NBER WORKING PAPER SERIES

DOES PERFORMANCE PAY ENHANCE SOCIAL ACCOUNTABILITY? EVIDENCE FROM REMOTE SCHOOLS IN INDONESIA

Arya Gaduh
Menno Pradhan
Jan Priebe
Dewi Susanti

Working Paper 30758
http://www.nber.org/papers/w30758

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2022

We thank our partners at the Indonesian Ministry of Education and Culture, the National Team for Acceleration of Poverty Reduction (TNP2K), and the district governments of Ketapang, Landak, Sintang, West Manggarai, and East Manggarai for their support and assistance. We also thank Usha Adelina, Emilie Berkhout, Kurniawati, Sharon Kanthy Lumbanraja, Marliyanti, and Indah Ayu Prameswari for their excellent research assistance. 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 PacDev Conference, 2019 MIEDC, 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, the Hong Kong University Business School, and the University of Arkansas for helpful comments and suggestions. We acknowledge financial support from the World Bank's Local Solutions to Poverty Trust Fund (DFAT) and Local Solutions to Development Trust Fund (USAID), and from the SMERU Research Institute (RISE Indonesia Study). This study was registered in the American Economic Association Registry for randomized control trials under trial AEARCTR-0003157. Dewi Susanti was an employee of the World Bank. Arya Gaduh, Menno Pradhan, and Jan Priebe were consultants for the World Bank. Menno Pradhan also worked as a co-principal investigator under the SMERU RISE project. 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Does Performance Pay Enhance Social Accountability? Evidence from Remote Schools in Indonesia

Arya Gaduh, Menno Pradhan, Jan Priebe, and Dewi Susanti

NBER Working Paper No. 30758
December 2022
JEL No. H52,I21,I25,I28,O15

ABSTRACT

Social accountability offers a viable alternative to top-down supervision of service delivery in remote areas when travel cost renders the latter ineffective. However, this bottom-up approach may not be effective when the community has weak authority relative to the service provider. This paper investigates whether giving communities authority over teacher performance pay improves the effectiveness of social accountability in Indonesia's remote schools. We tested incentive contracts based on either camera-verified teacher presence or community ratings of teacher performance. Social accountability had the strongest and most persistent impact on student learning when combined with the former. The results indicate that when the principal (community) has weak authority vis-à-vis the agent (regular teachers), increasing that authority using an incomplete but verifiable contract works better than using a more comprehensive but subjective one.

Arya Gaduh
Department of Economics
University of Arkansas
402 Business Building
Fayatteville, AR 72701
and NBER
AGaduh@walton.uark.edu

Menno Pradhan
VU University / University of Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
m.p.pradhan@vu.nl

Jan Priebe
German Institute of Global
and Area Studies
jpriebe@uni-goettingen.de

Dewi Susanti
Global School Leaders 12329
Culver Blvd
Los Angeles, CA 90066-6221
dewi@globalschoolleaders.org

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/3157

1 Introduction

Where a child grows up profoundly affects her intergenerational mobility (Chetty and Hendren, 2018; Laliberté, 2021). In low- and middle-income countries (LMICs), growing up in more remote areas hampers upward mobility in educational attainment (Alesina et al., 2021; van der Weide et al., 2021). While place-based policies in more disadvantaged regions have reduced disparities in access to basic education with measurable impacts on intergenerational mobility (see e.g., Akresh et al., 2018), learning deficiencies in remote areas of LMICs remain large (World Bank, 2018).

Improving learning in remote areas is much harder than improving access. Remoteness increases the cost of top-down supervision, lowering accountability for public spending (Muralidharan et al., 2017). Social accountability initiatives offer a bottom-up alternative to top-down supervision: by providing citizens with information, voice, and influence, they enable citizens to hold teachers directly accountable and reduce the need for costly top-down supervision. However, evaluations of this approach identified communities' lack of authority vis-a-vis schools and teachers as a key constraint for success (Muralidharan, 2017).

This paper investigates whether expanding citizen authority over the incentives of regular teachers improves the effectiveness of social accountability interventions in education in remote areas. Most social accountability initiatives rely on social pressure to improve teacher effort, but do not affect extrinsic rewards such as teacher pay or tenure. There is evidence that giving communities authority over short-term contract teachers can improve learning outcomes (Duflo et al., 2015). However, whether this approach can work on regular teachers remains an open question. First, regular teachers in public schools are usually hired by the government; local communities have little authority in their hiring or firing. Moreover, regular teachers are more educated than the average parent in these rural communities. Consequently, regular teachers have stronger informal authority and are more able to resist attempts to hold them accountable.

We evaluate a "social accountability only" and two "social accountability plus" treatments that targeted regular teachers in a large sample of mostly public primary schools in Indonesia's remote areas. The Social Accountability Mechanism (SAM) treatment only had the social accountability component. Parents and teachers formulated a joint agreement specifying their respective responsibilities to improve student learning after being informed of the children's learning level. Communities used the agreement to formulate teacher-specific scorecards, and established a user committee (UC) to monitor and evaluate their implementation. Every month, the UC evaluations were discussed and finalized in public meetings, and sent to the district education office. The social accountability component improved the flow of information on teacher performance to communities and provided the UC with a direct channel to report to school administrators.

The two "plus" treatments — SAM+Cam and SAM+Score — included incentive contracts that targeted the Teacher's Special Allowance (TSA), a government hardship allowance equal to the base salary for regular (and mostly civil service) teachers working in remote areas. Both treatments penalized this allowance for poorly performing TSA-receiving teachers. In SAM+Cam, the penalty was based solely on the teacher's presence indicator, which was verified every month by the community based on the selfies that the teacher took at the start and end of every workday using a specially provided smartphone

camera. Meanwhile, in SAM+Score, the penalty was based on the overall score on the scorecard.

These treatment variations allow us to evaluate the effectiveness of the different ways communities could hold teachers accountable. SAM relied on social pressure during the monthly evaluation meetings, where teachers were reminded of their commitments in the service agreement. SAM+Cam and SAM+Score used an incentive contract, which strengthened the UCs' authorities by enabling them to financially penalize teachers for poor performance. These SAM+ treatments varied in the indicators used to measure performance: SAM+Cam used a single verifiable indicator, while SAM+Score used a compound score based on (more subjective) evaluations of teacher effort by the UC. We use this treatment variation to study the role of performance indicators in the design of incentive contracts to enhance social accountability in remote areas.

We conducted the study in 270 remote schools in 5 disadvantaged districts in the East Nusa Tenggara and West Kalimantan provinces of Indonesia between October 2016 and May 2019. In its first year, an outside facilitator led the village-level implementation. The facilitator left at the end of 2017 and a trained cadre recruited from the village continued to perform those duties. The endline data for the one-year impact evaluation were collected in early 2018. Moreover, to study the sustainability of these impacts in the absence of outside facilitators, we collected a follow-up survey in all but the SAM+Score schools in early 2019.

