestingly, even though learning impacts of SAM+CAM declined in the second year, parental satisfaction with learning increased over time for this treatment. This, together the persistent positive effects on parental satisfaction, suggest that parents had not yet become aware of the fading teacher efforts in the second year.

Can the teacher responses be explained by increased top-down supervision or informal community pressure? Our evidence suggests that both mechanisms are at play in delivering results. Across the board, principals increase their supervision of teachers and these continued into the second year. Supervision from district officials and school inspectors increased in SAM+Cam in the first year. These effects however did not persist into the second year, and even turned negative for SAM. Communities only make sparse use of the tools they have been provided to hold teachers accountable. Teacher rating scores are generally very high at around 95 percent of the maximum score.[4] The average salary cut for teachers who received the TSA was 5 percent.

To explore the mechanisms, we develop a simple model of parent and teacher efforts (which are inputs to learning) when schools do not receive external supervision (e.g., school inspectors). We show in the model that if parental assessments can penalize teachers' TSA and parents can commit to assessing teachers truthfully, they can induce teachers to provide high efforts. However, when evaluations are more subjective (which was the case in SAM+Score relative to SAM+Cam), which would lead to more disagreements between teachers and parents, and potential retaliations from teachers, it would be harder for parents to induce a high effort from teachers.

We find empirical evidence consistent with this model. First, we find that a stronger commitment for a truthful assessment (and a willingness to punish poor performance) leads to better outcomes. We measure this commitment using a lab-in-the-field experiment to estimate local punishment norms.

[3]These differential treatment effects are in line with the existence of social preferences as found in Breza et al. (2018). They found that Indian garment workers reduced their efforts if they were paid more for no clear reason. If however the higher payment was resulting from an objective performance evaluation, these effects disappeared. In our case, the camera monitoring is an objective evaluation, and we do not see negative spillovers on those who do not receive the TSA. The community rating in SAM+Score might seem more arbitrary, and therefore result in lower effort on those who do the remote area allowance.

[4]High average performance rating also common place in many firms. They are often explained by a fear of a reduction in morale and effort following unjustified low performance ratings (Macleod, 2003; Marchegiani et al., 2016).

We find larger student learning gains and positive improvements in teacher behavior in schools with a higher propensity to punish free riders. Second, we also find evidence that subjectivity can lead to more disagreements between parents and teachers. In their qualitative study, Bjork and Susanti (2020) report more teacher dissatisfaction of the role assigned to the UC in SAM+Score compared to that in SAM+Cam. They also report a higher incidence of teacher pressure to user committees to improve scores in SAM+Score. At the same time, evaluation scores in SAM+Score are somewhat higher than in SAM and SAM+Cam despite the lack of independently observed outcome to corroborate these scores.

The cost-effectiveness of our interventions are comparable to other interventions that aim to improve learning in developing countries evaluated using RCTs. SAM+Cam, which was the most successful among our interventions, improved learning outcomes by 0.2 standard deviation (s.d.) at the cost of USD 44 (in current 2017 dollar) per student. This cost is somewhere in the middle of the distribution of the cost-effectiveness of the various interventions reported in JPAL (2019).[5] However, it is worth noting that our interventions were implemented in areas that are harder (and costlier) to reach than typical education interventions.

Our paper builds on a rich literature on how to improve learning in schools in LMICs through social accountability. Interventions in this area often include project components that aim to increase demand for better education by raising awareness for low learning (Lieberman et al., 2014; Afridi et al., 2020), empower community groups through training and grants (Banerjee et al., 2010; Pradhan et al., 2014; Barrera-Osorio et al., 2020), and enhance channels for communication between communities and service providers so they can pressure for better services (Björkman and Svensson, 2009). While our SAM intervention broadly follows this approach, it is unique in that it forges a service agreement between teachers and communities with teacher-specific indicators, targets, and scoring. The use of salary incentives based on teacher presence as captured by a camera (Duflo et al., 2012) or recorded by the head master (Cilliers et al., 2018) increased learning in studies from India and Uganda respectively. The contribution of our study lies in combining social accountability with teacher incentives, and testing the impact of social accountability with and without teacher incentives in the same context.

We also contribute to the question on the role of (the type of) performance measures in incentive contracts. Several studies have shown that both subjective and objective teacher evaluation correlate with student achievement (Rockoff and Speroni, 2011). The use of subjective or objective measures in performance contract could however also affect performance (Baker et al., 1994). Our study, along with a recent study of (n.d.), estimate the causal effect of subjective versus objective teacher evaluation on student performance. (n.d.)'s study was done in the context of private schools in urban Pakistan, where arguably incentives to perform and school management are stronger than in our context. They tested whether a teacher monetary incentive based on student value added (test scores) performed better than one that was based on supervisor evaluations of performance against agreed targets. While both interventions increased learning outcomes, the latter did not lead to a reduction in effort of non-incentives tasks. So in their study the subjective (supervisor evaluations) outperformed the objective (student test scores) evaluation while in this study, the objective (SAM+CAM) outperformed the subjective (SAM+Score) evaluation. The apparent contradiction however hides a common conclusion that what matters is to

[5]When converted to current 2011 dollars for comparability, this implies a cost of USD 22 per 0.1 s.d. learning improvement for SAM+Cam.

what extent the performance measure is perceived as noisy and distorted by teachers. In that respect, student value added fell short of supervisor evaluations in the (n.d.) study, and community evaluations fell short of teacher presence in this study.

Scaling up successful pilots has proven to be hard and it has been difficult to pinpoint where the problem arises (Raffler et al., 2019; Bold et al., 2018). The problem is that many things, such as the context, budget, and implementing agency, often change when a pilot is scaled up. This paper contributes to this policy question by including a second year in the study during which an initial step towards scale up of the most successful interventions, handing over tasks form project facilitators to local cadre, was taken. The fact that we worked with a government allowance, backed up by regular budgets and legalizing regulations, probably helped to keep the administrative processes intact during the second year. The impacts on learning and teacher effort however reduced, which stresses the importance of in person support to projects that build incentives based on community monitoring.

The rest of the paper is as follows. The next section discusses the context and the experimental design, including how the interventions were implemented in the field and how communities respond to the interventions. Section 4 describes the data collection and empirical strategy. The following two sections discusses the impact of the treatments on student learning outcome (Section 5). To explore the mechanisms, we first present a conceptual model followed by the empirical results on teacher effort, parental engagements, and school management (Section 6). Section 7 discusses the extent of local support to these interventions and report their cost effectiveness. Section 8 concludes.

2 Teacher Accountability in Indonesia's Remote Areas

Overall, Indonesia has an adequate number of teachers: its student-teacher ratio for primary schools stood at 16:1 in 2016, one of the lowest in Southeast Asia (Kesuma et al., 2018). About 60 percent of teachers are civil servant teachers, whose hiring and salary standards are substantially higher than the rest, namely teachers under temporary contracts. Yet, many schools in remote areas face a shortage of qualified teachers (Heyward et al., 2017). Teacher absenteeism is also higher in remote areas at 19.3 percent compared to the national average of 9.4 percent (Usman et al., 2004; ACDP, 2014).

Government efforts to improve education quality have mostly focused on improving teacher welfare. In 2005, the Teacher Law introduced two new teacher allowances: *Tunjangan Profesi Guru* (the Teaching Profession Allowance) for teachers meeting professional standards and *Tunjangan Khusus Guru* (Teacher's Special Allowance, hereafter TSA) for teachers working in specially designated areas, including remote areas. None of these allowances are tied to teacher performance or student learning and evidence suggests they do not improve quality. Studies find that TSA recipients were more likely to be absent relative to non-recipients in the same school (SMERU, 2010) and that the Teaching Profession Allowance had no impact on learning (de Ree et al., 2018).

Our interventions work with the TSA, which is a non-permanent hardship allowance for teachers working in disadvantaged areas. Its value could go up to one times the teacher's base salary per month. Government-hired teachers (either civil servant or under a temporary contract) are eligible to receive the TSA if they worked in villages that are designated as remote and very disadvantaged.[6] Villages are

[6]There are three types of teacher status: permanent, contract, or school-contracted teacher. Permanent teachers are tenured

designated as very disadvantaged and remote by the central government based on a national index. The government adopted this more systematic approach to distributing the TSA in 2017, right before our interventions. The new approach improved on a more subjective and discretionary system that often led to inadequate fiscal allocations for the TSA.[7] In doing so, it expanded the coverage of the TSA.

In addition to the challenge of teacher retention, teachers in remotely located schools are also harder for local government agencies to supervise. Travel distance makes on-site supervisions costly and as a result, poor teacher performance can go by unnoticed by local authorities. Indonesia's experience with community-driven development (CDD) programs suggest that a community-based approach to monitoring service delivery offers a viable solution.[8] A common feature of these programs is the provision of community block grants accompanied by facilitation to ensure that grant money is spent in a transparent manner and in accordance to local needs. The success of these programs can, in part, be attributed to the long history Indonesia has in mobilizing community contributions for rural development programs (see p.71, Mansuri and Rao, 2012). Recent studies have investigated how CDD programs could be harnessed to increase use of health and education services (Olken et al., 2014). The design of our interventions build on the successful examples set forth by these programs.

3 Experimental Design

The *KIAT Guru* interventions set out to empower communities to hold teachers accountable.[9] Its design was informed by international evidence on how community-based approaches can improve service performance by strengthening the accountability relationships between principals (i.e., the government and beneficiaries) and agents (i.e., the service providers) (World Bank, 2003; Pritchett, 2015). These elements include: (i) having a standard to which the service providers will be accounted for; (ii) improving communities' access to information, including their basic rights to services; (iii) giving communities the means to influence and voice concerns to service providers; and (iv) providing routes to sanction poorly performing service providers (Joshi, 2013; Ringold et al., 2012). There is also some evidence that locally-defined and agreed-upon service standards are more effective than nationally-defined service standards in improving performance (World Bank, 2014, p.48).

This study follows up on Pradhan et al. (2014), which tested different ways to strengthen school committees in rural Central Java. It showed the importance of involving local leadership and ensuring that

civil servants, while contract teachers are hired either by district or provincial governments under annual contracts. School-contracted teachers are hired by the schools with a temporary employment status. The monthly pay range is the highest for permanent teachers (between around USD 108 and USD 408 depending on seniority), followed by contract teachers (between around USD 73 to 146), and school-contracted teachers (between USD 22 and 51).

[7] In the previous system, the national budget for TSA was determined based on proposals from the districts. The TSAs were then distributed to the districts, who would distribute it at their discretion. The amount of TSA received by districts often fell short of the number of teachers working in the disadvantaged and remote areas. For example, in 2013, there were 449,776 primary school teachers in disadvantaged districts but the TSA quota for that year was only for 53,038 teachers. To make do, districts would rotate the TSA recipients or distribute the TSA equally across all teachers. This had an effect of delinking the allowance from work in these disadvantaged areas: 42 percent of teachers working in such areas had no knowledge of the TSA, and only 26 percent both knew about it and were able to cite the amount they were entitled to (SMERU, 2010).

[8] Indonesia's CDD programs were developed following the Asian Economic Crisis and the fall of the Suharto regime in 1998. They were a response to the backlash against centrally-managed programs that were often associated with rampant corruption. These programs were initially financed through World Bank loans and in 2006, were eventually merged into the National Program for Community Empowerment (PNPM).

[9] KIAT Guru is the Indonesian abbreviation of the name of the project, i.e., "Teacher Performance and Accountability".

community involvement leads to concrete actions that improve education. It also underlined the difficulty of inducing increased efforts of teachers if there are no incentives attached to community action. A pathway analysis suggested that the positive effects on learning in this study were mostly a result of increased inputs of the community and not teacher effort.

The final design for KIAT Guru was informed by an operational pilot conducted in 31 schools in very remote villages in Keerom, Kaimana, and Ketapang districts of Indonesia, from June 2014 to December 2015. The operational pilot tested the implementation of key processes (e.g., facilitation of community meetings, pay-for-performance mechanisms), the legal and administrative regulations, process-monitoring instruments, and the survey instruments. Key lessons learned from the operational pilot set the parameter for the implementation of the study, particularly on district and village selections.[10]

3.1 Experimental Treatments

There are two core components of our treatments: (i) the social accountability mechanism (SAM) to formulate local service standards and form a user committee to monitor their adherence; and (ii) a set of performance pay mechanisms that links evaluation results to cuts to teacher pay. All treatments include the former, but vary in terms of the latter. We first describe each component, followed by the variation that defines the different treatments below.

3.1.1 Social Accountability Mechanism (SAM)

The SAM comprises a local service standard and a community monitoring institution that were established through a facilitator-driven set of meetings. The first of these meetings was an orientation meeting, attended by parents, community members and leaders, and school management (including teachers) to inform them about the pilot and their rights to participate in education service delivery. Subsequently, three separate meetings with representatives of students and alumni, parents and community members, and teachers discussed how to improve the learning environment at school and at home, and what role each of them should play. Afterward, everyone came together to formulate the service agreement. The service agreement listed a set of actions to improve the learning environment that parents, community leaders, teachers, and the school principal would commit to.

This service agreement became the basis for the principal- and teacher-specific scorecard. Between 5 and 8 indicators with targets that the principal and teachers committed to in the service agreement were made part of the scorecard. Meeting participants were free to choose the included indicators; however, the scorecard must always include the presence indicator. Afterward, participants assigned a weight to each indicator that reflected (their belief of) its importance to improve learning. These weights must add up to 100. A meeting of the user committee would then elaborate how each indicator would be scored. A scorecard would therefore consist of a set of indicators and targets, each was accompanied with a weight and a scoring guideline.

[10] Among others, we find that the success of the program requires commitments at multiple levels. Community needs to be willing to contribute time and resources and demand better education services. Both district and school managements need to be sufficiently transparent about their finances. Finally, the district bureaucracy needs to be reform-minded enough to fully support program implementation.

To monitor and evaluate teacher compliance to his/her scorecard, a user committee (UC) was established. The UC must have a minimum of nine members, with at least half of them female and marginalized group representatives. It also includes parents to represent each of the grade levels. The facilitation manual encouraged overlapping memberships between the UC and existing village and school communities. However, we did not find many incidences of overlapping memberships in our data.

The facilitator also recruited a village cadre who would be prepared to take over the role of a facilitator once the project was completed. The village cadre organized monthly village meetings and facilitated the meetings. Seventy five percent of the village cadres were appointed at the first village meeting. They co-organized and co-facilitated the aforementioned set of meetings to establish the SAM with the facilitators.

The UC is responsible for conducting monthly meetings to review the implementation of the service agreement and evaluate the scorecard. In these meetings, participants discussed their view on the service agreement and suggested potential improvements to the indicators. The UC presented their monthly evaluation of the scorecard and allow each teacher an opportunity to respond. Once the score for each teacher was finalized, UC members and the teacher/principal signed off on the evaluation results. These evaluation results were then posted or announced in a village meeting and dispatched to the district government.

A key activity that was conducted during the implementation of the SAM was to inform the community of how their children were doing in terms of learning. This was done a few months after the beginning of the implementation in a village-wide meeting whose objectives were, among others, to reassess the current service agreement, scorecard, and UC memberships. Prior to the meeting, the village cadre and UC members who had undergone training administered a learning diagnostic test that identifies students' skills in basic literacy and numeracy along a learning continuum of the national curriculum. The diagnostic test was administered to a random sample of six students per grade level. Results from the diagnostic test were shared at the beginning of this evaluation meeting.

3.1.2 Varying the Enforcement Mechanisms

The treatments vary in how the UC evaluations can affect the amount of TSA that eligible teachers received. Table 1 presents the three experimental treatments in our study: SAM, SAM+Cam, and SAM+Score. These treatments vary in how performance evaluations affect the amount of TSA allowance received. The SAM treatment lacks a performance pay mechanism (PPM): eligible teachers always received their full TSA amount. SAM+Cam and SAM+Score differ in the indicators and tools that were used to penalize poor performance with cuts to the TSA allowance. In all treatments, non-TSA teachers were evaluated the same way as TSA teachers, but the evaluation had no effect on their income.

In the SAM+Cam treatment, only the teacher presence indicator affected the TSA amount. Teachers in the SAM+Cam schools are given a tamper-proof smartphone camera to record their presence. They take pictures at the beginning and end of a school day and document their arrival and departure times on a manually entered teacher attendance form. At the end of each month, the UC verifies these records and any letters provided by teachers to account for their absences. Based on these daily records, the UC penalizes teachers for each partial presence (up to 1.5 presence points), excused absence (2 points), or unexcused absence (5 points). If a teacher maintains a presence score of 85 or above, they would receive

Table 1: Summary of the Treatments

	Control	SAM	SAM+ Cam	SAM+ Score
SAM: Scorecards and user committee	No	Yes	Yes	Yes
PPM: Presence indicator	No	No	Yes	Yes
PPM: Indicators other than presence	No	No	No	Yes
Tamper-proof camera	No	No	Yes	No
Number of schools	67	68	68	67

the share of the TSA equal to that score; otherwise, they will lose their TSA for that month.[11]

In the SAM+Score treatment, the scores used to determine the amount of TSA received were based on the entire scorecard. Three things distinguish SAM+Score from SAM+Cam. First, in SAM+Score, cuts to the TSA are based on the total weighted scores of all indicators in the scorecard, and not only the presence indicator. Second, unlike in SAM+Cam, there was no cut-off score below which a teacher would not receive the allowance: If a teacher received a score of 79 for that month, they would receive 79 percent of their TSA allowance. Finally, since no camera was provided for the SAM+Score schools, the mandatory presence indicator needed to be proactively monitored by the UC following the steps suggested during in their training.

We made sure that the TSAs were uniformly and reliably disbursed across control and treatment schools. The TSA was paid on a quarterly basis. Civil servant teachers were paid by the district governments, while non-civil servant teachers were paid directly by the education ministry. All payments were made through direct transfers into the teacher's bank account.

3.2 District and School Selection

We worked in willing districts with significant problems of teacher absenteeism in their remote, disadvantaged villages. Based on lessons learned from the operational pilot, we exclude districts with very weak governance and with transitory communities (i.e. fishing and the bush communities). To ensure manageable implementation costs, we excluded districts with very high transportation costs.[12] We also excluded conflict-prone areas and districts that were part of many other education pilots. Finally, we limit the districts to those that had at least 40 primary schools in rural areas that fulfill our definition of eligible schools described below. Our final list included three districts in West Kalimantan (Ketapang, Sintang, and Landak) and two districts in East Nusa Tenggara (Manggarai Barat and Manggarai Timur).

We only include schools under the MOEC that satisfy four eligibility requirements. First, each school must have a minimum of 70 registered students. Second, since the PPM interventions link evaluations to the TSAs, at least 3 of its teachers must receive the TSA in 2017. Third, schools must satisfy a remoteness criterion of being located in a village that was at least one-hour drive away from the district capital.

[11] To accommodate the use of the smartphone camera, the facilitators held an additional training to use it during the monthly community meeting. Moreover, SAM+Cam schools added verification of the camera reports to its monthly meeting agenda.
[12] For example, we exclude Papua, and certain districts in East Nusa Tenggara and Central Sulawesi.

Our data suggest that on average, participating schools are located around 40 km (and about a two-hour travel time) from the subdistrict office. Finally, we allowed for a maximum of two primary schools (instead of one) per village to be part of the project due to budgetary reasons.[13] More than 90 percent of the schools in our sample are public schools.

3.3 Treatment Assignment and Compliance

We use a stratified-random assignment procedure to assign schools to control and treatment groups. Each stratum has four villages. The similarity of schools within each stratum is determined by the following variables: village access to a mobile phone signal, the total number of teachers in the school, the share of teachers with the teacher registration number — which is a TSA prerequisite — and the exit-exam test scores obtained from the ministry. Villages with two schools were, to the extent possible, grouped with other villages with 2 schools resulting in strata with 8 schools. The last stratum with fewer than 4 two-school villages was assigned single-school villages instead to complete the assignment. This ensures that two schools in the same village always received the same treatment. Except for this stratum, all other strata had villages with equal number of schools. We detail the stratification procedure in Appendix A.

During the baseline survey, we discovered that three schools in Manggarai Barat were not in the villages indicated by the administrative data used for the initial treatment assignment. In all three cases, these schools were in villages with a school already participating in the study. Since all schools in the same village should be assigned to the same treatment group, we randomly reassigned the treatment status for schools in the three affected villages. The reassignment took place before the start of the intervention.

Moreover, a few weeks before the intervention started, the education ministry changed its mechanism to define eligible TSA locations. It used a national index instead of district head recommendations to determine eligibility, where all registered teachers working in these villages would automatically be eligible. This change took away the TSA eligibility of three villages. These affected schools were all part of the control group. As we discuss below, we control for these three schools in our empirical analysis.

3.4 Details on Implementation

Before discussing our results, we discuss some additional implementation details and report community response to the interventions. We derive most of the materials in this section from data collected from the process monitoring, as part of project management.

[13] To maintain a reasonable implementation budget, we excluded sub-districts (kecamatan) with less than four eligible primary schools and those requiring costly additional travel requirements (e.g. using boat/plane just to reach that specific sub-district). We found less than 270 villages with eligible primary schools. To obtain 270 schools, we needed to have more than 1 school in some of the villages. We therefore randomly chose 170 villages to have a single school participating, and 50 villages to have 2 schools participating in KIAT Guru. In two-school villages, our randomization procedure ensured that both schools received the same treatment. Furthermore, in villages with more than the assigned number of schools, we randomly selected the participating school(s).

3.4.1 The Social Accountability Intervention

The set of seven meetings to set up the service agreement and the user committee were conducted between November 2016 and June 2017. Details on these meetings were retrospectively collected in 166 schools during monitoring visits. An average meeting took 3.3 hours and the seven meetings were completed in an average of 38 days. While each facilitator was assigned to between four and six schools, the initial set of meetings in 57 percent (95 of 166) schools were facilitated by two or more facilitators due to personnel safety reasons and different strategies taken to encounter various logistical and geographical challenges. The meeting(s) to formulate the service agreement and teacher-specific scorecards took the longest time.[14] The process monitoring and several focus group discussions with facilitators throughout the implementation did not identify differences in how the facilitators conducted these meetings in all treatments.

Service Agreement and the Scorecard. Initially, the second-most common indicator (after the requisite teacher-presence indicator) was a safe environment free of physical and verbal abuse — an indicator whose importance was emphasized during the socialization process. Other indicators were on improving learning (e.g., implement various ways to teach and enhance understanding; improve reading, writing, and counting; provide additional lessons, provide feedback to students), motivating students, introducing students to social and cultural norms, communicating with parents, and improving teacher behaviors and conducts. Appendix Figure F.1 shows an example of the scorecard.

Following a meeting designated to allow UCs to revise their indicators (around August 2017), we find an increase in indicators that focused on the student learning process from 33 to 48 percent.[15] At the same time, we find the committees were most likely to drop the corporal punishment indicator that teachers felt was too difficult to implement.[16] The scorecard revision meetings took place in all treatment schools between July 2017 and January 2018. At the end of 2017, the project facilitators left the villages, and starting in 2018, the village cadres facilitated the monthly and evaluation meetings. The cadre-facilitated meetings in treated villages continued (at least) until our follow-up survey in mid-2019.

Figure 2 shows the evolution of the mean scores over time between August 2017 and March 2019. Average scores are generally high, in the range from 94 to 98 on a 100-point scale. The scores given for SAM+Score are slightly higher than those given in SAM and SAM+Score. The trends indicate that average scores gradually increase over time.

User Committee and the Monthly Evaluation Meetings. Most village cadres and UC members did not change until the endline. The follow up survey only collected data from the UC in SAM and SAM+Cam treatments, with three UCs reported as inactive and 26 percent members being replaced. Compared to the endline, the follow up survey found improvement in female membership of the UC, from 46 to 48 percent, and in those with more than a secondary school education, from 27 to 31 percent. From the

[14] In 40 percent of the schools, this process took between three to seven hours, and in the rest of the schools, the process required two or more meetings.

[15] These learning-oriented indicators include, among others, actions to improve student literacy and numeracy skills, and teachers making lesson plans and using various learning tools and props.

[16] Some of the difficulties arose from deeply entrenched cultural norms. Information collected from the qualitative research and process monitoring indicate that when corporal punishment was not allowed, teachers and parents alike found it difficult to discipline students. Since the project did not train parents or teachers on strategies to conduct positive discipline for children, they could not find alternative strategies to address the situation.

endline survey, 26 percent of the village cadres were female, with the majority having a high-school degree or higher.

We find variations in the way monthly meetings were conducted. In some villages, UC members and teachers conducted face-to-face evaluation of the scorecards. In others, the UC members gave the scorecard results to the village cadres, to be delivered to the teachers.[17] By the end of 2017, meeting facilitation was fully managed by the village cadres except in one school. Village-level implementation ended in 2017 and from 2018 to mid-2019, the project only provided district-level technical support; however, the UCs continued to send monitoring reports to the district officials which in SAM+Cam and SAM+Score continued to determine cuts to the TSA. In 2017, 83 percent of the treatment schools received funding from village heads to provide operational costs for monthly meetings and incentives for the village cadres and UC members. By 2018, all village heads allocated funding for all treatment schools. The amount and allocation of funding provided by village heads ranged widely.[18]

3.4.2 Performance Pay

Two issues affected the early implementation of the performance pay mechanism. First, administrative holdups delayed the implementation of the incentive payments for approximately 15 percent of the 135 SAM+Cam and 3 schools. Out of 135 schools, 113 had their first evaluation meeting between April and May 2017, and received their first incentive payments in April 2017. The remaining 22 schools affected by the holdup held their first meeting between June and July 2017. By October 2017, all 135 schools have received their incentive payments. Second, due to the end-of-year budgetary account closure, TSAs for the second half of November and December 2017 and 2018 were paid in full irrespective of the scorecard.

We find clear evidence that the scorecard did determine cuts to the allowance as stipulated by these treatments. TSA teachers in SAM+Score received an average pay cut of around 6.9 percent, whereas teachers in SAM+Cam received an average cut of 10.1 percent. Furthermore, we find strong evidence of compliance of the pay-for-performance rule for SAM+Cam. In SAM+Cam, TSA teachers will receive no allowance if their presence score fell below 85 percent and will receive an allowance whose share is a linear function of their presence score at 85 percent and above. A plot of the payment cut against the presence score shows that the payment schedule was applied correctly 97 percent of the time (Figure 3).

4 Data and Empirical Strategy

4.1 Data Collection

Student Learning Assessments. The research team developed its own student learning assessments (SLA) instruments to assess basic functional literacy (in Indonesian) and numeracy competencies along the learning continuum standards set in the 2006 national curriculum. They are designed based on frameworks and findings from other assessment tools (ASER Centre, 2014; Uwezo, 2012; Gove and Wetterberg, 2011; Platas et al., 2014) and they consist of: (i) a diagnostic test which aims to quickly capture

[17] Focus group discussions with facilitators suggest that village-specific idiosyncracies —e.g., cultural norms and initial resistance from teachers to have their performance being evaluated so openly — drove these differences.

[18] The average per district ranged from IDR 1.471 million in Sintang to IDR 9.022 in East Manggarai. Within the same district, for example in Sintang, the range starts from minimum of IDR 750,000 to IDR 6.4 million.

students' competencies in literacy and numeracy; and (ii) an evaluation test which maps students' more specific abilities along the literacy and numeracy learning continuum.

Separate test booklets were developed for each elementary grade level with multiple-choice items consisting of 15 percent grade-level, 65 percent one-grade-below, and 20 percent two-grade-below. Overlapping items across grades made it possible to vertically link scores across grades and thus assess these tests using item response theory (IRT). For the baseline survey, the evaluation test was administered for all students in grades 1 to 5 in participating schools, on a one-on-one basis for grades 1 and 2, and on a group basis for grades 3 to 5. At the endline, another evaluation test was administered to the same set of students, the majority of whom were in grades 2 to 6, as well newly enrolled grade 1 to 6 students who did not participate in the baseline survey. Due to budgetary constraints, the follow-up survey was administered to all students for grades 3 to 6 only.[19]

Teacher Absence Survey (TAS). The instrument originated from the World Bank's multi-country teacher absence survey (Chaudhury et al., 2006), which calls for an unannounced visit to schools during normal school hours to obtain a representative estimate of teacher absence from school. The instrument has since been adapted for various TAS implementations in Indonesia. We adapted the design and methodology of the TAS were from the Analytical and Capacity Development Partnership (2014) study in Indonesia, with additional inputs from the instruments used in UNICEF (2012) study in Papua and West Papua. In its implementation, the enumerators implemented the TAS on the day of arrival which were unannounced.

Survey Instruments. In addition to the SLA and the TAS, we interviewed: (i) school principals; (ii) teachers; (iii) a random sample of 20 households of children in primary-school-age-attending school (4 from each of grades 1 to 5 at baseline) and all panel parents; (iv) school committee; (v) the village head; and (vi) the user committee (for the endline and follow-up surveys). We collected a rich set of measures to capture their characteristics, perceptions of the education quality and other education stakeholders, as well as the relationships between parents, teachers, school committee members, and the school principal. For parents, we collected detailed information on their monetary and time investments in their children's education. The questionnaires were adapted from previous surveys conducted by the World Bank and others (Hasan et al., eds, 2013; Chang et al., 2014; World Bank, 2015, 2016; ACDP, 2014).

Behavioral Experiment. We also conduct a lab-in-the-field behavioral experiment to measure school-level norms related to the willingness to provide public goods and punish free riders. The experiment was implemented in 182 randomly-selected schools out of the 270. The experiment involved between 16 and 20 teachers and parents associated with each school playing a simple Public Good Game followed by a Public Good Game with Punishment using paper-and-pencil instrument (Fehr and Gächter, 2000; Barr et al., 2012). Appendix Section D provides details of the implementation of the experiment.

Data Collection Timeline. An independent survey team collected the survey data, while project facilitators and project implementation team collected the monitoring data. Figure 1 shows the implementation timeline. The baseline survey was conducted in October and November 2016 for 213 schools and

[19]For grades 1 and 2 students, the evaluation test was only conducted for students who were part of the earlier surveys.

completed in February 2017 for the remaining 57 schools. An endline survey was conducted in February until mid-April 2018, soon after the facilitators handed over facilitation to village cadres at the end of 2017.[20]

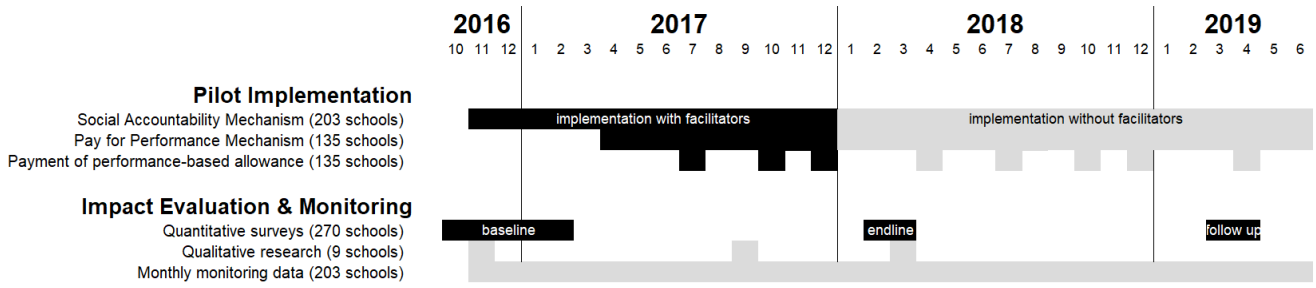


Figure 1: Implementation Timeline

In March to May 2019, we conducted a follow-up survey to investigate whether the impact of the interventions persisted more than a year after the implementation ended. The follow-up survey coincided with a plan by district governments to expand the SAM+Cam treatment to other schools in the 2019/2020 academic year. However, given budgetary limitations and the government's expressed interest in SAM+Cam, we only collected the follow-up survey in schools that were part of the SAM and SAM+Cam treatments, as well as the control group. As mentioned above, we also did not collect learning outcomes from new grade 1 and 2 students.

4.2 Baseline Characteristics and Covariate Balance

Table 2 presents the summary statistics of student, teacher, and parent characteristics at baseline for the control and treatment groups. We observe poor literacy and numeracy among students in participating schools. Their mean scores from the Indonesian and mathematics learning assessments at baseline were 37.5 and 37.7 (out of 100). The student population was 53 percent male and more than 80 percent of students have parents with only a primary education or less.

Teacher accountability — indicated by teacher absenteeism rate and observed in-school activities — is low. Our baseline teacher absence survey recorded almost 20 percent absenteeism rate. Among those who were in school, only three quarters of the teachers were observed to be doing work and a quarter of teachers were not teaching when they were observed in class.[21]

Despite low teacher efforts, parents were not aware of these problems. At baseline, about 90 percent of parents believed that the quality of their children's school is either good or very good. Furthermore, only slightly more than one in five parent respondents reported teacher absence as one of the three main problems afflicting education in their community. At the same time, we find limited parental supervision of their children education: at home, children were accompanied when learning for about 2.5 hours a

[20] An alternative we considered was to wait with the data collection until Oct 2018. We decided to conduct this round earlier rather than later because we were concerned about fade-outs as a result of the Ramadan, holidays, and class transitions which followed right after (May through June 2018). In addition, we would have lost a cohort of students if we had to wait until after the class transition. The downside of the decision was that the baseline and endline were administered in different months, which could result in seasonality affecting our results.

[21] We define "teaching" as performing teaching and other academic activities such as grading or giving quizzes.

week.

Appendix Tables F.4–F.5 present the balance tables for student, teacher, and parent characteristics. The tables show that the covariates are mostly balanced across control and treatment groups. We find a few statistically significant differences from the control group for a particular treatment and a particular outcome, which is to be expected from a random assignment. Our preferred specification includes these covariates as control variables.

5 Impact on Student Learning Outcomes

Our primary outcome of interest is student learning.[22] We estimate the treatment effects by regressing the following model:

$$Y_{ijt} = \alpha_k + \sum_{r \in R} \gamma_r T_{rj} + X_{ij}^{'} \beta + \delta Y_{ij0} + \lambda \bar{Y}_{j0} + \varepsilon_{ijt} \quad (1)$$

where Y_{ijt} = the student learning outcome for individual i in school j at time $t \in \{1, 2\}$ (i.e., baseline, endline, and follow-up); α_k = the fixed effects for strata k; and X = control variables. T^{rj} is the dummy variable for school j's treatment regime r, and γ is the average treatment effect. Our preferred specification controls for student and school characteristics, as well as individual (Y_{ij0}) and school-mean outcomes ($\bar{Y}_{j0}$) at baseline. We also include dummy variables to account for individuals with missing control variables. Standard errors are clustered at the school level. We also report the *p-values* from a randomization inference test of the sharp null of no effect for each individual treatment, holding other treatments' assignments constant.[23]

5.1 Main Results

Table 3 presents the impacts of the treatments on learning outcomes and grade repetition. For the learning outcomes, we facilitate comparison across studies by constructing grade-adjusted standardized scores. The mean outcomes for students in the control schools (including the raw unstandardized scores) are presented along with tests for cross-treatment differences in coefficient estimates in the panel below the coefficient estimates. Odd-numbered columns present estimates of the one-year (2018) impacts for all treatments and even-numbered columns present the two-year (2019) impacts for SAM and SAM+Cam.[24] As a robustness check, Appendix Table F.6 presents the results of regressions without the control variables.

We find that the one-year impacts on student learning outcomes were positive for all treatments, but they were stronger and more persistent for SAM+Cam. Columns 1 and 3 show the one-year impacts for Indonesian and Mathematics. SAM treatment improved Indonesian and mathematics outcomes by 0.09 and 0.07 standard deviation (s.d.) respectively. SAM+Score yielded similar learning impacts of around 0.12 and 0.10 s.d. for Indonesian and mathematics. However, SAM+Cam yielded impacts that were three halves and twice as large (0.19 and 0.20 for Indonesian and mathematics).

[22] The pre-analysis plan for this study is documented in Bjork et al. (2018).
[23] The test is based on the user-written Stata command *ritest* (see Heß, 2017).
[24] As we discussed in Section 4.1, we did not survey the SAM+Score schools in 2019 due to budget constraints.

We also find that the SAM+Cam impacts were persistent going into the second year. Column 2(4) shows that the two-year impact of SAM+Cam on Indonesian (mathematics) fell by close to 50 (15) percent to 0.09 (0.17) s.d. However, these comparisons may have underestimated persistence of SAM+Cam. Our estimate by grade (Appendix Figure F.3) shows that the impacts of SAM+Cam are stronger for lower grades. Hence, our follow-up data collection strategy that mostly excluded grades 1 and 2 students may have artificially biased the 2019 average treatment effect estimates downward. Appendix Table F.8 presents the results when we restrict the sample to students who would have been at grades 3–6 in 2018 and 2019. Here, we find a much weaker decay in the impact on Indonesian and no decay in the impacts for mathematics.

In Appendix C, we show that we can decompose the two-year impacts of SAM+Cam into knock-on impacts from the first-year implementation and new impacts in the second year. Appendix Table C.1 show that about half of the two-year impacts of SAM+Cam was new impacts that were realized in the second year. This is an important result. Since project facilitators had left these communities at the end of the first year, this finding confirms that the institutional setup inherited by the SAM+Cam treatment continued to improve learning in these communities.

We did not find similarly persistent impacts for SAM. The decays in learning impacts — using either the full or restricted grades 3–6 sample — were steeper for the SAM treatment. Using the full sample, the two-year impact on Indonesian was negligible (0.01 s.d.) while its two-year impact on mathematics was small (0.04 s.d.), close to half of its one-year impact and not statistically significant.

5.2 Retention and Attrition

We do not find evidence that our treatments affected the school's grade retention strategy. Columns 7 and 8 suggest that there was no one- or two-year impact of any of the treatments on students' grade repetition. This result provides a reassurance that our estimated learning impacts were obtained as students went through their normal progression through primary school, instead of being driven by the impact of the interventions on the school's grade retention strategy. Appendix Table F.7 shows that the learning impact estimates are robust to an IRT correction that accounts for grade retention.

Another potential bias could arise from systematic attrition. If schools in the treated groups selectively encouraged better students to take the SLAs in the endline and follow-up, our findings could be biased upward. We therefore use data on the universe of students who participated in the SLAs across periods to examine their attrition pattern. Table 5 presents the regressions of the student's attrition on his/her schools' treatment status. Columns 1 and 3 show that students in the SAM+Score treatment are less likely to attrit compared to the control schools. However, using interactions between the student's SLA performance at baseline with his/her school's treatment status (columns 2 and 4), we find no evidence of selective attrition based on academic ability.

5.3 Heterogeneity Analysis

In this section, we study the heterogeneous impacts of our interventions. To estimate these heterogeneous impacts, we estimate the following regression:

$$Y_{ijt} = \alpha_k + \gamma_h Z_{ij0} + \sum_{r \in R} \gamma_r T_{rj} + \sum_{r \in R} \gamma_{rh}(T_{rj} \times Z_{ij0}) + X_{ij}^{'}\beta + \delta Y_{ij0} + \lambda \bar{Y}_{j0} + \varepsilon_{ijt} \quad (2)$$

where Z_{ij0} is the baseline variable we use for the heterogeneity analysis, γ_{rh} is the differential impact for the subsample of individuals defined by Z, while the other variables are as defined in Equation 1.