We find all treatments increased learning outcomes, with SAM+Cam showing the largest improvements. After one year, SAM and SAM+Score improved learning by between 0.08 and 0.11 standard deviation (sd). By comparison, SAM+Cam approximately doubled the effect sizes with a learning improvement of 0.20 sd To put these magnitudes in perspective, a meta-study of randomized experiments in primary schools in LMICs estimated the learning impacts of interventions that improved information, school management, and student/teacher incentives at, respectively, 0.05, 0.06, and 0.09 s.ds. (McEwan, 2015). Overall, the impacts of our interventions do not differ by gender, but are more positive for students in earlier grades and those who performed better at baseline. For SAM+Cam, the learning effects persisted into the second year and, importantly, were not merely knock-on impacts from the first-year.

The overall effects of the treatments on teacher presence and self-reported work hours were negligible. However, teachers shifted effort toward learning enhancing activities in SAM and SAM+Cam, and parents reported meeting with teachers more often. Looking at the effects on TSA and non-TSA teachers separately, we find that the salary incentive treatments reduced the presence of non-TSA teachers relative to TSA teachers. No such effect was found for the SAM treatment. 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, parental satisfaction, and aspirations. Almost all these effects persisted into the second year. Parents met with teachers more often, although these effects 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 were not yet aware of fading teacher efforts in the second year.

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).[1] We show in the model that if parental assessments can penalize the TSA and parents can commit to assessing teachers truthfully, they can induce teachers to exert a high efforts. However, when evaluations are more subjective (which was the case in SAM+Score relative to SAM+Cam) and teachers can retaliate, parents will tend to overrate teachers and thus find it more difficult to induce a high effort from teachers.

We find empirical evidence consistent with this model. First, we find that a stronger parental 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. We find larger student learning gains and positive improvements in teacher behavior in communities with a higher propensity to punish free riders. Second, we also find evidence that the more subjective performance measures in SAM+Score led to disagreements between teachers and the UC. Teachers in SAM+Score treatment schools were more likely to put pressure on the UC to increase their score.[2] In SAM+Cam, such disagreements were tempered by the hard evidence provided by the camera. At the same time, evaluation scores in SAM+Score are somewhat higher than in SAM and SAM+Cam even though our independent measures of teacher effort and student learning did not corroborate these scores.

Remarkably, even though we worked in places that are more difficult to reach and therefore costlier to manage, the cost-effectiveness of our interventions is comparable to other interventions that aim to improve learning in LMICs. SAM+Cam, which was the most successful among our interventions, improved learning outcomes by 0.2 sd at the cost of USD 44 (in current 2017 dollars) per student. This cost is somewhere in the middle of the distribution of the cost-effectiveness of the various interventions reported in JPAL (2019).

Our paper contributes to the evidence on the effectiveness of participatory programs to improve learning in LMICs. Past studies have shown mixed results from programs designed to increase demand for better education by raising awareness about learning (Banerjee et al., 2010; Lieberman et al., 2014; Afridi et al., 2020; Barrera-Osorio et al., 2020a,b) and/or empowering community groups through training and grants (Banerjee et al., 2010; Pradhan et al., 2014; Barrera-Osorio et al., 2020b). While these programs often induced behavioral changes in parents, they generally did not translate into changes in teacher effort — in part, because parents were insufficiently empowered vis-a-vis teachers (Muralidharan, 2017). Our paper offers the first evidence that incentive payments based on community monitoring can strengthen the effectiveness of a self-standing bottom-up accountability intervention in improving learning. This is particularly relevant for environments where top-down supervision is costly.

We also contribute to the question on the role of performance measures in incentive contracts. Much research in contract theory focuses on the role of subjective measures to correct the distortionary effects of (incomplete) objective measures (Holmstrom and Milgrom, 1991; Prendergast and Topel, 1993; Baker et al., 1994). However, there is limited evidence on how the principal's perceived authority can affect the effectiveness of such measures. Macleod (2003) shows that when the agent does not trust the principal's assessment, the optimal contract would pool most of the agents and could only discriminate the

[1] We find some evidence that in the first year, SAM+Cam increased external supervision from district officials. These effects did not persist into the second year. We did not find similar effects for SAM or SAM+Score.

[2] The qualitative study of Bjork and Susanti (2020) corroborated this result. They report more teacher dissatisfaction regarding the role assigned to the UC in SAM+Score compared to that in SAM+Cam.

worst performing agents — in part, due to the principal's desire to avoid conflict ex post. Our evidence suggests that when the principal's authority is weak relative to the agent, an incomplete but verifiable performance measure yields a better outcome than more comprehensive subjective measures.[3] Teachers bargain more over subjective evaluations and our model shows how bargaining dampens the effectiveness of the performance pay contract. Our results highlight a crucial design element when incorporating incentive contracts into community-based monitoring initiatives.

Scaling up successful pilots has proven to be difficult, and it is not easy to pinpoint where the problem arises (Bold et al., 2018; Raffler et al., 2019). 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 toward scaling up — to wit, the handing over of tasks from the project facilitator to a local cadre — was taken. We find that although the administrative processes were sustained, impacts, especially on teacher efforts, were weakened. The fact that we worked with a government allowance, backed up by regular budgets and legalizing regulations, helped to keep processes intact during the second year. The results also indicate that communities need outside support to sustain impacts.

The rest of the paper proceeds as follows. The next section discusses the context for our experiment. Section 3 details the experimental design, including both the implementation and data collection timelines. Section 4 describes the data collection instruments and summarizes baseline characteristics of students, parents, and teachers in the sample. Section 5 presents the results on the primary outcome, namely student learning outcomes. Section 6 explores the mechanisms for our results. We first present a conceptual model followed by the empirical results on teacher effort, parental engagements, and school management. Section 7 discusses various aspects related to the interventions' potential sustainability, including a discussion on 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 servants, whose hiring and salary standards are substantially higher than the others, 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 (Usman et al., 2004). In 2014, it stood at 19.3 percent compared to the national average of 9.4 percent (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

[3]Our result contrasts with that of Andrabi and Brown (2021) who study the choice of performance measures in the context of private schools in urban Pakistan. They find that an incentive contract based on a comprehensive performance evaluation by the school principal has similar effects on test scores as a contract based on test scores only, but performs better in terms of socio-emotional outcomes. This difference could be explained by the stronger authority of school principals in their study vis-a-vis the user committees in our study.

evidence suggests they do not improve quality. Studies find that TSA recipients are 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 be up to the teacher's monthly base salary. Starting in 2017 (right before our interventions), the government reformed its approach to distributing the TSAs.[4] Teachers hired by the government — either as a tenured civil servant or under a fixed-term contract — are eligible to receive the TSA if they are working in villages that are designated as remote and very disadvantaged.[5] Villages are designated as very disadvantaged and remote by the central government based on a national index. This new approach improved both the coverage and reliability of TSA distribution.

In addition to the challenge of teacher retention, it is harder for local government agencies to supervise teachers in remotely located schools. Travel distance makes on-site supervisions costly and as a result, poor teacher performance can go 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.[6] 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 (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.[7] 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

[4]This approach improved on a more subjective and discretionary system. Previously, the national budget for TSAs was determined based on proposals from the districts. The TSAs were then distributed to the districts, which would distribute them 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 funding 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).