Table 4 presents these results with heterogeneous variables as the column headers. Columns 1–2 show that the impacts of these interventions are gender neutral. We also examine whether a student's exposure to a TSA teacher strengthened the impact of the interventions. Since the SLAs were deployed in the middle of the second semester of an academic year, students could potentially be taught by two different teachers between two survey waves. In columns 3–4, we use the total number of years (from the baseline year) a student were taught by a TSA teacher at each respective year to capture the heterogeneous impact of an additional year of exposure to a TSA teacher. Our results suggest having an extra year of TSA teacher did not strengthen the benefits of the interventions.

We find that these interventions tend to have stronger positive effects on more able students. Columns 5–6 show that students whose baseline SLAs were above median *within their school* benefited more in terms of learning improvements, especially from the SAM+Cam treatment. Columns 6–7 suggest qualitatively similar, albeit more noisily estimated, effects among students whose baseline SLAs were above median *across all schools* for both SAM and SAM+Cam treatments, but not in the SAM+Score treatment.

Interestingly, we also find some evidence that the positive impacts of our interventions are stronger for weaker schools. Columns 8 shows that the one-year impact of the interventions do not differ by school quality (as measured by the school-level average of the baseline standardized SLA scores). However, our estimate of the two-year impact in column 9 suggests the temporal decay of the learning impacts of SAM and SAM+Cam treatments primarily occurred in above-median quality schools at baseline. For the below-median quality schools, the two-year impact of SAM only decayed slightly relative to the one-year impact (from 0.11 to 0.08 s.d.), while the two-year impact of SAM+Cam increased relative to its one-year impact (from 0.16 to 0.19 s.d.). The decay in the above-median schools were larger, in part driven by a large increase in the performance of above-median schools in the control group.

6 On the Mechanisms: Teachers, Parents, and School Management

Our institutional innovations are designed to improve parent, teacher, and school inputs into student learning. This section studies how our interventions affect these intermediate outcomes. To fix ideas, we begin with a simple model of parent-teacher interactions to frame how our interventions affect parent and teacher inputs into student learning. We draw from the literature on sustaining cooperation under weak institutions (Gerber and Wichardt, 2009; Han, 2016) and performance contracts (Baker et al., 1994; Holmstrom and Milgrom, 1991; Macleod, 2003). The model provides a framework for the empirical estimates that follow.

6.1 A Model of Parent-Teacher Interactions

6.1.1 Efforts under Weak External Supervision

Consider a remotely located school where government inspectors are unable to monitor teacher performance. Parents and teachers receive positive utility from student's expected level of learning. Learning is a function of parent and teacher efforts that incur utility costs. There is effort complementarity in the learning production function: if teachers put more effort, the marginal return of parental effort increases. We assume a linear production and cost functions, and two levels of effort (high and low) that teachers and parents can put into improving student learning. We also assume that the utility cost of effort for parents is lower than that for teachers.

Appendix B shows that under these assumptions, parent and teacher payoffs under the status quo can be summarized in Table 6a. T(P) indicates the payoff for teachers (parents) and the number indicates its magnitude, with 1 indicating the lowest payoff and 4 indicating the highest. Parents receive the highest payoff when both parents and teachers put in high effort. Note that these payoffs are based on the observed effort levels rather than some objective measure of student learning — an assumption that will be important when we introduce imperfections in the measurement of teacher efforts in Section 6.1.3. Conditional on parental effort, teachers obtain a higher payoff when they put in little effort. This asymmetry arises because the utility cost of effort is lower for parents. Under the status quo, the pure Nash equilibrium is when both teachers and parents put in low effort.

6.1.2 The Role of Commitment Contracts

The purpose of the SAM component of our interventions is to commit teachers and parents to providing a high effort level through the service agreement. It calls upon the moral responsibility of teachers and parents for student learning to commit them to increase their effort levels that are higher than those at baseline. Because the willingness of teachers to cooperate with this process will vary, the standards of what constitutes a high effort is left as part of the negotiation process.

As in Gerber and Wichardt (2009), we can model this commitment as an upfront payment. Teachers' and parents' commitment to providing a high effort documented in the service agreement can be modeled as upfront payments (or "deposits") d_t and d_p respectively. If they break their commitment by providing a low effort, they lose these deposits. Because $P4 > P3$, the parental commitment (d_p) necessary to maintain the high-learning Nash equilibrium is zero as long as teachers put in a high effort. For ease of exposition, we set $d_p = 0$. Table 6b presents the payoff matrix under this setup. Teacher commitment, d_t, needs to be greater than $T4 - T3$ to make the high-learning equilibrium feasible. For $d_t > T2 - T1$, the high-learning is the only equilibrium.

The interventions vary the way teachers are penalized for falling short of their commitments. In the SAM intervention, breaking the commitment will be costly to the teacher's reputation. Teachers participate in the formulation of the service agreement, which is publicly announced and subsequently monitored. In a remote community where the teacher is also often a respected member of the community, this "naming and shaming" may be a sufficient threat. Meanwhile, by adding performance pay components, SAM+Cam and SAM+Score interventions raise the stakes for TSA-receiving teachers who could also lose (part of) their allowance if they received a low score on absenteeism and/or other indi-

cators. Ceteris paribus, these enforceable commitment contracts would increase the chance for a high effort equilibrium.

6.1.3 Imprecise Measurements and Retaliations

The model so far generates no clear hypothesis on the relative effectiveness of SAM+Cam v. SAM+Score in incentivizing high efforts. On the one hand, SAM+Score could have higher impacts on learning because all aspects of the service agreement count for the teacher incentives. Assuming that the score cards were formulated to maximize learning, SAM+Score should be most effective in improving learning outcomes. In contrast, by putting stronger incentives on teacher absenteeism, SAM+Cam could result in teachers neglecting other aspects of their jobs that are important for learning (Holmstrom and Milgrom, 1991). As this incentive effect is well understood, we do not incorporate it into our model.

On the other hand, SAM+Score could perform worse because it relies to a greater extent on subjective indicators. This causes two problems. First, the measurements of subjective indicators tend to be imprecise, which could result in parents and teachers having conflicting assessments of teacher effort. This introduces noise in the assessment, which weakens the incentive (Baker et al., 1994). Second, since subjective indicators open up a room for negotiation, teachers can try to pressure parents to increase their ratings. Anticipating such behavior, and depending on the bargaining position of teachers, this could result in parent ratings that are too lenient (Macleod, 2003; Marchegiani et al., 2016). SAM+Cam largely avoids these problems because the additional incentive is based on absenteeism as recorded by a camera, which leaves little room for subjectivity.

We introduce three features to the model to study the potential effects of subjective indicators. First, we model the consequences of the imprecision of subjective indicators by introducing two probabilities, π_o and π_u, which are taken as given. To simplify, let us assume that teachers can precisely measure their own effort. Let π_u be the probability that the parent *under*-rates the teacher, that is, the teacher provides high effort while the parent assesses it as low. Similarly, the π_o be the probability that the parent *over*-rates the teacher.[25] More subjective indicators will have higher π_o's and/or π_u's. Second, we allow for teachers to retaliate if they were underrated. The exogenous utility cost of retaliation to parents is indicated by R. Teachers with a stronger bargaining power can incur a higher R. Finally, we allow parents to choose either a strict or lenient assessment regime. In a strict assessment regime, parents always report what they observe; otherwise, they always report a high teacher effort.

Figure 4 shows a version of this extended model in sequential form. We consider the more interesting case where parents can credibly commit to the assessment regime ex-ante, before effort levels are realized.[26] Hence, parents will first choose between the strict and lenient assessment regime. Next, teachers and parents simultaneously decide on their effort. With positive π_o and π_u, parents might over-/underrate teacher effort. Parents' payoffs are based on the perceived effort of teachers; teachers' payoffs are based on their actual effort level. If a teacher was underrated, s/he could retaliate by imposing a utility cost R to parents.

[25] Of course, teacher measurement of their own effort could be upward biased. In such a case, we can define under/over-rating as the divergence between parental assessments and what a teacher believes to be the effort level that s/he has provided.

[26] Alternatively, parents may not be able to credibly commit to the strictness of their assessment regime ex-ante and instead, decide on it ex post after the realization of effort levels. In that case, parents would always be lenient to avoid the risk of retaliation, and thus there would be no incentives for teachers to perform.

Table 6c presents the payoff matrix under imperfect monitoring with possible teacher retaliation under the different assessment regimes. If parents choose to assess leniently, then teachers will never be punished and thus will never retaliate; otherwise, under the strict assessment regime, $d_t > 0$ and $R > 0$. When parents choose to assess leniently, the Subgame Perfect Equilibrium (SPE) yields a low effort of teachers. For parents, their tendency to overrate the teacher, π_o, would determine their effort level: If parents assess teacher effort correctly ($\pi_o = 0$), they will provide a low effort; if they always overrate ($\pi_o = 1$), then parents will provide a high effort. When parents choose to assess strictly, positive π_o and π_u weakens the relation between teacher effort and punishment and thus the likelihood that teachers will put in a high effort.

Retaliation could affect the likelihood that parents choose a strict assessment regime. To see the effect of retaliation, consider the case where d_t and π_u/π_o allows for either the low/low and high/high effort equilibrium. Parents will choose the strict regime if their high/high-equilibrium payoff under that regime exceeds their low/low-equilibrium payoff under the lenient regime, to wit, $(1-\pi_u).P4+\pi_u.(P1-R) > (1-\pi_o).P2+\pi_o.P3$. A higher level of retaliation R and a higher probability to underrate π_u will increase the likelihood that parents choose to assess leniently.

In summary, the model predicts that a greater emphasis on subjective indicators would diminish the chances for the high/high equilibrium. This is because the use of subjective indicators weaken the relationship between the effort level and punishments for teachers while increasing the likelihood for teachers to become disgruntled (because they are underrated) and retaliate against parents. This would incentivize parents to monitor leniently, which would eliminate the incentive for teachers to put in a high effort. If parents correctly assess teacher effort, they will also reduce their effort; otherwise, if they (mistakenly) overrate teacher effort, they might still put in a high effort.

6.2 Teacher Effort

Did the interventions improve teacher effort? First, we examine teacher activities during school hours. We use three variables from the TAS data set — collected through enumerator observation during unannounced visits (see Section 4.1) — as outcomes, namely whether: (i) a teacher is present when s/he is scheduled to be; (ii) a teacher who is present is observed to be working; and (iii) a teacher who is in class is observed to be teaching. We limit our teacher sample to classroom teachers who were responsible for teaching Indonesian and mathematics for these primary school students.[27] Furthermore, since SAM+Cam and SAM+Score interventions incentivizes TSA teachers, we estimated the heterogeneous impact of the TSA status on teacher effort.

Table 7 presents the one-year and two-year impacts. Panel A suggests presents the overall treatment effects on teacher effort ranged from negative to weak positive. In particular, after a year of implementation, SAM+Score reduced overall teacher presence and whether teacher were observed to be working in school by 6.3 and 7.6 percentage points respectively.

When we examine the effects by teacher status, we find that a lot of these negative effects were driven by non-TSA teachers. Panel B shows the regression results when we interacted the treatment status with the teacher's TSA status. Column 1 shows suggests that the non-TSA teachers were driving

[27] In other words, we exclude subject teachers, who typically are physical education or religion teachers.

the one-year mean effect estimates toward zero (or negative), especially in SAM+Score. We find a small but imprecise treatment effect (of 0.05 percentage point with a p-value=0.15) on the presence of TSA teachers in SAM+Cam, but not in other treatments. By the second year, however, the effect on TSA teachers' presence disappeared (column 2).

Columns 3–6 of Panel B exhibited similar patterns with regard to observable teacher efforts. They suggest that in the first year, non-TSA teachers tended to decrease their efforts relative to TSA teachers. The reductions in efforts among non-TSA teachers were more salient for SAM+Cam and SAM+Score treatments, where the TSAs were incentivized. The one-year impacts on TSA teachers' efforts were more muted but somewhat more positive for the two performance pay interventions.[28] Similar to the effects on presence, they were: (i) stronger and more precisely estimated for SAM+Cam; and (ii) the positive effects on TSA teachers virtually disappeared for SAM and SAM+Cam by the second year.

A potential drawback of the TAS data is that they are collected from a single spot-check visit for each survey period. As an alternative measure of teacher effort, we used data from the teacher survey and constructed a proxy of teacher inputs into student learning using their self-reported hours allocated to various school-related activities. We first identify activities that are positively correlated with student learning at baseline.[29] Once we have identified these learning-enhancing activities, we estimated the impact of the interventions on the total hours that teachers spent on activities that are positively correlated with learning.

Table 8 present the results. We first examine the impact of the treatments on the time that teacher spent on school-related activities. Columns 1–4 show that the treatments have no impact on the total number of weekly hours teachers spent on school-related activities. However, we show in Column 5 that in the first year, SAM and SAM+Cam increased the number of hours spent on learning-enhancing activities by between 8-8.5 percent out of a mean of 15.1 weekly hours. The strongest impact was observed for SAM+Cam. Column 7 suggests that the increase in the hours spent on learning-enhancing activities did not differ by teachers' TSA status. However, similar to the impacts on teacher activities observed in the TAS, these positive improvements disappeared by the second year.

6.3 Parental Investments in Education

One of the objectives of the SAM component is to encourage parents to participate more actively in their children's education. In this section, we report the impacts of our interventions on parents' financial and time investments in their children education. To measure parental investments, we look at their education expenditure, children's participation in paid work or family business, the total number of hours their children were accompanied when they were learning, and their engagement with the school.

[28] The one-year treatment effects (clustered standard errors in the parentheses) of SAM, SAM+Cam, and SAM+Score on the likelihood that TSA teachers were observed to be working were respectively 0.04(0.04), 0.08(0.04), and -0.02(0.04) s.d.; while the effects on whether TSA teachers were observed to be teaching when in class were respectively 0.07(0.04), 0.08(0.05), and 0.07(0.05).

[29] To identify activities that are positively correlated with learning, we estimated a regression of student learning outcomes on their teacher's allotted time to different activities at baseline. We use a post-double-selection lasso procedure (following Belloni et al., 2014) to determine the controls included in the regression. Activities whose coefficient is positive and statistically significant at 10% in included in the set of "learning-enhancing activities," namely: (i) in-school teaching; (ii) out-of-school additional intra-curricular lessons; (c) out-of-school scientific publications; and (d) out-of-school innovative activities (develop teaching tools, etc.).

Table 9 presents our results. After one year, all interventions showed some evidence of increased parental investments in their children, with SAM+Cam exhibiting the strongest impact. Education expenditures increased by about Rp 27,000 (approximately US$2) for SAM+Cam relative to the control-group average of Rp 324,580 (US$23), an 8.3 percent increase (column 1). The impact was smaller for SAM and the smallest for SAM+Score, and neither was statistically significant. In the same period, the interventions increased parents' willingness to forgo their children's contributions to the household economy: children participation in the labor market at both the extensive and intensive margins fell in all treatments (columns 3 and 5). Parents also increased the number of hours their children were accompanied when they were studying at home (column 7) and the number of meetings with the schoolteachers (column 9).

Most of these impacts persisted for both SAM and SAM+Cam well into the second year. Education expenditure, the number of hours of accompanied learning, and the number of parent-teacher meetings were higher in the SAM and SAM+Cam schools than those in the control group schools. SAM and SAM+Cam also reduced the number of hours their children participated in the labor market (column 6). However, column 4 suggests that by the second year, the impacts of SAM and SAM+Cam on the likelihood that children worked disappear.

6.4 Intervention Effectiveness and Punishment Norms

Our model highlights the importance of parents' credible commitment to a strict assessment regime to reach the equilibrium where both parents and teachers provide a high effort. Such a commitment importantly depends on whether the community is willing to punish violations to an agreed standard. Different societies may exhibit different willingness to punish standard violations (Ensminger and Henrich, eds, 2014). Societies that are unwilling to punish will not be able to effectively use performance-pay tools to induce accountability among teachers. To examine this hypothesis, we conducted a lab-in-the-field experiment at baseline to measure the different communities' willingness to punish and examine whether the punishment norm predicts the effectiveness of the interventions.

Using a public good game with punishment (similar to Fehr and Gächter, 2000), we construct a school-level continuous measure that captures the community's willingness to punish individuals with below-average public good contributions.[30] We conducted this experiment in 182 schools that were randomly selected from the 270 participating schools. Appendix D provides details on the design and implementation of this lab-in-the-field experiment and how we construct the school-level willingness-to-punish measure. Using this continuous measure, we then categorized schools into those with above-/below-median punishment norm.

We find that that the punishment norm plays an important role in the short-run effectiveness of the interventions. Table 10 presents the heterogenous impact of our interventions by the baseline punishment norm. Column 1 shows that our interventions had no impact on TSA teachers' presence in communities with below-median punishment norms. Instead, the one-year impacts on TSA teachers' presence primarily occurred in communities with above-median punishment norms. The marginal impact of having a stronger punishment norm was largest for SAM+Cam and weakest (and imprecisely estimated) for

[30] This measure captures the school-specific elasticity of the punishment with respect to how far below a session-mean a partner contributed.

SAM+Score.[31] However, column 2 suggests that these heterogeneous impacts on teacher presence did not persist. Columns 3–4 show the lack of heterogeneous impact of having an above-median punishment norms on the presence of non-TSA teachers and serve as a placebo check. We also found that learning improvements for the performance pay interventions were primarily driven by communities with stronger punishment norms and this differential impact persisted for SAM+Cam (columns 5–6).

The weaker impact on SAM+Score is consistent with our model on how subjective indicators and teacher retaliation could weaken the incentive for teachers to provide a high effort. We have evidence that subjectivity in the assessment created tensions between teachers and the user committee. A qualitative study in our treated schools suggest that teachers in SAM+Score often questioned the user committee evaluations, whose members typically were were less educated than these teachers (World Bank, 2020). The teachers' higher social stature in the community put them in a position to pressure user committee members to improve their score. Indeed, we find corroborating evidence from our survey of user committee members: Table 11 shows that user committee members in SAM+Score schools are more likely to be pressured to increase the evaluation scores and received threats for a low score than those in the other treated schools. Moreover, the intimidation and threats were more likely to happen in SAM+Score in schools whose principal believed that teachers in the school would be hard to replace.

6.5 External and Internal School Management

The credibility of the assessment regime becomes stronger if external and internal school management aligns its actions with monitoring outcomes of the user committee. Indeed, this was part of our theory of change: Monitoring results that were discussed in monthly meetings at the school were conveyed to higher authorities (such as the school inspector at the district education office), so that they could act on the information. Moreover, these interventions might introduce school principals with a more systematic way to monitor and evaluate their teachers. Tables 12 and 13 present our results on the impacts of the interventions on how external actors manage the schools and how the school principals manage their teachers.

Table 12 shows that only SAM+Cam meaningfully increased, albeit temporarily, external engagement and supervision. Column 1 shows that SAM+Cam increased the number of meetings with the subdistrict office by 1 out of a base of 2.2 meetings per year. It also led to a significant increase in the number of annual supervision visits by 0.8 from a base of 1.4 (column 3). We have a qualitatively similar, but quantitatively much smaller effects from the SAM intervention, while SAM+Score only increased the number of supervisor visits. However, the increases in external engagement and supervision did not persist into the second year.

However, we find that our interventions had persistent effects on how school principals evaluated teachers (Table 13). By the first year, all three interventions led to increases in the share of teachers who received any or routine interventions, the frequency of evaluation, and the likelihood that teachers were observed while teaching. For SAM and SAM+Cam, these evaluation practices persisted into the second year. We were not able to reject that the impacts of the different interventions were different from each

[31] In Appendix Table F.9, we show the results for other teacher effort variables, to wit whether they were working (teaching) when observed in school (class). The patterns of heterogeneous impacts by punishment norms across treatments were qualitatively similar to that on teacher presence.

other in both 2018 and 2019.