[5]In addition to the two types of government-hired teachers, there are also school-contracted teachers with a temporary employment status. The monthly base salaries are highest for civil servant teachers (between around USD 108 and USD 408 depending on seniority), followed by fixed-term contract teachers (between around USD 73 to 146), and school-contracted teachers (between USD 22 and 51).

[6]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).

[7]*KIAT Guru* is the abbreviation of the project's name in Indonesian, that is, "*Kinerja dan Akuntabilitas Guru*" meaning "Teacher Performance and Accountability".

and beneficiaries) and agents (i.e., the service providers) (World Bank, 2003; Pritchett, 2015). The program's elements include: (i) having a standard to hold service providers accountable; (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 (Ringold et al., 2012; Joshi, 2013). 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).

The design of our interventions builds on the lessons of Pradhan et al. (2014), which tested different ways to strengthen school committees in rural Central Java, Indonesia. That study shows the importance of involving local leadership and ensuring that community involvement leads to concrete actions that improve education. It also underlines the difficulty of inducing increased teacher efforts if there are no incentives attached to community action. Its pathway analysis suggested that the positive effects on learning were mostly a result of increased inputs of the community and not teacher effort.

KIAT Guru's final design was informed by an operational pilot in very remote villages of Indonesia. The pilot tested the implementation of key processes (e.g., facilitation of community meetings, pay-for-performance mechanisms), legal and administrative regulations, process-monitoring instruments, and the survey instruments. Key lessons learned from the pilot helped refine the design, particularly on district and village selection criteria.[8]

3.1 Experimental Treatments

We introduced three experimental treatments. The common component of all three treatments was the social accountability mechanism to establish community monitoring of locally-formulated standard. Two treatments enhanced social accountability with a performance pay component that employed different ways to use the community evaluation results to incentivize teacher pay. Below, we describe each component and the variations that define each treatment.

3.1.1 Social Accountability Mechanism

The social accountability component helped communities formulate a local service standard and establish a community monitoring institution known as the user committee (UC). To establish a local service standard, project facilitators organized a set of meetings involving parents, community members and leaders, and school management to discuss and agree on the role that each of them should play to improve children learning. Based on that discussion, they formulated a service agreement comprising the set of actions to improve the learning environment that each of them would commit to.

This service agreement was the basis for the principal- and teacher-specific scorecard. The scorecard had between 5 and 8 indicators, each with a target that the principal and teachers committed to. ''Presence during office hours" was always included, but otherwise meeting participants could freely choose

[8] Among other things, we find that the success of the program required commitments at multiple levels. The community needed to be willing to contribute time and resources, and to demand better education services. Both district and school managements needed to be sufficiently transparent about their finances. Finally, the district bureaucracy needed to fully support program implementation. The operational pilot was conducted from June 2014 to December 2015.

what to include. Each indicator was assigned a weight (for a total of 100) based on participants' belief of its importance to improving learning. The UC would devise a scoring guideline for each indicator.

The community could revise the scorecard in a meeting held a few months later. To help inform communities of their children's literacy and numeracy skills, the UC with the village cadre administered a set of learning diagnostic tests developed by the project.[9] The 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 village-wide meeting.

The UC was responsible for monitoring teacher compliance to their scorecard. The UC had a minimum of nine members, and at least half of them were female. It also included parents who represented each of the grade levels.[10] The UC conducted monthly meetings to review the implementation of the service agreement and evaluate the scorecard. The UC presented its monthly evaluation of the scorecard and gave each teacher an opportunity to respond. Once the scores were 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.

To maintain the sustainability of the interventions, we recruited a village cadre from each village who would undertake the facilitator responsibilities once the project was completed. In the first year, each village cadre co-organized and co-facilitated the various meetings with the facilitator. Along with UC members, they were also trained to implement the aforementioned learning diagnostics. Seventy-five percent of the cadres were appointed at the first village meeting.[11]

3.1.2 Varying the Enforcement Mechanisms

The treatments varied in how the UC evaluations could 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. SAM had no performance pay component and eligible teachers always received their full TSA amount. SAM+Cam and SAM+Score differed in the indicators and tools that were used to penalize poor performance by cutting the TSA. In all treatments, non-TSA teachers received similar evaluations by the UC, but their evaluations had no effect on their income.

Table 1: Summary of the Treatments

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

[9] The learning diagnostics measure children's skills along a learning continuum based on the national curriculum.

[10] 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.

[11] Marliyanti et al. (2022) document the tiered process to train the key actors in the social accountability intervention.

In SAM+Cam, only the teacher presence indicator affected the TSA amount. Teachers in the SAM+Cam schools were given a tamper-proof smartphone camera to record their presence. They took 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 the teachers to account for their absences. Based on these daily records, the UC penalized teachers for each partial presence (up to 1.5 presence points), excused absence (2 points), or unexcused absence (5 points), and calculate the teacher's presence score for that month. To establish a norm for the maximum number of acceptable absences, we set a cutoff score of 85 (out of 100), below which a teacher would lose their full TSA for that month. Those who maintained a presence score of 85 or above received the share of the TSA equal to that score.[12]

In SAM+Score, the TSA amount was determined by the total weighted scores of all indicators in the scorecard. Unlike in SAM+Cam, there was no cut-off score below which a teacher would receive zero allowance. Because SAM+Score relies on a compound score of subjective indicators to measure performance, we wanted to avoid introducing a focal point for negotiations between teachers and the UC. Therefore, SAM+Score has a continuous penalty schedule in which the percentage of the TSA allowance received was equal to the evaluation score for that month. Furthermore, no camera was provided for the SAM+Score schools. Hence, the mandatory presence indicator needed to be proactively monitored by the UC following the steps suggested in their training.

We made sure that the TSAs were uniformly and reliably disbursed across control and treatment schools. Teachers were paid the TSA 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.

3.1.3 The Second-Year Implementation

At the end of the first year, village-level facilitator support ended. Village cadres took over the facilitator's role of organizing evaluation meetings. All other relevant SAM and performance-pay processes continued in the second year. These include the dispatch of monthly evaluation reports to district offices and the incentivization of the TSAs for schools in the performance-pay treatments.

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 bush communities). We also excluded districts with very expensive transportation costs for budgetary reasons, as well as conflict-prone areas and districts that were part of many other education pilots.[13] Finally, we only included districts with at least 40 rural primary schools that fulfilled the school eligibility criteria below. Our final list included three districts in West Kalimantan (Ketapang, Sintang, and Landak) and two districts in East Nusa Tenggara (East and West Manggarai).

[12] To accommodate the use of the smartphone camera, the facilitators held an additional training on its use during the monthly community meeting. Moreover, SAM+Cam schools added verification of the camera reports to its monthly meeting agenda.

[13] For example, we exclude Papua, West Papua, and certain districts in East Nusa Tenggara and Central Sulawesi.

We only included schools under the Ministry of Education and Culture that satisfied four eligibility requirements. First, each school must have had a minimum of 70 registered students. Second, at least 3 of its teachers must have received the TSA in 2017. Third, schools must satisfy a remoteness criterion — being located in a village that was at least a one-hour drive from the district capital. Our data suggest that on average, participating schools were located around 40 km (and about two hours 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.[14] More than 90 percent of the schools in our sample were public schools.

3.3 Treatment Assignment and Compliance

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

During the baseline survey, we discovered that three schools in East Manggarai 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 that was 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 for defining eligible TSA locations. It used a national index instead of district head recommendations to determine eligibility, and 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. We control for these three schools in our empirical analysis.

3.4 Timeline

Panel A of Figure 1 shows the implementation timeline. Seven meetings to set up the service agreement and the user committee were conducted between November 2016 and June 2017. A few months afterward (between July 2017 and January 2018), UCs held a meeting to reevaluate and revise the scorecard

[14] To maintain a reasonable implementation budget, we excluded sub-districts (kecamatan) with less than four eligible primary schools and those requiring costly additional travel (e.g. using boat/plane just to reach that specific sub-district). There were 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).

indicators. Facilitator support for village implementation concluded at the end of 2017. For 84 percent of the schools in the performance pay treatments, TSAs began to be incentivized in April 2017. By October 2017, all 135 schools under the performance-pay scheme had their TSAs incentivized. Appendix B provides additional details on implementation.

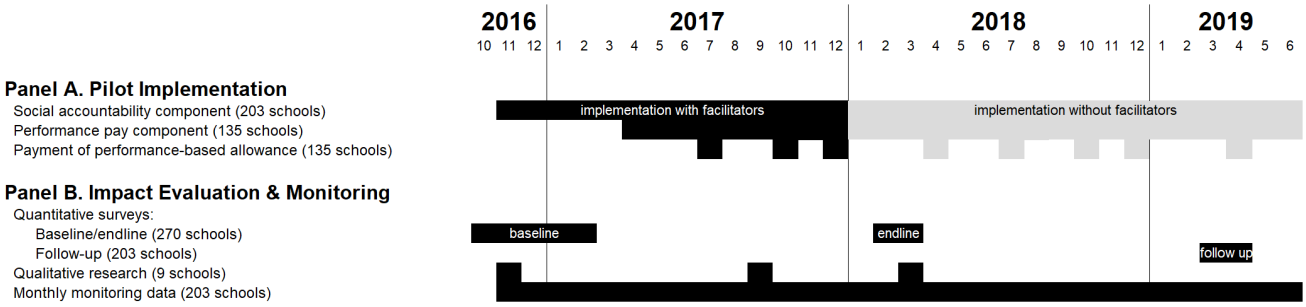


Figure 1: Implementation and Data Collection Timeline

Panel B shows the data collection timeline. For the impact evaluation, we collected three waves of the data. We conducted the baseline survey from October 2016 to February 2017. We conducted an endline survey to evaluate the one-year impacts from February until mid-April 2018, soon after the end of village-level facilitator support at the end of 2017.

The third wave, a follow-up survey to evaluate the second-year impact of the interventions, was collected in March to May 2019. This follow-up survey coincided with the plan by district governments to expand the SAM+Cam treatment to other schools in the 2019/2020 academic year. However, given budget constraints and the governments' expressed interest in SAM+Cam, we had to scale down the third wave in two ways. First, we did not collect data in the SAM+Score schools. Second, we did not collect learning outcomes from all grade 1 and grade 2 students. Since learning outcomes for grades 1 and 2 were collected on a one-to-one basis (instead of in a class setting) and were therefore more expensive to collect, we opted to include only grades 1 and 2 students who were part of an earlier survey.

4 Data

4.1 Instruments

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 (see Lumbanraja and Prameswari, 2021). They were designed based on frameworks and findings from other assessment tools (Gove and Wetterberg, 2011; Uwezo, 2012; Platas et al., 2014; ASER Centre, 2014) and they consist of: (i) a diagnostic test that rapidly captures students' competencies in literacy and numeracy; and (ii) an evaluation test that maps students' 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. For the baseline survey, the evaluation test was administered to 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 students in grades 1 to 6 who did not participate in the baseline survey.

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 from the Analytical and Capacity Development Partnership (2014) study in Indonesia, with additional inputs from the instruments used in the UNICEF (2012) study in Papua and West Papua. In its implementation, the enumerators implemented the TAS on the day of arrival which was 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 with children in participating schools (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 of 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; ACDP, 2014; Chang et al., 2014; World Bank, 2015, 2016).

Behavioral Experiment. At baseline, we conducted 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 E provides details of the implementation of the experiment.

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 an absenteeism rate of almost 20 percent. Class observations found that a quarter of teachers who were scheduled to teach did not teach.[15]

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

Despite low teacher efforts, parents were not aware of these problems. Our examination of the baseline data (not reported in the table) suggests that about 90 percent of parents believed that the quality of their children's school was 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. We also find limited parental supervision of their children education: Panel C of Table 2 reports that children were accompanied when learning at home for about 2.5 hours a week.

Appendix Tables G.1–G.2 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.[16] We estimate the treatment effects by regressing the following model:

$$Y_{ijt} = \alpha_k + \sum_{r \in R} \gamma_r T_{rj} + X'_{ij}\beta + \delta Y_{ij0} + \lambda \bar{Y}_{j0} + \varepsilon_{ijt}, \tag{1}$$

where Y_{ijt} = the student learning outcome for individual i in school j at time $t \in \{0, 1, 2\}$ (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.[17]

5.1 Main Results

Columns 1–4 of Table 3 present the impacts of the treatments on learning outcomes. To facilitate comparisons across studies, we use the mean of grade-adjusted standardized scores for Indonesian and mathematics as the measure of learning outcomes. The mean and standard deviation of the raw unstandardized scores are presented along with tests for cross-treatment differences in coefficient estimates in the panel below the coefficient estimates. We also included the p-values from a randomization inference procedure in the bottom panel. The table presents regression results with the control variables; as a robustness check, Appendix Table G.3 presents the regression results without these controls.

Columns 1–2 present the results for the full sample of students. Columns 3–4 present the results for students who would have been in grades 3–6 in each respective year to allow for consistent comparisons across years given the exclusion of almost all students in grades 1–2 in the follow-up survey.

[16] The pre-analysis plan for this study is documented in Bjork et al. (2018).
[17] The test is based on the user-written Stata command *ritest* (see Heß, 2017).

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.[18]

Column 1 shows positive one-year impacts on student learning outcomes for all treatments, with an impact that was strongest for SAM+Cam. SAM and SAM+Score treatments improved the mean learning outcome by 0.08 and 0.11 sds respectively and these effects are not statistically distinguishable from each other. These are at the higher end of the mean effect sizes of RCT-evaluated learning interventions in primary schools in LMICs that improved information flow (0.05 sd), school management (0.06 sd), and student/teacher incentives (0.09 sd) (McEwan, 2015). In contrast, SAM+Cam yielded a learning impact of 0.20 sd, almost twice as large as that of SAM+Score.