6.6 Summary Remarks

We find that even though all three interventions improved learning outcomes, the intervention that combined locally defined service agreement, community empowerment, and performance pay that was based on a narrow, objectively verifiable indicator (SAM+Cam) yielded the strongest and persistent impact. The one-year impacts of SAM and SAM+Score were comparable. The interventions yielded little lasting effects on observable teacher efforts; however, they improved the internal management by school principals by improving the regularity and frequency of teacher evaluation. They also increased parental investment in their children' education and engagement with teachers.

On the one hand, SAM+Cam suggests that performance pay mechanisms can effectively complement community monitoring to improve learning. On the other hand, the performance pay in SAM+Score offered no marginal improvement over SAM. The more subjective evaluation standards in SAM+Score combined with the opportunity for teacher to retaliate may have a role in dampening its impact. We provide a theoretical framework and empirical evidence that this may indeed be the case.

7 Discussion

7.1 Local Support for Reform

Successful policy reforms need political support. Even when a policy is extremely effective in improving learning, it is unlikely to be adopted if it leads to widespread dissatisfaction. In this section, we use measures of teacher and parent attitudes to examine the extent of local support for our interventions. We first look at whether our interventions made teachers feel unappreciated or reduced their job satisfaction. We then examine their impacts on parents' satisfaction or their children's learning and school quality, and whether the interventions altered their aspiration for their children's education.

Teacher Satisfaction. Under the status quo, teachers were likely aware of the (minimum) performance expected for their remunerations; at the same time, most were also aware that they could treat these standards as discretionary. Introducing routine evaluations that were tied to a performance pay mechanism could have heterogenous impacts of ambiguous directions on teacher satisfaction. On the one hand, TSA teachers who felt entitled to the allowance might consider these pay reforms unfair and feel less appreciated. On the other hand, non-TSA teachers — who were paid less for similar efforts and were less satisfied under the status quo — might consider such reforms fairer. Finally, regardless of TSA status, intrinsically motivated teachers could see these reforms as an affirmation of the importance of standards and, hence, an appreciation of the (intrinsic) worth of their job.

Table 14 presents the impacts of our treatments on various aspects of teacher satisfaction. Columns 1–4 of Panel A show that on average, all interventions led teachers to be more satisfied of the appreciation from district education officials and other villagers. For SAM and SAM+Cam, this increased satisfaction of outside appreciation persisted into the second year. Our analysis of the heterogeneous impacts by TSA status (Panel B) suggests that there was no statistically significant difference in satisfaction of outside appreciation between TSA and non-TSA teachers.

Columns 5–6 of Panel A suggest that after one year, all three interventions improved teacher satisfaction of their salary. Interestingly, the heterogeneous impact analysis in Panel B suggests that the one-year impacts on salary satisfaction were much stronger for non-TSA teachers, even though their salaries were unaffected by our treatments. Furthermore, the impacts on non-TSA teachers' salary satisfaction were stronger in the two performance pay interventions. The (total) impacts on TSA teachers were more muted, but somewhat positive for the two performance pay interventions.[32] For SAM and SAM+Cam treatments, the second-year impacts on salary satisfaction were qualitatively similar.

Meanwhile, columns 7–8 suggest overall increases in job satisfaction across all treatments that were primarily driven by non-TSA teachers. These improvements among non-TSA teachers were again stronger for the two performance pay interventions. In contrast, the total impact of the interventions on TSA teachers' job satisfaction were negligible.[33] However, by the second year, the positive impact of SAM+Cam on non-TSA teachers' job satisfaction had completely disappeared.

Overall, these results alleviate concerns that performance pay schemes would lead to widespread dissatisfaction among affected teachers.[34] If anything, our results suggest that incorporating SAM and performance pay mechanisms into the hardship allowances made teachers feel more appreciated by officials and their community. These reforms, especially those with performance pay components, improved teacher satisfaction about their remunerations — and interestingly, even more so among non-TSA teachers whose remunerations were unaffected by these allowances. This last finding suggests that these conditionalities might have made allowances seem fairer to non-recipients.

Parent Satisfaction and Aspirations Table 15 presents the impact estimates on parents' satisfaction with the their children's school and learning, as well as on the education aspirations for their children. All three interventions improved parents' view of the school quality which were generally high: among control schools, 91 percent of parents rated their children's school as either good or very good. Columns 1–2 show that overall, the interventions increased this by about 5 p.p. after one year, and these positive improvements persisted into the second year for SAM and SAM+Cam.

However, the immediate impacts on parental view of the school quality did not immediately translate into satisfaction on their children's learning. Columns 3 and 5 shows that the interventions had no one-year impact on whether parents were satisfied with their children's learning results in Indonesian and mathematics. For SAM+Cam, their satisfactions of their children's learning results were significantly improved by the second year; however, we did not find any effect for SAM (columns 4 and 6).

All three interventions improved parents' educational aspirations for their children. Column 7 shows that parents' stronger agreement to the statement that they would prefer their children to go to college instead of working. These effects were not very different across interventions. Column 8 shows that these effects persisted for SAM and SAM+Cam: in the second year, the effects of these two interventions remained positive, but were smaller.

[32]The total, one-year impacts on TSA teachers' salary satisfaction for SAM, SAM+Cam, and SAM+Score were 0.024(0.151), 0.202(0.149), and 0.538(0.152) respectively (with clustered standard errors in the parentheses).

[33]The total one-year impacts on TSA teachers' job satisfaction for SAM, SAM+Cam, and SAM+Score were 0.049(0.054), 0.049(0.053), and 0.052(0.054) respectively (with clustered standard errors in the parentheses).

[34]The absence of aversion toward performance pay is in line with evidence elsewhere. Using the 1987-8 School and Staffing Survey, a comprehensive survey of about 9,300 public and 3,500 private schools in the United States, (Ballou and Podgursky, 1993) find a similar lack of hostility toward merit pay systems among teachers in districts that implemented them.

7.2 Cost Effectiveness

The investment cost of implementation for project facilitators was at USD 5,058 per school or USD 40 per student, which includes all costs made over the period of this study.[35] The cost was USD 506 per school or USD 4 per student higher for Group 2 schools, to cover for the purchase of mobile phones and the maintenance of the application. After one year of intervention, SAM+Cam improved learning outcomes by 0.2 standard deviation, at USD 44 per student. This means it costs USD 22 per student per 0.1 standard deviation increase. Starting in 2018, the annual cost to sustain SAM was USD 2,182 per school or USD 17 per student. Details on the cost calculation can be found in Appendix Section E.

Compared to other rigorously evaluated interventions in education that improved learning outcomes, the cost of KIAT Guru are on par with interventions that adopted similar approaches. To make our cost figure comparable to those reported in Glewwe and Muralidharan (2016) and JPAL (2019), we convert our cost to 2011 US dollars using US GDP deflators from 2011 and 2017. USD 22 in 2017 is equivalent to USD 20 in 2011. For SAM, the most comparable study is Pradhan et al. (2014), which was most successful in strengthen school committees in Indonesia through a combination of democratic elections of committees and facilitating joint planning with the village council, which costed USD 7.50 per 0.1 standard deviation increase in learning.[36] Three studies on Conditional Cash Transfer (CCT) grants improved learning outcomes with costs averaging USD 77 per 0.1 standard deviation increase. For PPM comparison, camera monitoring and teacher-presence-based payment in India costs USD 44 per 0.1 standard deviation increase, excluding the cost of staff, transportation, and monthly meetings. A teacher incentive intervention in Kenya costs USD 16 per 0.1 standard deviation increase, while another in India costs USD 1 per 0.1 standard deviation increase.

8 Conclusion

We show that interventions that incorporate a community-driven SAM and performance pay were effective in improving student learning. While the SAM treatment was effective on its own in improving learning, adding an appropriate performance pay contract to allow parents to hold teachers accountable can improve its effectiveness. This result suggests that a strategic approach (conceptualized, inter alia, by Fox, 2015b) that integrates social accountability with measures to increase public sector responsiveness outperforms a tactical approach that relies on information alone to generate collective action and influence public sector performance. We demonstrate that such a strategic approach can be implemented using an existing (district- or national-) government allowance. While implementing a pay reform on a government allowance all-at-once nationally may be daunting, our study suggests that it can be implemented piecemeal using asymmetric government regulations — in our case, issued only for our study districts. A similar approach can be utilized by governments in regions facing last-mile frontline service delivery challenges.

This study also showed that teachers, and other stakeholders in the community, accept performance

[35] Cost figures in Rupiah were converted to US dollars at an exchange rate of IDR 13,490 per USD, the average market exchange rate over the implementation period.

[36] This result is conditional upon receiving a grant of USD 870 per school committee. All school committees in the comparison, including the controls, were provided the grant. The grant by itself had no significant impact on learning outcomes.

pay. There were no large pushbacks and the surveys show positive impacts on satisfaction rates, suggesting that a scale up is also politically feasible. We also note that the treatments increased the satisfaction of non-TSA teachers the most. Perhaps teachers appreciate the fact that their TSA-eligible colleagues had to perform to receive the allowance. This is consistent with results from a separate survey of Indonesian schools that finds that individual teachers prefer performance-based over seniority-based pay (Perez-Alvarez et al., 2020).

Our findings also suggest that not all performance-pay contracts improve the effectiveness of a social accountability intervention. Policy makers need to pay attention to how features of its design could (de-)motivate service providers. We find that a simple contract based solely on monitoring presence complements social accountability better than a more comprehensive but less well-specified one. This is an important question that arises in many labor contracts (Baker et al., 1994; Khan et al., 2016). Note that this intervention was implemented in a context where the teacher (agent) has a higher social status, and is more knowledgeable about education, than the community (principal). We think this is one of the reasons why incentives based on presence worked best. It was an indicator where both community member and teachers felt comfortable with their assigned roles and there was little risk of disagreement over the rating. This simple contract also leads to less divergence between the performances of TSA v. non-TSA teachers — suggesting that teachers might perceive it to be fairer — and less conflict at the community level. Hence, we find that in our context of remote schools, a simple transparent rule that targets an incomplete but verifiable measure of performance works better than a comprehensive evaluation which is more prone to subjectivity.

Our results also show that impacts weakened when the facilitators left the village. While monitoring reports continued to be gathered, SAM did not have sustained impacts on learning while incremental effect of SAM+Cam weakened. This largely seems to be due to teacher efforts dissipating by the second year. Appendix Table F.10 suggests that school principals might have undermined their teachers' presence-based contract by providing them with excuses that will minimize penalties from their absences.[37] This result is neither unique nor surprising: A study of Indian nurses working in public health facilities similarly found that the administration was allowing them to claim more "exempt days" (Banerjee et al., 2008). It nonetheless suggests that policymakers need to prepare for the possibility that service providers would try to find loopholes that render the conditionalities ineffective. On the other hand, the impacts of SAM+Cam on parent inputs, school management, and importantly, on learning outcomes persisted — suggesting that some of the other changes to the learning environment were able to sustain the learning impact, despite the loopholes.

Given these results, the policy relevant question is how one can sustain the positive impacts that were achieved during project implementation at substantially lower cost. The follow-up results indicate that it is difficult to sustain effective teacher monitoring without external support. Some visits every couple of months are probably needed to energize stakeholders and to signal that local policy makers are taking this seriously. This begs the question to what extent these functions can be carried out by existing institutions affiliated with the school system. The interventions presented so far depended on the work of project facilitators, who worked with communities to establish a user committee — a novel

[37] Appendix Table F.10 shows that principals in SAM+Cam schools were more likely to report "off-school assignments" which would yield zero penalty to the TSA (Panel A) in place of excuses that would lead to non-zero penalty (Panels B and C).

institutional arrangement that did not exist prior to these interventions. It is an open question whether school supervisors and an existing institution within the school system, such as the school committee, can take its place and still replicate the success of this intervention.

References

Afridi, F., B. Barooah, and R. Somanathan, "Improving learning outcomes through information provision: Experimental evidence from Indian villages," *Journal of Development Economics*, 2020, *146*, 102276.

Analytical and Capacity Development Partnership, "Study on Teacher Absenteeism in Indonesia 2014," Technical Report, Ministry of Education and Culture 2014.

ASER Centre, "Annual Status of Education Report (Rural) 2013," Technical Report, ASER Centre, New Delhi 2014.

Baker, G., R. Gibbons, and K. J. Murphy, "Subjective Performance Measures in Optimal Incentive Contracts," *The Quarterly Journal of Economics*, November 1994, *109* (4), 1125–1156.

Ballou, Dale and Michael Podgursky, "Teachers' Attitudes toward Merit Pay: Examining Conventional Wisdom," *ILR Review*, October 1993, *47* (1), 50–61.

Banerjee, Abhijit V., Esther Duflo, and Rachel Glennerster, "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System," *Journal of the European Economic Association*, April 2008, *6* (2-3), 487–500.

Banerjee, A.V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani, "Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India," *American Economic Journal: Economic Policy*, 2010, *2* (1), 1–30.

Barr, Abigail, Frederick Mugisha, Pieter Serneels, and Andrew Zeitlin, "Information and collective action in community-based monitoring of schools: Field and lab experimental evidence from Uganda," 2012.

Barrera-Osorio, F., K. Gonzalez, F. Lagos, and D.J. Deming, "Providing performance information in education: An experimental evaluation in Colombia," *Journal of Public Economics*, 2020, *186*, 104185.

Belloni, A., V. Chernozhukov, and C. Hansen, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, April 2014, *81* (2), 608–650.

Bjork, Christopher, Arya Gaduh, Menno Pradhan, Jan Priebe, and Dewi Susanti, "Improving education in remote and isolated areas in Indonesia," *AEA RCT Registry*, 2018.

Bjork, Christopher Brian and Dewi Susanti, "Community Participation and Teacher Accountability: Improving Learning Outcomes in Remote Areas of Indonesia," Technical Report, The World Bank 2020.

Björkman, M. and J. Svensson, "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda," *The Quarterly Journal of Economics*, 2009, *124* (2), 735–769.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, December 2018, *168*, 1–20.

Breza, Emily, Supreet Kaur, and Yogita Shamdasani, "The Morale Effects of Pay Inequality," *The Quarterly Journal of Economics*, May 2018, *133* (2), 611–663.

Brixi, H., E. Lust, and M. Woolock, *Trust, voice, and incentives : Learning from local success stories in service delivery in the Middle East and North Africa*, Washington DC: World Bank, 2015.

Bruns, B., D. Filmer, and H.A. Patrinos, *Making schools work: New evidence on accountability reforms*, World Bank, 2011.

Chang, Mae Chu, Sheldon Shaeffer, Samer Al-Samarrai, Andrew B. Ragatz, Joppe De Ree, and Ritchie Stevenson, *Teacher reform in Indonesia: the role of politics and evidence in policy making*, Washington, D.C: World Bank, 2014.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers, "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *Journal of Economic Perspectives*, February 2006, *20* (1), 91–116.

Cilliers, Jacobus, Ibrahim Kasirye, Clare Leaver, Pieter Serneels, and Andrew Zeitlin, "Pay for locally monitored performance? A welfare analysis for teacher attendance in Ugandan primary schools," *Journal of Public Economics*, November 2018, *167*, 69–90.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia," *The Quarterly Journal of Economics*, May 2018, *133* (2), 993–1039.

Duflo, Esther, Rema Hanna, and Stephen P Ryan, "Incentives Work: Getting Teachers to Come to School," *American Economic Review*, June 2012, *102* (4), 1241–1278.

Ensminger, Jean and Joseph Patrick Henrich, eds, *Experimenting with social norms: fairness and punishment in cross-cultural perspective*, New York: Russell Sage Foundation, 2014.

Fehr, Ernst and Simon Gächter, "Cooperation and punishment in public goods experiments," *The American Economic Review*, 2000, *90* (4), 980–994.

Fox, J.A., "Social accountability: What does the evidence really say?," *World Development*, 2015, *72*, 346361.

Fox, Jonathan A., "Social Accountability: What Does the Evidence Really Say?," *World Development*, August 2015, *72*, 346–361.

Gerber, Anke and Philipp C. Wichardt, "Providing public goods in the absence of strong institutions," *Journal of Public Economics*, April 2009, *93* (3-4), 429–439.

Glewwe, P. and K. Muralidharan, "Improving Education Outcomes in Developing Countries," in "Handbook of the Economics of Education," Vol. 5, Elsevier, 2016, pp. 653–743.

Gove, Amber and Anna Wetterberg, "The Early Grade Reading Assessment: Applications and Interventions to Improve Basic Literacy," Technical Report, RTI Press, Research Triangle Park, NC October 2011.

Grandvoinnet, H.M., G. Aslam, and S. Raha, *Opening the black box : the contextual drivers of social accountability*, World Bank, 2015.

Han, The Anh, "Emergence of social punishment and cooperation through prior commitments," in "Proceedings of the conference of the American Association of Artificial Intelligence" Phoenix, AZ 2016, pp. 2494–2500.

Hasan, Amer, Marilou Hyson, and Mae Chu-Chang, eds, *Early childhood education and development in poor villages of indonesia: strong foundations, later success* Directions in development : human development, Washington, D.C: World Bank, 2013.

Heß, Simon, "Randomization Inference with Stata: A Guide and Software," *The Stata Journal: Promoting communications on statistics and Stata*, September 2017, *17* (3), 630–651.

Heyward, Mark, Aos Santosa Hadiwijaya, Mahargianto, and Edy Priyono, "Reforming teacher deployment in Indonesia," *Journal of Development Effectiveness*, April 2017, *9* (2), 245–262.

Holmstrom, B. and P. Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, January 1991, *7* (special), 24–52.

Joshi, Anuradha, "Do They Work? Assessing the Impact of Transparency and Accountability Initiatives in Service Delivery," *Development Policy Review*, July 2013, *31*, s29–s48.

JPAL, "Conducting Cost-Effectiveness Analysis (CEA)," 2019.

Kesuma, Ratna, Anuja Utz, Petra W. Bodrogini, and Ruwiyati Purwana, *Efficient Deployment of Teachers*, World Bank, Washington, DC, August 2018.

Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken, "Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors," *The Quarterly Journal of Economics*, February 2016, *131* (1), 219–271.

Kosec, K. and L. Wantchekon, "Can information improve rural governance and service delivery?," *World Development*, 2020, *125*, 104376.

Lieberman, E.S., D.N. Posner, and L.L. Tsai, "Does information lead to more active citizenship? Evidence from an education intervention in rural Kenya," *World Development*, 2014, *60*, 6983.

Macleod, W. Bentley, "Optimal Contracting with Subjective Evaluation," *American Economic Review*, February 2003, *93* (1), 216–240.

Mansuri, Ghazala and Vijayendra Rao, *Localizing Development: Does Participation Work?*, The World Bank, November 2012.

Marchegiani, Lucia, Tommaso Reggiani, and Matteo Rizzolli, "Loss averse agents and lenient supervisors in performance appraisal," *Journal of Economic Behavior & Organization*, November 2016, *131*, 183–197.

Molina, E., L. Carella, A. Pacheco, G. Cruces, and L. Gasparini, "Community monitoring interventions to curb corruption and increase access and quality in service delivery: a systematic review," *Journal of Development Effectiveness*, 2017, *9* (4), 462–499.

Olken, Benjamin A., Junko Onishi, and Susan Wong, "Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia," *American Economic Journal: Applied Economics*, October 2014, *6* (4), 1–34.

Perez-Alvarez, Marcello, Jan Priebe, and Dewi Susanti, *Teacher Accountability and Pay-for-Performance Schemes in (Semi-) Urban Indonesia*, World Bank, Washington, DC, January 2020.

Platas, L., L. Ketterlin-Gellar, A. Brombacher, and Y. Sitabkhan, "Early Grade Mathematics Assessment (EGMA) Toolkit," Technical Report, RTI International 2014.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha, "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia," *American Economic Journal: Applied Economics*, April 2014, *6* (2), 105–126.

Pritchett, Lant, "Creating Education Systems Coherent for Learning Outcomes: Making the Transition from Schooling to Learning," Technical Report RISE-WP-15/005 December 2015.

Raffler, P., D.N. Posner, and D. Parkerson, "The weakness of bottom-up accountability: Experimental evidence from the Ugandan health sector," *mimeo*, 2019.

Ringold, Dena, Alaka Holla, Margaret Koziol, and Santosh Srinivasan, *Citizens and Service Delivery: Assessing the Use of Social Accountability Approaches in the Human Development Sectors* Directions in Development, Washington, DC: World Bank, 2012.

Rockoff, J.E. and S. Speroni, "Subjective and objective evaluations of teacher effectiveness," *American Economic Review Papers & Proceedings*, 2011, *100* (2), 261–266.

SMERU, "Teacher Absenteeism and Remote Area Allowance: Baseline Survey," Technical Report 2010.

UNICEF, ""We Like Being Taught" – A Study on Teacher Absenteeism in Papua and West Papua," Technical Report 2012.

Usman, S., Akhmadi, and Daniel Suryadarma, "When Teachers are Absent: Where Do They Go and What is the Impact on Students?," Technical Report, SMERU 2004.

Uwezo, "Are Our Children Learning? Annual Learning Assessment Report," Technical Report, Twaweza East Africa, Nairobi 2012.

World Bank, *World Development Report 2004: Making Services Work for Poor People*, The World Bank, September 2003.

_ , *World Development Report 2015: Mind, Society, and Behavior*, The World Bank, December 2014.

World Bank, "Assessing the Role of the School Operational Grant Program (BOS) in Improving Education Outcomes in Indonesia," Technical Report AUS4133, World Bank, Washington DC 2015.

_ , "Teacher certification and beyond: An empirical evaluation of the teacher certification program and education quality improvements in Indonesia," Technical Report 94019-ID, World Bank, Washington DC 2016.

_ , *World Development Report 2018: Learning to realize education's promise*, Washington DC: World Bank, 2018.

_ , "Community Participation and Teacher Accountability: Improving Learning Outcomes in Remote Areas of Indonesia," Technical Report, World Bank, Washington DC 2020.

Tables

Table 2: Baseline Summary Statistics

	Mean	Standard deviation	N
	(1)	(2)	(3)
	Panel A. Student Characteristics		
Male	0.53	0.50	25701
Age	10.68	2.01	25457
Share having mothers with:			
...no education	0.09	0.29	24252
...primary education	0.73	0.44	24252
...more than primary education	0.18	0.38	24252
Share having fathers with:			
...no education	0.07	0.26	24479
...primary education	0.69	0.46	24479
...more than primary education	0.23	0.42	24479
Baseline learning assessment score:			
Indonesian	37.46	20.75	26580
Mathematics	37.65	21.64	26580
	Panel B. Teacher characteristics		
Age	37.38	10.69	2297
Male	0.52	0.50	2297
Married	0.85	0.35	2297
Bachelor's degree or higher	0.55	0.50	2297
Share of teachers observed to be ...:			
...present	0.80	0.40	2212
...working	0.74	0.44	2212
...teaching	0.73	0.44	1688
(Self-reported) hours spent monthly:			
...preparing lessons	17.51	16.65	2021
...teaching curricular materials	64.84	22.05	2021
...assessing student work	12.85	11.61	2021
...teaching extra-curricular materials	4.22	6.04	2021
...on off-own-school employment	17.12	32.64	2048
	Panel C. Parent characteristics		
Mother is the respondent	0.45	0.50	4427
Education expenditures in last academic year	365,624	233,063	4427
Hours of children's accompanied learning (last week)	2.47	2.94	4427
Meetings with principal or teacher in academic year	1.40	4.66	4427
	Panel D. School characteristics		
Number of teachers	8.52	2.29	270
Number of civil servant teachers	3.97	1.66	270
Number of students	106.63	45.62	270
Private school	0.08	0.27	270

Table 3: Impact on Student Learning Outcomes

	Indonesian		Mathematics		Average Score		Grade Repetition	
	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SAM	0.094	0.014	0.073	0.042	0.084	0.028	0.010	-0.000
	(0.037)**	(0.027)	(0.040)*	(0.044)	(0.036)**	(0.032)	(0.010)	(0.008)
SAM+Cam	0.190	0.096	0.202	0.176	0.198	0.133	0.004	0.014
	(0.036)***	(0.028)***	(0.041)***	(0.046)***	(0.036)***	(0.034)***	(0.010)	(0.008)
SAM+Score	0.122		0.094		0.110		0.009	
	(0.034)***		(0.038)**		(0.033)***		(0.010)	
Control group mean							0.08	0.04
Control group raw-score mean	47.13	38.12	47.03	44.04	47.08	41.08		
Test of equality (P-val)								
SAM v. SAM+Cam	0.013	0.003	0.003	0.003	0.003	0.002	0.565	0.117
SAM+Cam v. SAM+Score	0.061		0.013		0.018		0.614	
SAM v. SAM+Score	0.447		0.602		0.474		0.963	
Randomization Inference								
(P-value, N = 1000)								
SAM	0.014	0.663	0.071	0.397	0.026	0.445	0.393	0.990
SAM+Cam	0.000	0.002	0.000	0.001	0.000	0.000	0.723	0.156
SAM+Score	0.001		0.016		0.003		0.424	
R2	0.330	0.109	0.302	0.184	0.390	0.192	0.139	0.073
Observations	31022	15611	31022	15611	31022	15611	24719	13257
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standardized scores are grade adjusted. Control variables include sex, age dummies, both parents' education, baseline outcome, dummy variables for missing controls (one for each control variable), school-level mean scores, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. The randomization inference tests the sharp null hypothesis of no effect for each individual treatment (holding other treatment assignments constant). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 4: Differential Impacts on Student Learning

	Male		Years with TSA teachers		Above-median student				Above-median school	
					in school		across all schools			
	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
SAM	0.060	0.026	0.078	0.021	0.070	-0.001	0.060	-0.006	0.096	0.084
	(0.039)	(0.038)	(0.052)	(0.065)	(0.040)*	(0.038)	(0.040)	(0.039)	(0.050)*	(0.046)*
SAM+Cam	0.180	0.120	0.176	0.128	0.130	0.104	0.130	0.107	0.175	0.190
	(0.040)***	(0.038)***	(0.055)***	(0.063)**	(0.041)***	(0.040)***	(0.044)***	(0.044)**	(0.055)***	(0.054)***
SAM+Score	0.091		0.098		0.066		0.079		0.135	
	(0.038)**		(0.051)*		(0.038)*		(0.042)*		(0.050)***	
Covariate: [...]	-0.152	-0.212	0.000	-0.008	0.109	0.108	0.155	0.095	0.031	0.197
	(0.021)***	(0.027)***	(0.024)	(0.018)	(0.027)***	(0.031)***	(0.033)***	(0.032)***	(0.070)	(0.068)***
... × SAM	0.034	0.011	0.000	0.006	0.012	0.050	0.031	0.055	-0.037	-0.117
	(0.029)	(0.037)	(0.035)	(0.033)	(0.035)	(0.035)	(0.047)	(0.044)	(0.072)	(0.061)*
... × SAM+Cam	0.012	0.020	0.009	0.002	0.071	0.056	0.064	0.045	0.018	-0.127
	(0.029)	(0.037)	(0.038)	(0.030)	(0.034)**	(0.039)	(0.044)	(0.042)	(0.079)	(0.073)*
... × SAM+Score	0.018		0.002		0.023		-0.008		-0.071	
	(0.030)		(0.032)		(0.035)		(0.043)		(0.072)	
Observations	31022	15297	31022	15297	24700	13655	24700	13655	31022	15297
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Control variables include sex, age dummies, both parents' education, baseline outcome, dummy variables for missing controls (one for each control variable), school-level mean scores, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 5: Student Attrition

	2018	2019	2018	2019
	(1)	(2)	(3)	(4)
SAM	-0.008	-0.004	-0.010	-0.002
	(0.006)	(0.007)	(0.008)	(0.008)
... × Above-median student			0.003	-0.002
			(0.008)	(0.012)
SAM+Cam	-0.008	-0.008	-0.012	-0.008
	(0.006)	(0.007)	(0.008)	(0.008)
... × Above-median student			0.006	0.000
			(0.007)	(0.011)
SAM+Score	-0.013		-0.017	
	(0.005)**		(0.007)**	
... × Above-median student			0.006	
			(0.007)	
Control group mean	0.08	0.07	0.08	0.07
Test of equality (P-val)				
SAM v. SAM+Cam	0.997	0.443	0.836	0.491
SAM+Cam v. SAM+Score	0.164		0.322	
SAM v. SAM+Score	0.208		0.296	
R2	0.495	0.090	0.493	0.081
Observations	26613	19044	26613	19044
Individual controls	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes

Notes: Control variables include sex, age dummies, both parents' education, dummy variables for missing controls (one for each control variable), and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 6: Payoff Matrix in the Parent-Teacher Interactions

		PARENT EFFORT	
		LOW	HIGH
TEACHER EFFORT	LOW	$(T2, P2)$	$(T4, P1)$
	HIGH	$(T1, P3)$	$(T3, P4)$

(a) No interventions

		PARENT EFFORT	
		LOW	HIGH
TEACHER EFFORT	LOW	$(T2 - d_t, P2)$	$(T4 - d_t, P1)$
	HIGH	$(T1, P3)$	$(T3, P4)$

(b) Perfectly monitored commitment contract

		PARENT EFFORT		ASSESSMENT REGIME
		LOW	HIGH	
TEACHER EFFORT	LOW	$(T2, (1-\pi_o).P2 + \pi_o.P3)$	$(T4, (1-\pi_o).P1 + \pi_o.P4)$	LENIENT
	HIGH	$(T1, (1-\pi_u).P3 + \pi_u.P2)$	$(T3, (1-\pi_u).P4 + \pi_u.P1)$	
TEACHER EFFORT	LOW	$(T2 - (1-\pi_o).d_t,$ $(1-\pi_o).P2 + \pi_o.P3)$	$(T4 - (1-\pi_o).d_t,$ $(1-\pi_o).P1 + \pi_o.P4)$	STRICT
	HIGH	$(T1 - \pi_u.d_t,$ $(1-\pi_u).P3 + \pi_u.(P2 - R))$	$(T3 - \pi_u.d_t,$ $(1-\pi_u).P4 + \pi_u.(P1 - R))$	

(c) Imprecisely monitored commitment contract with teacher retaliation

Table 7: Impact on Teacher Presence and Activities

	Teacher is [...]					
	present		working		teaching	
	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Overall impact					
SAM	0.007	-0.025	0.020	0.010	0.047	0.036
	(0.024)	(0.028)	(0.029)	(0.034)	(0.035)	(0.039)
SAM+Cam	0.024	-0.015	0.031	-0.003	0.016	0.020
	(0.026)	(0.024)	(0.030)	(0.031)	(0.039)	(0.038)
SAM+Score	-0.063		-0.076		-0.006	
	(0.027)**		(0.033)**		(0.036)	
	Panel B. Impact by TSA status					
SAM	-0.010	-0.033	-0.009	0.016	0.011	0.017
	(0.038)	(0.041)	(0.041)	(0.048)	(0.045)	(0.053)
SAM+Cam	-0.019	-0.015	-0.059	0.022	-0.088	0.034
	(0.046)	(0.037)	(0.047)	(0.045)	(0.051)*	(0.051)
SAM+Score	-0.132		-0.161		-0.113	
	(0.044)***		(0.052)***		(0.051)**	
TSA-receiving teacher	-0.041	0.004	-0.079	0.029	-0.122	0.042
	(0.046)	(0.042)	(0.046)*	(0.047)	(0.050)**	(0.051)
... × SAM	0.029	0.014	0.048	-0.011	0.063	0.030
	(0.055)	(0.054)	(0.055)	(0.059)	(0.059)	(0.060)
... × SAM+Cam	0.069	-0.000	0.143	-0.045	0.171	-0.026
	(0.060)	(0.054)	(0.061)**	(0.058)	(0.064)***	(0.060)
... × SAM+Score	0.115		0.141		0.180	
	(0.059)*		(0.061)**		(0.065)***	
Control group mean	0.84	0.84	0.80	0.79	0.76	0.76
Observations	1711	1234	1711	1234	1531	1148
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Includes the sample of class teachers. Individual controls include sex, age, education, and the baseline outcome. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 8: Impact on Teachers' Time Allocation for School-Related Activities

	Total hours				Hours of learning-enhancing activities†			
	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SAM	-0.530	0.447	0.027	1.314	1.218	0.302	1.605	0.694
	(0.817)	(0.828)	(1.375)	(1.401)	(0.449)***	(0.493)	(0.756)**	(0.837)
SAM+Cam	0.190	0.543	0.719	1.313	1.305	0.073	1.701	0.148
	(0.812)	(0.827)	(1.421)	(1.468)	(0.446)***	(0.491)	(0.780)**	(0.875)
SAM+Score	-0.116		-0.962		0.553		0.262	
	(0.822)		(1.408)		(0.451)		(0.774)	
TSA-receiving teacher			2.492	2.549			1.696	0.997
			(1.303)*	(1.323)*			(0.718)**	(0.792)
... × SAM			-0.907	-1.422			-0.631	-0.628
			(1.688)	(1.702)			(0.929)	(1.019)
... × SAM+Cam			-0.859	-1.276			-0.638	-0.177
			(1.731)	(1.767)			(0.952)	(1.057)
... × SAM+Score			1.144				0.359	
			(1.726)				(0.950)	
Control group mean	26.31	25.35	26.31	25.35	15.14	16.33	15.14	16.33
Observations	1418	950	1418	950	1418	950	1418	950
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Includes the sample of class teachers. †The outcome variable for columns 5–8 is the total weekly hours of teacher activities that are positively correlated with learning outcomes at baseline. The underlying correlation is estimated using a specification that is determined through a post double-selection lasso process. Variables that are positively correlated with learning (at 10% significance level) at baseline are: (i) in-school teaching; (ii) out-of-school additional intra-curricular lessons; (c) out-of-school scientific publications; and (d) out-of-school innovative activities (develop teaching tools etc). Individual controls include sex, age, education, and the baseline outcome. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels.

Table 9: Parent Investments

	Education expenditure		Child's employment†				Hours of accompanied learning		Number of meetings with teachers‡	
			Child is employed		Hours per week					
	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
SAM	13816.3	15992.0	-0.0806	-0.0176	-0.786	-0.318	0.242	0.258	1.043	0.628
	(13376.4)	(12622.4)	(0.0183)***	(0.0184)	(0.224)***	(0.242)	(0.194)	(0.174)	(0.213)***	(0.304)**
SAM+Cam	27666.1	28593.7	-0.0444	0.0210	-0.294	-0.359	0.292	0.337	1.218	0.937
	(13998.3)**	(12096.0)**	(0.0185)**	(0.0203)	(0.205)	(0.244)	(0.193)	(0.186)*	(0.222)***	(0.286)***
SAM+Score	8808.3		-0.0370		-0.431		0.263		1.067	
	(14221.3)		(0.0186)**		(0.191)**		(0.196)		(0.244)***	
Control group mean	323867.6	347148.3	0.403	0.353	1.477	1.703	2.458	2.135	1.199	1.415
Test of equality (P-val)										
SAM v. SAM+Cam	0.302	0.335	0.049	0.041	0.024	0.857	0.773	0.660	0.440	0.150
SAM+Cam v. SAM+Score	0.182		0.689		0.492		0.871		0.556	
SAM v. SAM+Score	0.713		0.018		0.097		0.905		0.921	
R2	0.731	0.752	0.235	0.278	0.108	0.112	0.427	0.370	0.230	0.218
Observations	5401	4166	5401	4185	5397	4128	5394	4160	5401	3563
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: †Child employment is defined as working for pay or is a family labor. ‡The total number of meetings is calculated as the maxima of the reported number of meetings between teacher and parents on various topics. Outcomes were constructed from the parent survey. Individual control variables include whether the respondent is the child's mother, as well as child characteristics (sex, age dummies, both parents' education), and the baseline outcome. School-level control variables include dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Controls also include dummy variables for missing controls (one for each control variable). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 10: Heterogeneous Impacts on Learning and Teacher Presence by Punishment Norms

	Teacher Presence				Learning Outcomes	
	TSA		Non-TSA			
	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)
SAM	0.003	-0.049	-0.055	-0.137	0.040	-0.026
	(0.046)	(0.086)	(0.064)	(0.063)**	(0.068)	(0.069)
SAM+Cam	-0.039	-0.066	-0.078	0.006	0.088	-0.008
	(0.052)	(0.086)	(0.059)	(0.060)	(0.063)	(0.067)
SAM+Score	-0.009		-0.198		0.005	
	(0.057)		(0.066)***		(0.064)	
Above-Median Punishment	-0.109	-0.036	0.048	0.024	-0.171	-0.087
	(0.063)*	(0.095)	(0.067)	(0.070)	(0.060)***	(0.065)
... × SAM	0.146	0.041	-0.023	0.085	0.078	0.010
	(0.075)*	(0.127)	(0.090)	(0.091)	(0.093)	(0.099)
... × SAM+Cam	0.238	0.004	-0.005	-0.115	0.253	0.204
	(0.091)***	(0.124)	(0.093)	(0.105)	(0.093)***	(0.095)**
... × SAM+Score	0.058		-0.096		0.179	
	(0.082)		(0.097)		(0.088)**	
Observations	714	467	469	375	22522	11229
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Treatment variables are interacted with a punishment norm variable based on a lab-in-the-field behavioral games in 182 out of 270 schools. The variable captures whether the average parent participants in the school imposed an above-median penalties to group members who had a below-average contribution in the public goods game. Teacher respondents include the sample of class teachers. Individual controls include sex, age, education, and the baseline outcome. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Controls also include dummy variables for missing controls (one for each control variable). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 11: User Committee Reports of Pressure from School

	Intimidated		Pressure to Increase Score		Threats for Low Score	
	(1)	(2)	(3)	(4)	(5)	(6)
SAM+Cam	0.021	-0.045	-0.005	-0.115	0.071	0.095
	(0.040)	(0.073)	(0.056)	(0.093)	(0.044)	(0.073)
SAM+Score	0.066	-0.028	0.119	0.128	0.165	0.131
	(0.041)	(0.068)	(0.056)**	(0.087)	(0.044)***	(0.069)*
Absent teacher is hard to replace†		-0.083		-0.103		0.007
		(0.070)		(0.090)		(0.071)
... × SAM+Cam		0.115		0.112		-0.089
		(0.103)		(0.132)		(0.104)
... × SAM+Score		0.164		-0.024		0.056
		(0.095)*		(0.121)		(0.096)
Constant	0.107	0.161	0.035	0.152	-0.038	-0.023
	(0.083)	(0.100)	(0.114)	(0.127)	(0.090)	(0.100)
Control group mean	0.030		0.075		0.000	
Observations	201	195	201	195	201	195
School-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Column (1) from the question "Did UC members feel intimidated to discuss evaluation results openly?"; column (2): "Did you feel any pressure from the school to give scores that are better than the teacher deserved; column (3): "Did any UC member ever receive threats from a teacher/principal to not give a low score?" From 203 treated schools, 1 school was missing because user committee members were unavailable for an interview after multiple visits, and 1 school was dropped for being a singleton within the strata. †Based on school principal assessment at baseline in 197 out of 203 treated schools. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. */**/*** denotes 10/5/1 percent significance levels

Table 12: External Supervision of the School Management

	Number of meetings with education officials		Number of supervisor visits	
	2018	2019	2018	2019
	(1)	(2)	(3)	(4)
SAM	0.361	-1.226	0.315	-0.014
	(0.487)	(0.488)**	(0.334)	(0.373)
SAM+Cam	1.035	-0.389	0.773	-0.309
	(0.489)**	(0.484)	(0.336)**	(0.370)
SAM+Score	-0.068		0.487	
	(0.491)		(0.337)	
Control group mean	2.24	2.54	1.42	2.21
Test of equality (P-val)				
SAM v. SAM+Cam	0.156	0.076	0.161	0.415
SAM+Cam v. SAM+Score	0.023		0.387	
SAM v. SAM+Score	0.367		0.598	
Observations	270	203	270	203
Controls	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes

Notes: All outcomes are recorded with respect to the current academic year based on the school principal survey. Columns 3–4 report the impacts on the number of monitoring visits by school inspectors and/or for private schools, a representative of the private foundation. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. We also include a dummy variable of whether the respondent is the school principal. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 13: Teacher Management by School Principal

	Share of teachers receiving [...] evaluation				Frequency of evaluation		In-class teaching observation	
	any		routine					
	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SAM	0.059	0.095	0.121	0.093	1.924	1.480	0.086	0.065
	(0.042)	(0.046)**	(0.050)**	(0.055)*	(0.437)***	(0.517)***	(0.029)***	(0.029)**
SAM+Cam	0.097	0.127	0.149	0.125	2.506	1.883	0.091	0.069
	(0.040)**	(0.043)***	(0.049)***	(0.052)**	(0.439)***	(0.517)***	(0.028)***	(0.028)**
SAM+Score	0.089		0.146		2.306		0.099	
	(0.040)**		(0.049)***		(0.425)***		(0.029)***	
Control group mean	0.73	0.71	0.42	0.45	2.79	3.44	0.67	0.70
Test of equality (P-val)								
SAM v. SAM+Cam	0.313	0.422	0.547	0.519	0.205	0.420	0.874	0.880
SAM+Cam v. SAM+Score	0.817		0.949		0.644		0.770	
SAM v. SAM+Score	0.410		0.576		0.384		0.651	
Observations	270	203	270	203	270	203	2021	1430
School-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	-	-	-	-	-	-	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Outcome variables come from the responses of individual class teachers to the teacher survey. The outcomes for columns 1–6 are aggregated at the school level, while those for columns 7–8 are at the individual level. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Individual teacher controls include age, sex, and whether the teacher is married. Standard errors are robust for school-level outcomes (columns 1–6) and clustered at the school level for individual-level outcomes (columns 7–8). */**/*** denotes 10/5/1 percent significance levels

Table 14: Teacher Satisfaction

	Appreciation from [...]				Salary		Current job in this school	
	district		village					
	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A. Overall impact							
SAM	0.125	0.459	0.243	0.349	0.231	0.280	0.068	-0.018
	(0.114)	(0.119)***	(0.098)**	(0.108)***	(0.116)**	(0.122)**	(0.041)*	(0.045)
SAM+Cam	0.360	0.418	0.321	0.515	0.436	0.500	0.130	-0.020
	(0.113)***	(0.118)***	(0.098)***	(0.107)***	(0.115)***	(0.120)***	(0.040)***	(0.045)
SAM+Score	0.474		0.370		0.682		0.098	
	(0.114)***		(0.098)***		(0.116)***		(0.041)**	
	Panel B. Impact by TSA status							
SAM	0.002	0.420	0.189	0.488	0.464	0.548	0.087	-0.019
	(0.171)	(0.177)**	(0.149)	(0.161)***	(0.173)***	(0.179)***	(0.061)	(0.068)
SAM+Cam	0.398	0.273	0.239	0.557	0.686	0.687	0.236	0.002
	(0.173)**	(0.178)	(0.150)	(0.161)***	(0.174)***	(0.180)***	(0.062)***	(0.068)
SAM+Score	0.506		0.166		0.824		0.150	
	(0.170)***		(0.147)		(0.171)***		(0.061)**	
TSA-receiving teacher	0.402	0.521	-0.049	0.414	1.092	1.104	0.145	0.099
	(0.179)**	(0.191)***	(0.155)	(0.173)**	(0.180)***	(0.193)***	(0.064)**	(0.073)
... × SAM	0.202	0.063	0.097	-0.248	-0.440	-0.493	-0.037	0.001
	(0.226)	(0.235)	(0.196)	(0.213)	(0.228)*	(0.237)**	(0.081)	(0.089)
... × SAM+Cam	-0.089	0.229	0.141	-0.088	-0.484	-0.367	-0.187	-0.042
	(0.228)	(0.236)	(0.197)	(0.214)	(0.229)**	(0.238)	(0.082)**	(0.090)
... × SAM+Score	-0.073		0.361		-0.285		-0.098	
	(0.226)		(0.196)*		(0.228)		(0.081)	
Control group mean	4.35	4.50	4.97	4.94	3.96	4.20	3.00	3.05
Observations	1773	1254	1773	1254	1773	1254	1773	1255
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Column 1–6 outcomes are measured using a 7-point scale while column 7–8 outcomes are measured on a 4-point scale. Individual controls include sex, age, education, and the baseline outcome. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Controls also include dummy variables for missing controls (one for each control variable). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 15: Parent Satisfactions and Aspirations

	Considers school to be good/very good		Satisfaction with results in				Prefers child pursues university over working	
			Indonesian		Mathematics			
	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SAM	0.050	0.051	-0.025	0.050	-0.046	0.094	0.094	0.045
	(0.019)**	(0.017)***	(0.074)	(0.068)	(0.078)	(0.068)	(0.027)***	(0.033)
SAM+Cam	0.054	0.053	0.029	0.335	-0.015	0.351	0.090	0.058
	(0.019)***	(0.017)***	(0.073)	(0.071)***	(0.076)	(0.070)***	(0.028)***	(0.031)*
SAM+Score	0.053		0.006		0.024		0.077	
	(0.019)***		(0.080)		(0.085)		(0.026)***	
Control group mean	0.911	0.901	4.747	4.924	4.579	4.730	3.510	3.500
Test of equality (P-val)								
SAM v. SAM+Cam	0.677	0.844	0.460	0.000	0.699	0.000	0.876	0.685
SAM+Cam v. SAM+Score	0.904		0.779		0.663		0.624	
SAM v. SAM+Score	0.766		0.685		0.431		0.504	
R2	0.999	0.999	0.992	0.996	0.989	0.995	0.977	0.977
Observations	5310	4164	5401	4165	5401	4165	5401	4165
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Columns 3–6 outcomes were measured on a 7-point scale, while columns 7-8 outcomes were measured on a 4-point scale. Student-level control variables include sex, age dummies, both parents' education, whether the respondent is the child's mother, and the baseline outcome. School-level control variables include dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Controls also include dummy variables for missing controls (one for each control variable). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

9 Figures

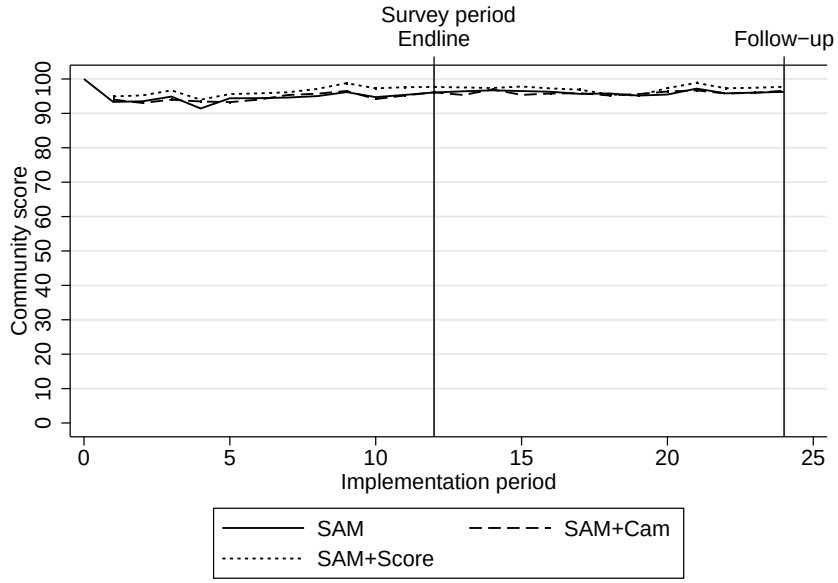
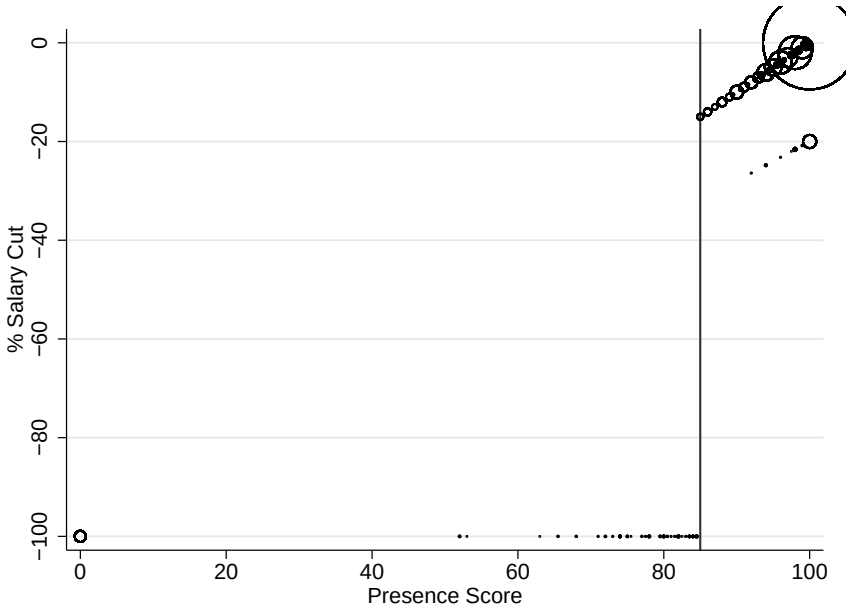


Figure 2: Average Scorecard Ratings Across Treatments



Notes: The salary cut is calculated as a percentage of the special allowance. The gray line indicates the cutoff score of 85. Markers are weighted by the number of observations in that point. The graph includes observations between August 2017 and March 2019, excluding December 2017 and 2018 when salaries were not cut.

Figure 3: Compliance of the 85 Percent Rule in SAM+Cam

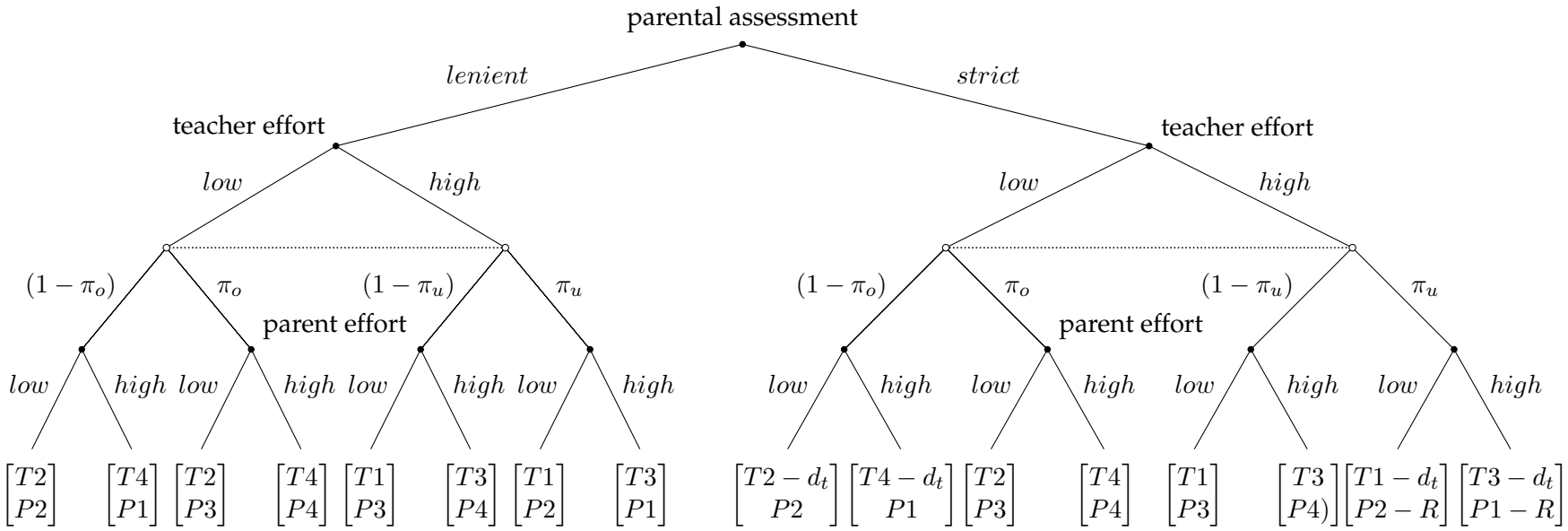


Figure 4: The Strictness of Parental Assessment, Imprecise Measurement, and Teacher Retaliation

Appendix

A The Stratification of Treatment Assignments

We use a simulation to construct groups of similar schools to form a stratum. We begin by constructing a measure of within-group dissimilarity for a particular random grouping of schools. For this, we first standardized all variables by subtracting the mean and dividing by the standard deviation. Then, we define a within-group absolute distance as

$$D(g) = \sum_{k} \sum_{i} \sum_{j,j<i} |x_{gkj} - x_{gki}|$$

where k indexes the underlying matching variable (e.g., the mobile phone signal), i and j denote the village id within the group g. Finally, we sum up the within-group absolute distances across all groups for this random sorting of villages to construct the within-group dissimilarity measure for a particular random grouping.

To determine the groups of schools with the smallest within-group dissimilarity, for each district, we randomly sorted villages, sequentially allocated them to groups, and calculated their total within-group dissimilarity. We then take another random draw and repeat this procedure. If the total distance in the new draw is smaller than any in the previous draws, we retain the grouping. We repeated the process 1,000 times. Because the procedure is implemented separately for each district, a group is always defined within a district.

B Model Appendix

In this section, we show that the payoff matrix as in Table 6a can be derived from a production function of learning that is linear in parent and teacher efforts with complementarities in their efforts. We further show that if the utility cost of effort is also linear, the payoff matrix presented in Table 6a implies that utility cost of effort for parents must be lower that for teachers.

Assume a linear production of learning:

$$L = \alpha_0 + \alpha_t.E_t + \alpha_p.E_p + \alpha_{tp}E_tE_p \tag{B.1}$$

where L is student learning, E is the effort put into student learning with subscripts t and p that respectively index teachers and parents. Let $E = 0$ denotes low effort and $E = 1$ high effort. All α's are assumed to be positive.

Parents and teachers derive utility from student learning and there is a utility cost of effort c. Utility is a linear function of learning and the utility cost of effort, to wit:

$$\text{Teachers:} \quad U_t(E_t, E_p) = L - c_t.E_t \tag{B.2}$$

$$\text{Parents:} \quad U_p(E_t, E_p) = L - c_p.E_p \tag{B.3}$$

Table 6a indicates the following ordering of the payoffs:

$$\text{Teachers:} \quad U_t(1,0) < U_t(0,0) < U_t(1,1) < U_t(0,1) \tag{B.4}$$

$$\text{Parents:} \quad U_p(0,1) < U_p(0,0) < U_p(1,0) < U_p(1,1) \tag{B.5}$$

Substituting B.1–B.3 into B.4 and B.5, these conditions together require that

$$\text{Teachers:} \quad c_t > \alpha_t + \alpha_p + \alpha_{tp} \tag{B.6}$$

$$\text{Parents:} \quad \alpha_p < c_p < \alpha_p + \alpha_{tp} \tag{B.7}$$

which implies that $c_t > c_p$, namely that the utility cost of effort is lower for parents than for teachers.

C Disentangling the Two-Year Impacts of SAM and SAM+Cam

In this section, we describe our empirical approach to disentangle the two-year impacts of our treatments into the knock-on impacts from the first-year implementation and additional impacts in the second year. Suppose student learning at time t can be described as:

$$y_t = \alpha_t + \beta_t T + \delta_t y_{t-1} + \varepsilon_t \tag{C.8}$$

where $y =$ student learning; $\beta_t =$ the (new) treatment effect of Treatment T at time t; and $\delta_t =$ the lagged learning coefficient. Learning at $t = 2$ can therefore be described as:

$$y_2 = \alpha_2 + \beta_2 T + \delta_2 y_1 + \varepsilon_2 \tag{C.9}$$

Replacing y_1 in Equation C.9 with an expression for y_1 based on Equation C.8, we obtain:

$$\begin{aligned} y_2 &= \alpha_2 + \beta_2 T + \delta_2(\alpha_1 + \beta_1 T + \delta_1 y_0 + \varepsilon_1) + \varepsilon_2 \\ &= \alpha_2 + (\underbrace{\beta_2 + \delta_2 \beta_1}_{\theta_2})T + \delta_2(\alpha_1 + \delta_1 y_0 + \varepsilon_1) + \varepsilon_2 \end{aligned} \tag{C.10}$$

where $(\beta_2 + \delta_2 \beta_2) = \theta_2 =$ the reduced form two-year impact estimates. In the absence of new second year impact of the treatment, $\beta_2 = 0$ and our reduced form estimates would be equal to $\delta_2.\beta_1$.

To test the null hypothesis of $\beta_2 = 0$, we need unbiased estimates of δ_2 and β_1. We obtain δ_2 by estimating Equation C.9 for the control schools. Meanwhile, β_1 (θ_2) is the one-year (two-year) impact estimates for each of the treatments. We estimate δ_2, β_1, and θ_2 in a Seemingly Unrelated Regression (SUR) framework with clustered standard errors.

Our results suggest that SAM+Cam, but not SAM, continued to improve learning above and beyond the knock-on impact from the first-year implementation. In the top panel of Appendix Table C.1, we show our estimates for the δ_2, β_1, and θ_2 of SAM and SAM+Cam for Indonesian, mathematics, and the mean standardized scores. In the middle panel, we use the delta method to construct $(\delta_2 \times \beta_1)$ for SAM and SAM+Cam. We then test for each treatment whether $(\delta_2 \times \beta_1) = \theta_2$ and present the p-value of that test in the bottom panel. Our finding suggests that we cannot reject the null hypothesis of $\beta_2 = 0$ for SAM, but we reject it for SAM+Cam. Moreover, column 3 suggests that almost half of the two-year learning impacts on the mean score can be attributed to new impacts in the second year.

Table C.1: Decomposition of the Two-Year Impacts of SAM and SAM+Cam

	Indonesian	Math	Mean Sore
	(1)	(2)	(3)
Lagged learning at follow-up (δ_2)	0.379	0.494	0.519
	(0.019)***	(0.020)***	(0.016)***
One-year impact (β_1):			
SAM	0.095	0.074	0.085
	(0.036)***	(0.043)*	(0.036)**
SAM+Cam	0.134	0.158	0.146
	(0.036)***	(0.044)***	(0.036)***
Two-year impact (θ_2):			
SAM	0.019	0.049	0.034
	(0.026)	(0.042)	(0.030)
SAM+Cam	0.099	0.182	0.138
	(0.028)***	(0.044)***	(0.033)***
Nonlinear Combinations:			
$\delta_2 \times \beta_1^{SAM}$	0.036	0.037	0.044
	(0.014)***	(0.021)*	(0.019)**
$\delta_2 \times \beta_1^{SAM+Cam}$	0.051	0.078	0.076
	(0.014)***	(0.022)***	(0.019)***
Test of equality (P-val)			
$(\delta_2 \times \beta_1^{SAM})$ v. θ_2^{SAM}	0.392	0.720	0.665
$(\delta_2 \times \beta_1^{SAM+Cam})$ v. $\theta_2^{SAM+Cam}$	0.030	0.002	0.011

Notes: The table reports coefficients from a SUR regression to estimate the coefficient on the lagged learning at follow up among the control schools (δ); the one-year impact of SAM and SAM+Cam; and the two-year impact of SAM and SAM+Cam (θ_2). The sample excludes students from grades 1 and 2 at the time the outcome variable was measured. Included control variables and fixed effects are identical to those in the main regression. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

D Lab-in-the-Field Experiment on Punishment Norms

To construct a measure of punishment norms, we employ a public goods game with punishment lab-in-the-field experiment (Fehr and Gächter, 2000). Budgetary constraints meant that we could only implement the experiment in 182 out of 270 schools. Furthermore, the baseline survey (and hence, the experiment) were conducted prior to the random assignment of the treatment arms. We therefore had to randomly selected the subset of schools that would participate in the lab-in-the-field experiment prior to the treatment assignment. As the result, we did not have perfect balance of the distribution of the included schools across the treatment arms: 42, 48, 45, and 47 participating schools were part of the Control, SAM, SAM+Cam, and SAM+Score respectively.