We also find that the SAM+Cam impact was persistent going into the second year. Because students from grades 1 and 2 were excluded in the follow-up survey, estimating the persistence of the learning impacts using the full sample (as in columns 1–2) may be biased if learning impacts were heterogeneous by grade. Appendix Figure G.1 shows that the impacts of SAM+Cam are stronger for lower grades. In columns 3–4, we addressed this issue by focusing on the sample of students in grades 3–6. Within this sample, there was a small decay in the learning impact of SAM+Cam in the second year of around one-sixth (from 0.17 to 0.13 sd). Notably, we show in Appendix Table G.4 (columns 5–8) that this temporal decay only occurred for Indonesian and not for mathematics.

In Appendix D, 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 D.1 shows that about half of the two-year impacts of SAM+Cam were 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 sd) while its two-year impact on mathematics was small (0.04 sd), 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 5 and 6 of Table 3 suggest that there was no one- or two-year impact of any of the treatments on students' grade repetition. This result reassures that our estimated learning impacts did not arise from the impact of the interventions on the school's grade retention strategy. Appendix Table G.5 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 at 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 4 presents the regressions of the student's attrition on their schools' treatment status. Columns 1 and 3 show that students in the SAM+Score treatment are less

[18] As discussed in Section 3.4, we did not survey the SAM+Score schools in 2019 due to budget constraints.

likely to attrit compared to the control schools. However, using interactions between the student's SLA performance at baseline with their school's treatment status (columns 2 and 4), we find no evidence of selective attrition based on academic ability.

5.3 Heterogeneity Analysis

To estimate these heterogeneous impacts, we use the following specification:

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

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, and the other variables are as defined in Equation 1.

Table 5 presents these results with the heterogeneity 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 was 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 with a TSA teacher did not strengthen the benefits of the interventions.

On the one hand, we find that better students benefited more from these interventions. 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.

On the other hand, the positive treatment impacts were more persistent for weaker schools at baseline. To examine persistence, we again focus on the sample of students in grades 3–6. Columns 8–9 present the heterogeneous treatment impact by school quality (measured by the school-level average of the baseline standardized SLA scores) for these students. Column 8 shows that the one-year impacts of the interventions do not differ by school quality.[19] However, column 9 shows that the temporal decay in the impacts of SAM and SAM+Cam were primarily experienced by the better schools at baseline. Among below-median quality schools, the impact of SAM decayed slightly in the second year (from 0.10 to 0.08 sd), while that of SAM+Cam increased over time (from 0.15 to 0.20 sd). We observe larger decays among the above-median schools, partly driven by a large increase in the performance of the 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 motivate

[19] In results not shown, we find similar results of no heterogenous one-year impact by treatment when estimating using the full sample.

our empirical analysis, 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 (Holmstrom and Milgrom, 1991; Baker et al., 1994; 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

We provide a stylized model of the theory of change that underlies the interventions. We view teacher and parent efforts as complementary inputs to student learning. Teachers and parents can provide either a low or high effort. With weak external supervision, the teacher's intrinsic motivation is insufficient to induce a high effort in the absence of interventions. Parents are more motivated to put in a high effort into their child's learning, but need the teacher's high effort to make it worthwhile. If the teachers put in a high effort, parents will do so as well.

Table 6a summarizes the strategic interactions between teachers and parents under these conditions.[20] 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.[21] Parents receive the highest payoff when both parents and teachers exert high efforts. Note that these payoffs are based on the observed effort levels (of the other agent) — 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 exert a low effort. Under the status quo, the pure Nash equilibrium is when both teachers and parents exert low efforts.

6.1.2 Parent-Teacher Agreement as a Commitment Contract

We can model the joint agreement between parents and teachers as a commitment contract. Following Gerber and Wichardt (2009), we model teachers' and parents' documented commitment to exerting high effort as an upfront payment (or "deposit") d_t and d_p respectively. If they exert a low effort, they lose these deposits. Because $P4 > P3$, the parental commitment (d_p) to maintain the high-learning Nash equilibrium is zero as long as teachers exert a high effort. We therefore 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 $d_t > T2 - T1$, high-learning is the only equilibrium. The following result therefore follows.

Result 1. *Teacher and parent efforts are increasing in the teacher's commitment cost d_t.*

The treatments vary the way teachers are penalized for falling short of their commitments. In SAM, the cost will be to their reputation. Meanwhile in SAM+Cam and SAM+Score, the performance pay

[20] In Appendix C, we show that such a payoff matrix could arise from a linear learning production function with positive complementarities in teacher and parent efforts, combined with simple utility functions that capture how teachers and parents weigh the benefit from student learning against the cost of putting in effort.

[21] Note that $T2 > T3$ gives rise to the same equilibrium. Because it does not change any of the results that follow, we will not consider it further.

components raise the stakes for TSA-receiving teachers whose performance would determine their allowance. All else being equal, these enforceable commitment contracts increase the chance for a high effort equilibrium.[22]

6.1.3 Imprecise Measurements and Retaliations

How might the learning impact of SAM+Cam differ from SAM+Score? On the one hand, by tying its incentive contract solely to teacher absenteeism, SAM+Cam might lead teachers to neglect other aspects of their job that are important for learning (Holmstrom and Milgrom, 1991). SAM+Score avoided this problem by allowing the other indicators of the service agreement to affect teacher incentives. As this distortion 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 use of subjective indicators could lead to discrepancies between teacher and UC assessments of teacher effort. Macleod (2003) shows that such discrepancies would weaken the incentive effects of a performance-pay contract as the principal becomes more lenient to avoid ex-post conflict. Second, subjective indicators open room for negotiation, allowing teachers to pressure parents to increase their ratings. Anticipating teacher behavior, parents might exhibit leniency bias in their evaluation that will once again weaken the effectiveness of the incentive contract (Macleod, 2003; Marchegiani et al., 2016). SAM+Cam largely avoided these problems because the incentive was based on verifiable camera evidence, leaving little room for interpretation or negotiation.

We introduced three features into the model to study the potential effects of subjective indicators. First, we model the divergence of parent and teacher assessments of the 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 *underrates* the teacher (i.e., a low effort rating for high teacher effort), while π_o is the probability that the parent *overrates* the teacher.[23] More subjective indicators will have higher π_o's and/or π_u's. Second, we allow underrated teachers to retaliate. The exogenous utility cost of retaliation to parents is indicated by R. Teachers with a stronger bargaining power can inflict 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 in which parents can credibly commit to the assessment regime ex-ante, before effort levels are realized.[24] 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 (π_u), parents might overrate

[22] In theory, teachers could reduce their effort commitment in response to the higher stakes to compensate its effects. We do not think this is a major concern. For the presence indicators, the working hours and scoring rules are set and not open for negotiation. The other indicators are more subjective, with both the standard setting and rating being less well defined. In all cases, the SAM+Cam and SAM+Score treatments provide parents with an opportunity to affect teacher salaries — a feature that was not present in the SAM treatment.

[23] Of course, teacher measurement of their own effort could be biased upward. In such a case, we can define underrating (or overrating) as the divergence between parental assessments and what a teacher believes to be the effort level that they have provided as in Macleod (2003).

[24] 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.

(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, they could retaliate by imposing a utility cost R to parents.

Result 2. *When parents assess teacher effort leniently, teachers will never provide a high effort.*

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. The subgame perfect equilibrium (SPE) of the model yields a low teacher effort.

Result 3. *The probability that parents exert a higher effort is increasing in their probability of overrating teacher effort (π_o) and decreasing in their probability of underrating teacher effort (π_u). The probability that teachers exert effort is decreasing in the probabilities of inaccurate assessments, π_o and π_u.*

Result 4. *The probability that parents choose to assess leniently is increasing in their cost from teacher retaliation (R) and the probability that parents underrate teacher effort (π_u).*

The probability of inaccurate assessments affect efforts. When teachers exert low effort and parents assess teacher effort correctly ($\pi_o = 0$), parents will exert low effort. However, if parents overrate teacher effort ($\pi_o = 1$), then parents will exert high effort. Similarly, a higher probability to underrate the teacher's high effort (π_u) will lower parental efforts. In a strict assessment regime, positive π_o and π_u weakens the relation between teacher effort and punishment and thus the likelihood that teachers will exert high effort.

Retaliation could affect the likelihood that parents choose a strict assessment regime. To see the effect of retaliation, consider the case in which d_t and π_u/π_o allows for either (low, low) or $(high, high)$ 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 retaliation level R and a higher underrating probability 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. The use of subjective indicators weakens the relationship between the effort level and punishments for teachers while increasing the likelihood that teachers become disgruntled (from being underrated) and retaliate against parents. This would incentivize parents to monitor leniently, which would eliminate the incentive for teachers to exert high effort. If parents correctly assess teacher effort, they will also reduce their effort; otherwise, if they (mistakenly) overrate teacher effort, they might still exert high effort.

6.2 Teacher Effort

We begin with the treatment impacts on teacher presence and in-school activities. We use three TAS variables that were collected during unannounced visits (see Section 4.1), namely whether: (i) a teacher is present when they are 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.[25] Furthermore, since SAM+Cam and SAM+Score interventions incentivize TSA teachers, we estimated the heterogeneous impact of the TSA status on teacher effort.

Table 7 presents the one-year and two-year impacts. Since the SAM component of the treatments required all teachers in treatment schools (regardless of TSA status) to be evaluated by the UC, we first show the impacts across all teachers in Panel A. We find that the overall treatment effects ranged from negative to weak positive. The impact on teacher behaviors in SAM schools are negligible across all outcomes. Interestingly, after a year of implementation, SAM+Score reduced overall teacher presence and teacher observed to be working in school by 6.3 and 7.6 percentage points respectively.

Panel B presents the heterogeneity analysis by the teachers' TSA status to shed some light on these mean effects. There are four key findings. First, there were no differential treatment effects by TSA status in SAM schools, where all teachers regardless of TSA status received similar treatments. Second, SAM+Cam improved observed TSA-teacher effort the most (and most consistently across measures) after one year, while SAM+Score improved the least. Third, the weak one-year average treatment effects in the two SAM+ treatments were driven by non-TSA teachers, who reduced their effort. Finally, the positive effects on TSA teachers virtually disappeared by the second year.

We are also interested in whether the treatments led teachers to redirect their activities toward those that improve learning. We 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.[26] Once we identified these learning-enhancing activities, we estimated the impact of the interventions on the total hours that teachers spent on activities that were positively correlated with learning.

Table 8 presents the results. We first examine the impact of the treatments on the total time spent on school-related activities. Columns 1–4 show that the treatments had 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 led teachers to reallocate their time on school-related activities toward learning-enhancing activities by between 1.2–1.3 hours, around 8–8.5 percent out of a mean of 15.1 weekly hours. The strongest impact was once again observed for SAM+Cam. Column 7 suggests that the increase in the hours spent on learning-enhancing activities did not differ by the teachers' TSA status. However, similar to the impacts on the TAS outcomes, these positive improvements disappeared by the second year.

Are these behavioral changes driven by the service agreement between communities and teachers? To address this question, we explore the potential role of the scorecard on teacher behavior by examining the correlations between the different weights assigned to its indicators on teacher effort. We first code all scorecard indicators from all treatment schools into categories that could be mapped into the

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

[26] 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 percent are included in our 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.).

teacher effort variables in our survey instruments.[27] We then regress each teacher effort variable on the interactions between the treatments and the weights assigned to or scores obtained for the indicator variables.[28]

Table 9 presents the coefficients on the treatment interaction with the indicator weight (Panel A) and score (Panel B) regressors. The only indicator category for which both the weight and the score was associated with higher teacher effort in all treatments was the request for additional intra-curricular teaching. We show above that additional intra-curricular teaching is positively correlated with learning at baseline (see footnote 26). The association between additional intra-curricular teaching and its weight in the scorecard was weakest for SAM+Score schools. Meanwhile, the parent engagement indicator is associated with more parent meetings only in the SAM+Score schools.

6.3 Parental Investments in Education

An objective of the social accountability component is to encourage parents to be more involved in education. We therefore study how these treatments affect parents' financial and time investments in their children's education. We measure these investments using 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 parents' engagement with the school.

Table 10 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 28,000 (approximately USD 2) for SAM+Cam relative to the control-group mean of Rp 325,000 (USD 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. All treatments increased parents' willingness to forgo their children's contributions to the household economy: children's 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 children were accompanied when studying at home (column 7) and the number of meetings with teachers (column 9).

Most 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 children's likelihood of working disappeared.

6.4 Punishment Norms and Retaliation

Our model in Section 6.1 identifies two determinants of the relative effectiveness of our treatments in producing results. Result 2 highlights the importance of parents' credible commitment to a strict assessment regime to reach the equilibrium where both parents and teachers exert high effort. Result 3

[27] In Appendix Table B.2 shows how we mapped the indicators to the teacher effort variables.

[28] These are not causal impact estimates of the weight of the indicators, as these weights are endogenous. Nonetheless, we show in Appendix Table B.1 that there is no systematic relation between the inclusion of an indicator and the intervention status.

suggests that the use of more subjective indicators in the incentive contract increases the room for disagreement, which might lead to teacher retaliation. We provide empirical evidence for each of these results below.

The Role of Strict Punishment Norms. Parents' commitment to a strict assessment regime depends on whether they are willing to punish violations of an agreed standard. Different societies may exhibit different willingness to punish standard violations (Ensminger and Henrich, eds, 2014). Societies that are unwilling to punish may not be able to effectively use incentive contracts 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.[29] We conducted this experiment in 182 schools that were randomly selected from the 270 participating schools. Appendix E 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 norms.

We find that the punishment norm plays an important role in the short-run effectiveness of the interventions. Table 11 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 SAM+Score.[30] However, column 2 suggests that these heterogeneous impacts on teacher presence did not persist.[31] Columns 3–4 show the lack of heterogeneous impact of having above-median punishment norms on the presence of non-TSA teachers and serve as a placebo check. We also find that learning improvements for the SAM+ interventions were primarily driven by communities with stronger punishment norms and this differential impact persisted for SAM+Cam (columns 5–6).