In each school, we invited a total of between 16 and 20 parents and teachers to participate in a set of public goods game. All sessions comprise three stages, with three rounds in each stage. Within each stage, participants played with the same set of individuals but groups are reshuffled at the beginning of each stage. In the first stage, participants anonymously play a standard public goods games where they contribute to a group account. All contributions are doubled and redistributed to all members. In the second stage, participants are informed of the teacher-parent composition of their groups and played the same public goods game.

We use data collected in the Stage 3 where we added a punishment component to the Stage 2 game, to construct our measure of school-level punishment norms. As in Stage 2, participants in Stage 3 know the teacher-parent composition of their group. In this stage, once participants observed the outcome of the first stage and the contribution of each group member, participants can purchase punishment tokens to penalize any member(s) of their group. Even though participants did not know the real identity of their group members, they were informed of whether a particular member of the group was a teacher or a parent. We also randomly allocated schools to two types of games, to wit, social and monetary punishments.[1]

We define the punishment norm as the willingness to punish below-(session-)average public good contributions along the specification of Fehr and Gächter (2000). To cleanly measure punishment norms without the potential effect of repeated interactions, we estimate our measure based on how participants play in the *first* round of Stage 3. School-level measurement norms are constructed by regressing the following specification:

$$P_{si} = \sum_s \beta_s^- (S_s \times D^-) + \sum_s \beta_s^+ (S_s \times D^+) + \gamma G + \eta_s + \varepsilon$$

where P_{si} is the total punishment received by individual i in school s; S_s is the dummy variables for each school; D^- is absolute value of the negative deviation of i's contribution from the session average contribution; D^- is the positive deviation of i's contribution; G is whether the school plays the social- or monetary-punishment game; and η_s is the school fixed effects. β_s^-, which is the *school-specific* elasticity of punishments with respect to under-contribution (relative to the session mean) is our measure of the school-specific punishment norm.

E Efficiency Analysis

The one-time investment cost for the Project facilitators to conduct KIAT Guru approach was a total of USD 1,026,759, or at USD 5,058 per school and USD 40 per student (Table E.2). This cost covered trainings, salaries and transportation costs of 41 Project facilitators working in 203 schools, with an average of 132 students per school. Over three calendar years of implementation, the cost of training and

[1] In the social-punishment game, punishment tokens sent to others resulted in a sticker that expressed dissatisfaction without any monetary consequence to the receiver. In the monetary-punishment game, punishment tokens reduced the receiver's private payoff.

workshop was USD 2,756,791, of which USD 431,667 was spent for training of Project facilitators. This boils down to an annual cost of training at USD 143,889, or at USD 709 per school, USD 6 per student. The average monthly salary for the facilitator was USD 815, bringing the average monthly total salaries of USD 33,432. The facilitators were employed over 15 months, with a total spending on salary USD 501,483, averaging USD 2,470 per school, USD 19 per student. Each facilitator visited a school with an average of 11 visits, and transportation cost averaging USD 112 per visit. Over the course of 15 months, the total transportation costs reached USD 252,888, with an average of USD 1,246 per school, USD 10 per student. The one-time investment cost of Initial Phase meetings at the village level was USD 633 per school, USD 5 per student. This cost covered seven meetings which resulted in the Service Agreement, the Community Score Card, and the establishment of the User Committee.

Table E.2: One-time Investment Cost to Introduce KIAT Guru

	One-time Investment (203 Schools)	Annual Cost per School	Annual Cost per Student
Training	143,889	709	6
Salary	501,483	2,470	19
Transport	252,888	1,246	10
Initial Meetings	128,499	6333	5
Total	1,026,759	5,058	40

KIAT Guru continued after the endline survey. Below we report the cost of sustaining KIAT Guru. As these costs were incurred after the endline survey, they have not been included in the cost benefit analysis in the main text. The average annual cost to sustain SAM was USD 2,182 per school or USD 17 per student (Table E.3), with additional USD 506 per school or USD 4 per student for Group 2 schools. This cost covers an annual refresher training, monthly meetings, and evaluation meeting. The annual cost of training per school at USD 709, or USD 6 per student. The annual cost to conduct monthly meetings was USD 834 per school, USD 7 per student. The average cost per village to conduct evaluation meetings at the end of every semester was USD 639 per school, USD 5 per student. In these evaluation meetings, User Committee and school providers reviewed the content of Service Agreement and Community Score Card indicators, and reappointed User Committee members. In 2017, 169 of 176 (96%) village governments provided the financial supports, and in 2018, all of them did, with an average of USD 674 (31% of total annual cost to sustain SAM). Given that these costs are only a very small fraction of Village Fund, the cost of maintaining KIAT Guru activities in the villages are completely sustainable. Group 2 schools had a total of USD 86,341 of additional cost to cover the purchase of one smart phone per school and salaries for two personnel to develop and maintain the application. On annual basis, this cost came down to USD 33,874, or at USD 506 per school, USD 4 per student.

Table E.3: Average Annual Cost per School for SAM

	Total Cost (203 Schools)	Annual Cost per School	Annual Cost per Student
Refresher Training	143,927	709	5
Monthly Meetings	169,302	834	7
Evaluation Meetings	129,717	639	5
Total	442,946	2,182	17
Group 2 Additional Cost	86,341	506	4
Total for Group 2	529,287	2,688	21

F Appendix Tables and Figures

I Tables

Table F.4: Balance Tables: Student Characteristics

	Mean (μ) (standard errors)				Differences = $\mu_{[...]} - \mu_{Control}$ (p-value)			Differences between $\mu_{[...]}$ and $\mu_{[...]}$ (p-value)		
	Control	SAM	SAM+ Cam	SAM+ Score	SAM	SAM+ Cam	SAM+ Score	SAM+Cam − SAM	SAM+Cam − SAM	SAM+Score − SAM+Cam
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Male	0.51	0.54	0.52	0.54	0.02**	0.01	0.02**	-0.02*	-0.00	0.02*
	(0.50)	(0.50)	(0.50)	(0.50)	(0.01)	(0.38)	(0.01)	(0.08)	(0.85)	(0.08)
Age	10.76	10.63	10.69	10.65	-0.13	-0.07	-0.11	0.06	0.02	-0.04
	(2.03)	(2.05)	(1.99)	(1.98)	(0.12)	(0.38)	(0.15)	(0.47)	(0.82)	(0.58)
Share having mothers with:										
...no education	0.09	0.07	0.11	0.09	-0.02	0.02	-0.00	0.05	0.02	-0.02
	(0.29)	(0.25)	(0.32)	(0.29)	(0.19)	(0.51)	(0.93)	(0.13)	(0.28)	(0.49)
...primary education	0.75	0.74	0.71	0.73	-0.01	-0.04	-0.02	-0.03	-0.02	0.02
	(0.43)	(0.44)	(0.45)	(0.44)	(0.85)	(0.26)	(0.45)	(0.37)	(0.60)	(0.65)
...more than primary education	0.16	0.19	0.18	0.18	0.03	0.02	0.02	-0.01	-0.01	0.01
	(0.36)	(0.39)	(0.38)	(0.39)	(0.22)	(0.46)	(0.28)	(0.64)	(0.78)	(0.82)
Share having fathers with:										
...no education	0.08	0.05	0.09	0.08	-0.03*	0.02	0.00	0.04*	0.03	-0.02
	(0.26)	(0.22)	(0.29)	(0.27)	(0.09)	(0.48)	(0.96)	(0.08)	(0.13)	(0.53)
...primary education	0.71	0.70	0.67	0.69	-0.02	-0.05	-0.03	-0.03	-0.01	0.02
	(0.45)	(0.46)	(0.47)	(0.46)	(0.59)	(0.13)	(0.30)	(0.34)	(0.67)	(0.55)
...more than primary education	0.21	0.25	0.24	0.24	0.04	0.03	0.03	-0.01	-0.01	-0.00
	(0.41)	(0.43)	(0.43)	(0.43)	(0.13)	(0.26)	(0.24)	(0.72)	(0.59)	(0.91)
Baseline learning assessment scores:										
Indonesian	37.83	36.94	38.46	36.56	-0.89	0.63	-1.27	1.52	-0.38	-1.91
	(21.26)	(20.24)	(20.74)	(20.66)	(0.65)	(0.74)	(0.54)	(0.40)	(0.85)	(0.33)
Mathematics	38.63	37.14	37.93	36.82	-1.48	-0.69	-1.81	0.79	-0.33	-1.12
	(22.45)	(21.32)	(21.16)	(21.50)	(0.49)	(0.72)	(0.43)	(0.70)	(0.89)	(0.61)
Mean score	38.23	37.04	38.20	36.69	-1.19	-0.03	-1.54	1.16	-0.36	-1.51
	(19.65)	(18.72)	(18.69)	(18.98)	(0.56)	(0.99)	(0.47)	(0.54)	(0.87)	(0.46)

Notes: Standard errors clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table F.5: Balance Tables

	Mean (μ) (standard errors)				Differences = $\mu_{[...]} - \mu_{Control}$ (p-value)			Differences between $\mu_{[...]}$ and $\mu_{[...]}$ (p-value)		
	Control	SAM	SAM+ Cam	SAM+ Score	SAM	SAM+ Cam	SAM+ Score	SAM+Cam − SAM	SAM+Cam − SAM	SAM+Score − SAM+Cam
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					Panel A. Teacher Characteristics					
Age	37.34	37.45	37.27	37.43	0.11	-0.07	0.09	-0.18	-0.02	0.16
	(10.96)	(10.69)	(10.67)	(10.48)	(0.87)	(0.92)	(0.89)	(0.80)	(0.98)	(0.82)
Male	0.52	0.53	0.51	0.52	0.01	-0.01	0.00	-0.03	-0.01	0.01
	(0.50)	(0.50)	(0.50)	(0.50)	(0.62)	(0.65)	(0.97)	(0.37)	(0.67)	(0.65)
Married	0.85	0.85	0.86	0.85	-0.00	0.01	-0.01	0.01	-0.00	-0.01
	(0.35)	(0.35)	(0.34)	(0.36)	(0.97)	(0.70)	(0.79)	(0.69)	(0.84)	(0.53)
Bachelor's degree or higher	0.52	0.54	0.56	0.57	0.02	0.05	0.05	0.03	0.03	0.01
	(0.50)	(0.50)	(0.50)	(0.50)	(0.62)	(0.21)	(0.18)	(0.39)	(0.33)	(0.85)
Share of teachers observed to be:										
...present	0.78	0.78	0.81	0.83	-0.00	0.03	0.05	0.03	0.05*	0.02
	(0.41)	(0.41)	(0.39)	(0.37)	(1.00)	(0.32)	(0.12)	(0.26)	(0.09)	(0.55)
...working	0.73	0.73	0.76	0.74	0.00	0.03	0.02	0.03	0.02	-0.02
	(0.45)	(0.45)	(0.43)	(0.44)	(1.00)	(0.36)	(0.69)	(0.31)	(0.67)	(0.66)
...teaching	0.71	0.74	0.75	0.73	0.02	0.03	0.01	0.01	-0.01	-0.02
	(0.45)	(0.44)	(0.43)	(0.45)	(0.55)	(0.36)	(0.78)	(0.73)	(0.82)	(0.60)
					Panel B. Parent Characteristics					
Mother is the respondent (baseline)	0.43	0.45	0.45	0.47	0.01	0.01	0.04	-0.00	0.03	0.03
	(0.50)	(0.50)	(0.50)	(0.50)	(0.66)	(0.68)	(0.18)	(0.97)	(0.37)	(0.33)
Education expenditures in last academic year	302,421	311,188	297,565	325,978	8,767	-4,856	23,558	-13,623	14,791	28,414
	(252,061)	(252,612)	(239,852)	(264,782)	(0.62)	(0.78)	(0.19)	(0.43)	(0.41)	(0.10)
Hours of accompanied learning in previous week	2.46	2.83	2.49	2.76	0.37**	0.03	0.31**	-0.34**	-0.06	0.28
	(2.95)	(3.26)	(2.75)	(3.15)	(0.02)	(0.84)	(0.05)	(0.04)	(0.71)	(0.10)
Meetings with principal or teacher in academic year	1.33	1.47	1.36	1.43	0.14	0.03	0.10	-0.11	-0.04	0.07
	(6.57)	(3.78)	(3.29)	(4.32)	(0.58)	(0.90)	(0.71)	(0.57)	(0.87)	(0.75)
					Panel C. School Characteristics					
Number of teachers	8.42	8.35	8.54	8.78	-0.06	0.13	0.36	0.19	0.42	0.23
	(2.05)	(2.11)	(2.11)	(2.82)	(0.86)	(0.73)	(0.40)	(0.60)	(0.32)	(0.59)
Number of civil servant teachers	3.97	3.90	3.87	4.16	-0.07	-0.10	0.19	-0.03	0.27	0.30
	(1.51)	(1.69)	(1.68)	(1.76)	(0.79)	(0.71)	(0.49)	(0.92)	(0.37)	(0.32)
Number of students	111.87	101.03	104.94	108.79	-10.84	-6.92	-3.07	3.91	7.76	3.85
	(52.14)	(42.31)	(39.81)	(47.64)	(0.19)	(0.39)	(0.72)	(0.58)	(0.32)	(0.61)
Private school	0.12	0.06	0.07	0.07	-0.06	-0.05	-0.04	0.01	0.02	0.00
	(0.33)	(0.24)	(0.26)	(0.26)	(0.22)	(0.37)	(0.39)	(0.73)	(0.72)	(0.98)

Notes: Standard errors clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table F.6: Impact on Student Learning Outcomes: No Individual Controls

	Indonesian		Mathematics		Average Score		Grade Repetition	
	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SAM	0.071	-0.010	0.054	0.002	0.066	-0.004	0.013	-0.000
	(0.045)	(0.042)	(0.044)	(0.061)	(0.040)*	(0.049)	(0.010)	(0.008)
SAM+Cam	0.191	0.115	0.186	0.155	0.190	0.135	0.005	0.013
	(0.048)***	(0.043)***	(0.049)***	(0.061)**	(0.044)***	(0.049)***	(0.010)	(0.008)
SAM+Score	0.090		0.066		0.085		0.012	
	(0.042)**		(0.044)		(0.038)**		(0.010)	
Control group mean							0.08	0.04
Control group raw-score mean	47.13	38.12	47.03	44.04	47.08	41.08		
Test of equality (P-val)								
SAM v. SAM+Cam	0.016	0.005	0.008	0.013	0.006	0.006	0.386	0.131
SAM+Cam v. SAM+Score	0.030		0.017		0.016		0.477	
SAM v. SAM+Score	0.676		0.805		0.652		0.881	
Randomization Inference								
(P-value, N = 1000)								
SAM	0.153	0.860	0.217	0.973	0.121	0.950	0.247	0.976
SAM+Cam	0.002	0.024	0.000	0.033	0.000	0.020	0.695	0.193
SAM+Score	0.063		0.161		0.045		0.288	
R2	0.282	0.064	0.276	0.113	0.355	0.116	0.106	0.059
Observations	31022	15611	31022	15611	31022	15611	24719	13257
Individual controls	No	No	No	No	No	No	No	No
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Control variables include sex, age dummies, both parents' education, baseline outcome, dummy variables for missing controls (one for each control variable), school-level mean scores, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. The randomization inference tests the sharp null hypothesis of no effect for each individual treatment (holding other treatment assignments constant). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table F.7: Impact on IRT-Corrected Student Learning Outcomes

	Uncorrected		IRT Corrected	
	2018	2019	2018	2019
	(1)	(2)	(3)	(4)
	Panel A. Indonesian			
SAM	0.094	0.014	0.095	0.016
	(0.037)**	(0.027)	(0.035)***	(0.026)
SAM+Cam	0.190	0.096	0.189	0.094
	(0.036)***	(0.028)***	(0.034)***	(0.028)***
SAM+Score	0.122		0.125	
	(0.034)***		(0.032)***	
	Panel B. Mathematics			
SAM	0.073	0.042	0.071	0.047
	(0.040)*	(0.044)	(0.040)*	(0.044)
SAM+Cam	0.202	0.176	0.205	0.181
	(0.041)***	(0.046)***	(0.040)***	(0.046)***
SAM+Score	0.094		0.099	
	(0.038)**		(0.038)***	
Observations	31022	15611	31022	15608
Individual controls	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes

Notes: Columns 1–2 are the main impact estimates for each respective subject from Table 3. Columns 3–4 are the standardized IRT-corrected scores where scores for students who did not advance to the next grade were replaced with a predicted score based on the IRT before being standardized. For mathematics, there was only one linked question between grades 3 and 4; therefore, for students who did not advance from grade 3, their actual mathematics score were used instead of the predicted score. Three students who were retained in grade 6 in 2019 were dropped in the IRT estimates because there was no grade 7 tests. Control variables include sex, age dummies, both parents' education, baseline outcome, dummy variables for missing controls (one for each control variable), school-level mean scores, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table F.8: Impact on Student Learning Outcomes for Grades 3-6†

	Indonesian		Mathematics		Average Score	
	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)
SAM	0.096	0.009	0.089	0.048	0.097	0.029
	(0.035)***	(0.027)	(0.042)**	(0.045)	(0.035)***	(0.032)
SAM+Cam	0.150	0.096	0.182	0.183	0.168	0.137
	(0.035)***	(0.028)***	(0.041)***	(0.047)***	(0.034)***	(0.034)***
SAM+Score	0.100		0.082		0.095	
	(0.034)***		(0.038)**		(0.032)***	
Control group mean						
Control group raw-score mean	49.17	37.63	46.78	43.63	47.97	40.63
Test of equality (P-val)						
SAM v. SAM+Cam	0.156	0.002	0.033	0.003	0.051	0.001
SAM+Cam v. SAM+Score	0.186		0.018		0.041	
SAM v. SAM+Score	0.924		0.863		0.969	
Observations	21448	15108	21448	15108	21448	15108
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standardized scores are grade adjusted. †The outcome variables are for students who would have been at grades 3-6 at each respective year. Control variables include sex, age dummies, both parents' education, baseline outcome, dummy variables for missing controls (one for each control variable), school-level mean scores, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table F.9: Heterogeneous Impacts on Learning and Teacher Presence by Punishment Norms

	Working		Teaching	
	2018	2019	2018	2019
	(1)	(2)	(3)	(4)
SAM	0.046	-0.038	-0.034	0.054
	(0.057)	(0.088)	(0.065)	(0.102)
SAM+Cam	-0.041	-0.064	-0.033	-0.062
	(0.061)	(0.082)	(0.073)	(0.104)
SAM+Score	-0.025		0.019	
	(0.068)		(0.069)	
Above-Median Punishment	-0.151	-0.016	-0.110	0.042
	(0.068)**	(0.096)	(0.088)	(0.107)
$\ldots \times$ SAM	0.098	0.025	0.197	-0.070
	(0.086)	(0.133)	(0.096)**	(0.144)
$\ldots \times$ SAM+Cam	0.290	-0.041	0.169	-0.051
	(0.096)***	(0.126)	(0.131)	(0.148)
$\ldots \times$ SAM+Score	0.036		0.071	
	(0.088)		(0.110)	
Observations	714	467	616	430
Controls	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes

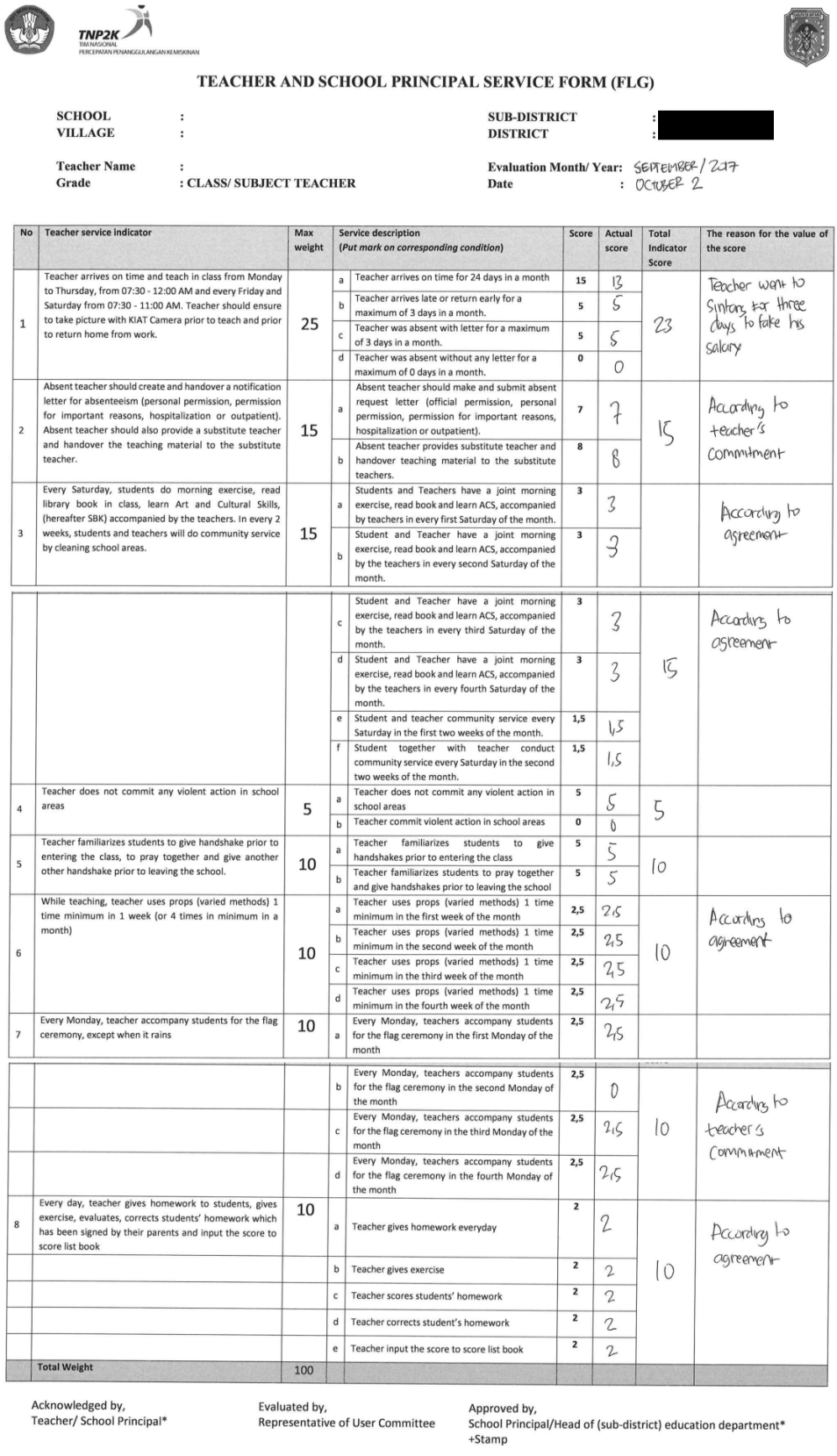
Notes: Treatment variables are interacted with a punishment norm variable based on a lab-in-the-field behavioral games in 182 out of 270 schools. The variable captures whether the average parent participants in the school imposed an above-median penalties to group members who had a below-average contribution in the public goods game. Teacher respondents include the sample of class teachers. Individual controls include sex, age, education, and the baseline outcome. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Controls also include dummy variables for missing controls (one for each control variable). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table F.10: Impact on School Principal's Reported Excuse for Teacher Absences

	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A. Off-school assignments (No penalty)				*Panel B. Late arrival, early departure (0.5-1.5% penalty)†*			
SAM	0.100	0.199	0.106	0.250	0.019	-0.140	0.095	-0.158
	(0.074)	(0.088)**	(0.119)	(0.115)**	(0.048)	(0.052)***	(0.085)	(0.070)**
SAM+Cam	0.084	0.206	-0.065	0.259	-0.032	-0.133	0.014	-0.194
	(0.089)	(0.090)**	(0.118)	(0.121)**	(0.053)	(0.053)**	(0.072)	(0.074)**
SAM+Score	0.112		0.108		0.004		-0.009	
	(0.073)		(0.099)		(0.056)		(0.063)	
TSA-receiving teacher			0.083	0.345			-0.023	-0.020
			(0.102)	(0.135)**			(0.118)	(0.070)
… × SAM			-0.008	-0.109			-0.126	0.041
			(0.171)	(0.180)			(0.124)	(0.080)
… × SAM+Cam			0.272	-0.130			-0.066	0.106
			(0.172)	(0.176)			(0.115)	(0.109)
… × SAM+Score			-0.015				0.044	
			(0.130)				(0.119)	
Control group mean	0.20	0.28	0.20	0.28	0.09	0.15	0.09	0.15
Observations	338	241	338	241	338	241	338	241
	Panel C. Sick and personal leave (0 - 2% penalty)‡				*Panel D. Others*			
SAM	-0.068	-0.052	-0.017	-0.113	-0.050	0.001	-0.173	0.027
	(0.070)	(0.069)	(0.089)	(0.099)	(0.095)	(0.088)	(0.127)	(0.127)
SAM+Cam	0.025	-0.144	0.115	-0.312	-0.077	0.063	-0.053	0.242
	(0.072)	(0.067)**	(0.097)	(0.120)**	(0.101)	(0.092)	(0.139)	(0.129)*
SAM+Score	0.041		0.066		-0.146		-0.149	
	(0.065)		(0.089)		(0.096)		(0.119)	
TSA-receiving teacher			0.090	-0.171			-0.151	-0.156
			(0.097)	(0.124)			(0.141)	(0.128)
… × SAM			-0.075	0.129			0.192	-0.059
			(0.127)	(0.137)			(0.200)	(0.167)
… × SAM+Cam			-0.163	0.309			-0.059	-0.286
			(0.126)	(0.166)*			(0.189)	(0.169)*
… × SAM+Score			-0.035				0.001	
			(0.121)				(0.164)	
Control group mean	0.21	0.22	0.21	0.22	0.51	0.34	0.51	0.34
Observations	338	241	338	241	338	241	338	241
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The outcomes are reasons for absent teachers reported by school pricipals. The daily penalty to the TSA for each excuse is in the parentheses. †Daily TSA penalties for late arrivals/early departures range from 0.5 percent (for less than 30 minutes) to 1.5 percent (for more than 1 hour). ‡Daily TSA penalties for sick and personal leaves range from 0 (e.g., under hospitalization, or the first 10 days of maternity leaves or as an outpatient) to 2 percent (a personal leave or sick leave without proper notice). Individual controls include sex, age, and education. School-level controls include school-level mean scores for the outcome, the total number of teachers and civil-servant teachers, the total number of students, and dummy variables for whether the school is a private school and whether it was among the three control schools who became TSA-ineligible due to the change in the government's definition of remoteness. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

II Figures



TNP2K
TIM NASIONAL PERCEPATAN PENANGGULANGAN KEMISKINAN

TEACHER AND SCHOOL PRINCIPAL SERVICE FORM (FLG)

SCHOOL :
VILLAGE :
SUB-DISTRICT : [redacted]
DISTRICT :

Teacher Name :
Grade : **CLASS/ SUBJECT TEACHER**
Evaluation Month/ Year: SEPTEMBER / 2017
Date : OCTOBER 2

No	Teacher service indicator	Max weight		Service description (*Put mark on corresponding condition*)	Score	Actual score	Total Indicator Score	The reason for the value of the score
1	Teacher arrives on time and teach in class from Monday to Thursday, from 07:30 - 12:00 AM and every Friday and Saturday from 07:30 - 11:00 AM. Teacher should ensure to take picture with KIAT Camera prior to teach and prior to return home from work.	25	a	Teacher arrives on time for 24 days in a month	15	13	23	Teacher want to Sintang for three days to take his salary
			b	Teacher arrives late or return early for a maximum of 3 days in a month.	5	5		
			c	Teacher was absent with letter for a maximum of 3 days in a month.	5	5		
			d	Teacher was absent without any letter for a maximum of 0 days in a month.	0	0		
2	Absent teacher should create and handover a notification letter for absenteeism (personal permission, permission for important reasons, hospitalization or outpatient). Absent teacher should also provide a substitute teacher and handover the teaching material to the substitute teacher.	15	a	Absent teacher should make and submit absent request letter (official permission, personal permission, permission for important reasons, hospitalization or outpatient).	7	7	15	According to teacher's commitment
			b	Absent teacher provides substitute teacher and handover teaching material to the substitute teachers.	8	8		
3	Every Saturday, students do morning exercise, read library book in class, learn Art and Cultural Skills, (hereafter SBK) accompanied by the teachers. In every 2 weeks, students and teachers will do community service by cleaning school areas.	15	a	Students and Teachers have a joint morning exercise, read book and learn ACS, accompanied by teachers in every first Saturday of the month.	3	3		According to agreement
			b	Student and Teacher have a joint morning exercise, read book and learn ACS, accompanied by the teachers in every second Saturday of the month.	3	3		
			c	Student and Teacher have a joint morning exercise, read book and learn ACS, accompanied by the teachers in every third Saturday of the month.	3	3	15	According to agreement
			d	Student and Teacher have a joint morning exercise, read book and learn ACS, accompanied by the teachers in every fourth Saturday of the month.	3	3		
			e	Student and teacher community service every Saturday in the first two weeks of the month.	1,5	1,5		
			f	Student together with teacher conduct community service every Saturday in the second two weeks of the month.	1,5	1,5		
4	Teacher does not commit any violent action in school areas	5	a	Teacher does not commit any violent action in school areas	5	5	5	
			b	Teacher commit violent action in school areas	0	0		
5	Teacher familiarizes students to give handshake prior to entering the class, to pray together and give another other handshake prior to leaving the school.	10	a	Teacher familiarizes students to give handshakes prior to entering the class	5	5	10	
			b	Teacher familiarizes students to pray together and give handshakes prior to leaving the school	5	5		
6	While teaching, teacher uses props (varied methods) 1 time minimum in 1 week (or 4 times in minimum in a month)	10	a	Teacher uses props (varied methods) 1 time minimum in the first week of the month	2,5	2,5	10	According to agreement
			b	Teacher uses props (varied methods) 1 time minimum in the second week of the month	2,5	2,5		
			c	Teacher uses props (varied methods) 1 time minimum in the third week of the month	2,5	2,5		
			d	Teacher uses props (varied methods) 1 time minimum in the fourth week of the month	2,5	2,5		
7	Every Monday, teacher accompany students for the flag ceremony, except when it rains	10	a	Every Monday, teachers accompany students for the flag ceremony in the first Monday of the month	2,5	2,5		
			b	Every Monday, teachers accompany students for the flag ceremony in the second Monday of the month	2,5	0	10	According to teacher's commitment
			c	Every Monday, teachers accompany students for the flag ceremony in the third Monday of the month	2,5	2,5		
			d	Every Monday, teachers accompany students for the flag ceremony in the fourth Monday of the month	2,5	2,5		
8	Every day, teacher gives homework to students, gives exercise, evaluates, corrects students' homework which has been signed by their parents and input the score to score list book	10	a	Teacher gives homework everyday	2	2	10	According to agreement
			b	Teacher gives exercise	2	2		
			c	Teacher scores students' homework	2	2		
			d	Teacher corrects student's homework	2	2		
			e	Teacher input the score to score list book	2	2		
	Total Weight	100						

Acknowledged by,
Teacher/ School Principal*

Evaluated by,
Representative of User Committee

Approved by,
School Principal/Head of (sub-district) education department*
+Stamp

Figure F.1: A Sample of the Community Scorecard

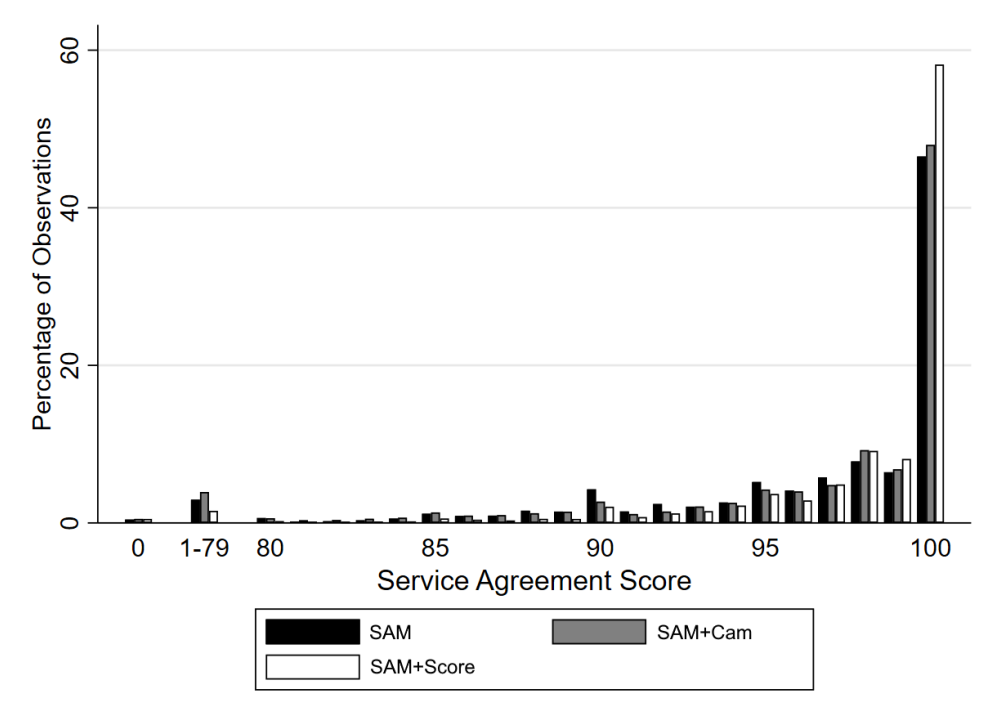
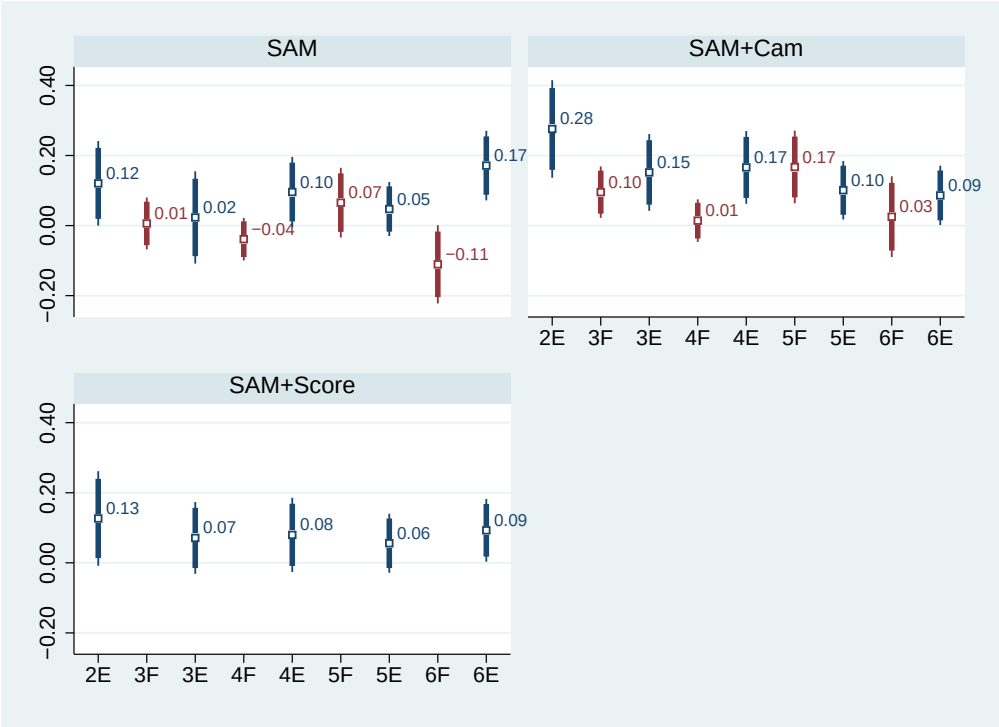
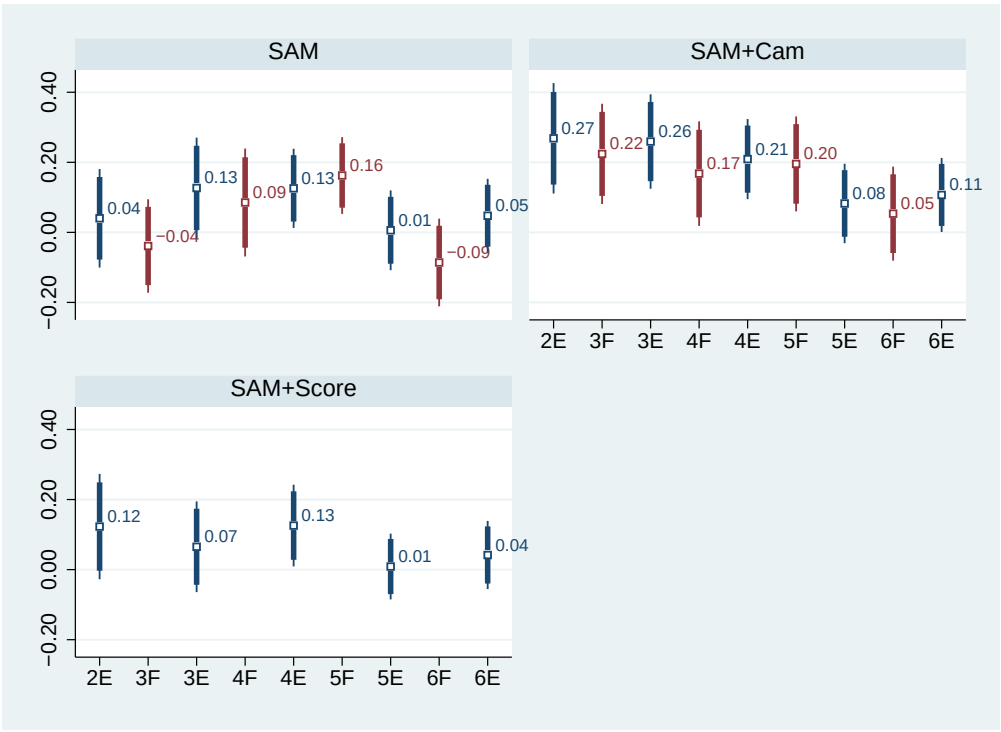


Figure F.2: The Distribution of the Service Agreement Scores by Treatment



(a) Indonesian



(b) Mathematics

Notes: The numbers on the horizontal axis refer to the grade at the time of measurement. E/F indicates whether the outcome was measured at endline/follow-up respectively. The outcome variable is the standardized mean of the Indonesian and Mathematics scores.

Figure F.3: Impact on Mean Scores at Midline and Endline by Baseline Grade