Subjective Performance Measures and Retaliations. 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 exert high effort. We have evidence that subjectivity in the assessment created tensions between teachers and the UC. A qualitative study in our treated schools suggest that teachers in SAM+Score often questioned the UC evaluations, as the UC members typically were less educated than these teachers (World Bank, 2020). The teachers' higher social status in the community put them

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

[30] In Appendix Table G.6, 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 those on teacher presence.

[31] We discuss in Section 7.1 that by the second year, school principals appeared to be more accommodating of teacher absence by providing excuses that do not penalize teacher allowance in the second year. We think this might explain why the community's willingness to punish no longer predicts teacher presence in the second year.

in a position to pressure UC members to improve their score. We find corroborating evidence from our survey of UC members: Table 12 shows that UC members in SAM+Score schools are more likely to be pressured to increase the evaluation scores and received more threats regarding a low score than those in the other treated schools.

6.5 External and Internal School Management

The credibility of the assessment regime becomes stronger if external and internal school management aligns their actions with monitoring outcomes of the UC. 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 to a more systematic way of monitoring and evaluating their teachers. Tables 13 and 14 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 13 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 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.

We also find that our interventions had persistent effects on how school principals evaluated teachers (Table 14). By the first year, all three interventions led to increases in the share of teachers who received any or routine supervision, 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 other in both 2018 and 2019.

7 Sustainability

In their seminal paper on community-based health programs, Shediac-Rizkallah and Bone (1998) identify three key categories of operational indicators to monitor sustainability. These categories are: (i) the persistence of benefits after the initial program completed; (ii) the institutionalization of project activities within an organization structure; and (iii) the recipient communities' continued capacity to execute their roles. Our research design offers insights into the sustainability question. We have presented evidence for the persistence of impacts beyond the first year. The following sections will discuss some evidence on the remaining two categories, followed by a section on cost effectiveness.

7.1 Institutionalization

The project planned for institutional sustainability. In the beginning, external facilitators helped set up the key implementing institutions at the village level, such as the UC and the evaluation meetings. Once they were setup, the key activity that needed to be maintained was the monthly evaluation meetings. To prepare for their eventual departure, external facilitators also trained a village cadre to perform this activity. In our implementation, once the village-level facilitator support expired after the first year, meeting facilitation was fully managed by the village cadres except in one private school that no longer had a TSA teacher. The UCs in all villages continued to send monitoring reports to the district officials; in SAM+Cam and SAM+Score villages, these reports continued to determine cuts to the TSA.[32]

However, even though the formal monitoring continued, it appears that school principals might have undermined their teachers' presence-based contract by providing them with excuses that would minimize penalties from their absences (Appendix Table G.7). Panel A shows that in the first year, school principals in treatment schools were not more likely to issue "off-school assignments" — the type of excuse that was the least scrutinized and would not result in a penalty in the SAM+ treatments — as an excuse for teacher absence. However, in the second year, this type of excuse was 20 percentage points more likely to be issued (relative to a control mean of 28 percentage points) in SAM and SAM+Cam schools. This result is neither unique nor surprising: for example, a study of Indian nurse working in public health facilities similarly finds that the administration was allowing them to claim more "exempt days" (Banerjee et al., 2008). This result nonetheless suggests that policymakers need to anticipate possible attempts by service providers to render the conditionalities ineffective.

7.2 Local Support for Reform

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 present the treatment impacts on parents' satisfaction of their children's learning and school quality, and on parental aspirations 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 15 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

[32]The institutionalization at the village level was reinforced/ enabled by the MoEC decrees, and supported by district level decrees with associated funding. There was also project coordinators at the district and national levels, so some level of external accountability was still happening.

from district education officials and other villagers. For SAM and SAM+Cam, this increased satisfaction of outside appreciation persisted into the second year. Panel B suggests that there was little differential satisfaction of outside appreciation between TSA and non-TSA teachers across all treatments.

Columns 5–6 of Panel A suggest that after one year, all three interventions improved teacher satisfaction of their salary. The heterogeneous impact analysis in Panel B suggests that the one-year impacts on salary satisfaction for the two SAM+ interventions were positive for the TSA teachers, albeit more muted for SAM+Cam. On the other hand, these treatments elicited a much stronger salary satisfaction responses for non-TSA teachers, even though their salaries were unaffected by our treatments. The overall and differential effects on salary satisfaction for SAM and SAM+Cam persisted well into the second year.

Similarly, column 7 of Panel A shows that overall job satisfaction increased across all treatments after one year. Panel B shows that these increases in job satisfaction were stronger for non-TSA teachers, especially in the two performance-pay treatments. Between the two performance-pay treatments, the differential impact was much stronger for SAM+Cam. We find a weak and statistically insignificant impact for the TSA teachers. Nonetheless, by the second year, the positive treatment impacts on job satisfaction had completely disappeared.

Overall, these results alleviate concerns that performance-pay schemes would lead to widespread dissatisfaction among affected teachers.[33] 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 conditions might have made allowances seem fairer to non-recipients.

Parent Satisfaction and Educational Aspirations Table 16 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 percentage points after one year, and these positive improvements persisted into the second year for SAM and SAM+Cam.

However, the immediate impacts on the parental view of the school quality did not immediately translate into satisfaction with 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 with 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

[33]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.

these effects persisted for SAM and SAM+Cam: in the second year, the effects of these two interventions remained positive, but were smaller.

Village Support. Over the course of the implementation, the village governments of the treatment schools gave small contributions to support the UCs. In 2017, 168 treatment schools received some support from the village budget; by 2019, all 203 treatment schools did so. The amount of funding provided by the village government varied widely. District averages of the annual amount allocated by each village ranged from IDR 1.471 million (USD 104) in Sintang to IDR 9.022 million (USD 646) in East Manggarai.[34] It also varied widely within each district: in Sintang, for example, the annual support ranged from IDR 750,000 (USD 54) to IDR 6.4 million (USD 460).

7.3 Cost Effectiveness.

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

Although these interventions were implemented in remote areas, their costs were on par with similar interventions that had been rigorously evaluated (see Glewwe and Muralidharan, 2016; JPAL, 2019).[36] For SAM, the most comparable study is that of Pradhan et al. (2014) in Indonesia. Its most successful intervention, which strengthened school committees through a combination of democratic elections of members and facilitating joint planning with the village council, cost USD 7.50 for 0.1 sd improvement in learning.[37] Three studies on conditional cash transfers improved learning outcomes with costs averaging USD 77 per 0.1 sd increase. For performance pay interventions, camera monitoring and teacher-presence-based payment in India costs USD 44 per 0.1 sd increase, excluding the cost of staff, transportation, and monthly meetings. A teacher incentive intervention in Kenya costs USD 16 per 0.1 sd increase, while another in India costs USD 1 per 0.1 sd increase.

8 Conclusion

This study investigates whether adding incentive payments based on community monitoring reports from a social accountability intervention could overcome some of the shortcomings of participatory programs noted in earlier studies. Our findings suggest that not all performance-pay contracts improve the

[34] Cost figures in Indonesian 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.

[35] For the exchange rate used, see Footnote 34.

[36] To make our cost figure comparable to those reported in Glewwe and Muralidharan (2016) and JPAL (2019), we converted our cost to 2011 US dollars using US GDP deflators from 2011 and 2017.

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

effectiveness of a social accountability intervention. 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 that is more prone to subjectivity.

Our finding answers an important question that arises in many labor contracts on the use of subjective versus objective performance measures (Baker et al., 1994; Khan et al., 2016). Note that this intervention was implemented in a context in which the teacher (agent) has a higher social status and is more knowledgeable about education than members of the community (principal). We think this is one of the reasons an incentive contract based on presence worked best. With a simple verifiable indicator, SAM+Cam made both community members and teachers feel comfortable with their assigned roles and minimized the risk of disagreement over the rating. This simple contract led to less divergences in the performances of TSA versus non-TSA teachers (suggesting that teachers might perceive it to be fairer) and less conflict at the community level.

This study also shows that teachers and other stakeholders in the community accept performance pay. There was no large pushback 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 results indicate that impacts weakened after the facilitators left the village. While monitoring reports continued to be gathered, SAM did not have sustained impacts on learning and the incremental effect of SAM+Cam slightly weakened. This largely seems to be due to dissipating teacher efforts by the second year. 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 lower teacher effort.

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. Periodic but infrequent external visits can help energize stakeholders, close loopholes, and signal that local policy makers are taking the program seriously. This raises the question of the extent to which 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 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.

Akresh, Richard, Daniel Halim, and Marieke Kleemans, "Long-term and Intergenerational Effects of Education: Evidence from School Construction in Indonesia," Technical Report w25265, National Bureau of Economic Research, Cambridge, MA November 2018.

Alesina, Alberto, Sebastian Hohmann, Stelios Michalopoulos, and Elias Papaioannou, "Intergenerational Mobility in Africa," *Econometrica*, 2021, *89* (1), 1–35.

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

Andrabi, T. and C. Brown, "Subjective versus Objective Incentives and Teacher Productivity," *mimeo*, 2021.

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, Abhijit V, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani, "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," *American Economic Journal: Economic Policy*, February 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.

Barrera-Osorio, Felipe, Paul Gertler, Nozomi Nakajima, and Harry Patrinos, "Promoting Parental Involvement in Schools: Evidence From Two Randomized Experiments," Technical Report w28040, National Bureau of Economic Research, Cambridge, MA October 2020.

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.

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.

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.

Chetty, Raj and Nathaniel Hendren, "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*," *The Quarterly Journal of Economics*, August 2018, *133* (3), 1107–1162.

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, Pascaline Dupas, and Michael Kremer, "School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools," *Journal of Public Economics*, March 2015, *123*, 92–110.

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.

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.

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.

Laliberté, Jean-William, "Long-Term Contextual Effects in Education: Schools and Neighborhoods," *American Economic Journal: Economic Policy*, May 2021, *13* (2), 336–377.

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*, 69–83.

Lumbanraja, Sharon Kathy and Indah Ayu Prameswari, *Student Learning Assessment : A Tool to Measure Primary Grade Student Learning Outcomes in Indonesia's Remote Areas - Background Paper*, Washington DC: World Bank Group, March 2021.

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.

Marliyanti, Usha Riyanto, and Dewi Susanti, *From Facilitation to Participation: Community Empowerment to Improve Education in Remote Areas of Indonesia.*, Jakarta: World Bank, 2022.

McEwan, Patrick J., "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments," *Review of Educational Research*, September 2015, *85* (3), 353–394.

Muralidharan, K., "Field Experiments in Education in Developing Countries," in "Handbook of Economic Field Experiments," Vol. 2, Elsevier, 2017, pp. 323–385.

Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal, "The fiscal cost of weak governance: Evidence from teacher absence in India," *Journal of Public Economics*, January 2017, *145*, 116–135.

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.

Prendergast, Canice and Robert Topel, "Discretion and bias in performance evaluation," *European Economic Review*, April 1993, *37* (2-3), 355–365.

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.

Shediac-Rizkallah, M. C. and L. R. Bone, "Planning for the sustainability of community-based health programs: conceptual frameworks and future directions for research, practice and policy," *Health Education Research*, March 1998, *13* (1), 87–108.

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.

van der Weide, Roy, Christoph Lakner, Daniel Gerszon Mahler, Ambar Narayan, and Rakesh Ramasubbaiah, "Intergenerational Mobility around the World.," Technical Report 9707, World Bank, Washington DC 2021.

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 Educa-

tion 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
Received TSA in 2017	0.62	0.48	2297
Share of teachers observed to be ...:			
...present	0.80	0.40	2212
...working	0.74	0.44	2212
...teaching (when scheduled)	0.73	0.44	1688
Self-reported weekly hours spent on:			
...preparing lessons	5.27	5.33	1796
...teaching curricular materials	20.14	8.78	1796
...assessing student work	4.08	3.91	1796
...teaching extra-curricular materials	1.25	1.92	1796
	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

	Mean Score		Mean Score (Grades 3–6)†		Grade Repetition	
	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)
SAM	0.084	0.028	0.097	0.029	0.010	-0.000
	(0.036)**	(0.032)	(0.035)***	(0.032)	(0.010)	(0.008)
SAM+Cam	0.198	0.133	0.168	0.137	0.004	0.014
	(0.036)***	(0.034)***	(0.034)***	(0.034)***	(0.010)	(0.008)
SAM+Score	0.110		0.095		0.009	
	(0.033)***		(0.032)***		(0.010)	
Control group mean					0.08	0.04
Control group raw-score:						
Mean	47.08	41.08	47.97	40.63		
Standard deviation	18.86	19.66	19.12	19.73		
Test of equality (P-val)						
SAM v. SAM+Cam	0.003	0.002	0.051	0.001	0.565	0.117
SAM+Cam v. SAM+Score	0.018		0.041		0.614	
SAM v. SAM+Score	0.474		0.969		0.963	
Randomization Inference (P-value, N = 10)						
SAM	0.000	0.300	0.100	0.400	0.200	1.000
SAM+Cam	0.000	0.000	0.000	0.000	0.700	0.100
SAM+Score	0.000		0.000		0.500	
R2	0.390	0.192	0.475	0.192	0.139	0.073
Observations	31022	15611	21448	15108	24719	13257
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. 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: 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 5: Differential Impacts on Student Learning

	Male		Years with TSA teachers		Above-median student				Above-median school (Grades 3–6)†	
					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	0.101	0.081
	(0.039)	(0.038)	(0.052)	(0.065)	(0.040)*	(0.038)	(0.040)	(0.039)	(0.053)*	(0.047)*
SAM+Cam	0.180	0.120	0.176	0.128	0.130	0.104	0.130	0.107	0.146	0.197
	(0.040)***	(0.038)***	(0.055)***	(0.063)**	(0.041)***	(0.040)***	(0.044)***	(0.044)**	(0.055)***	(0.056)***
SAM+Score	0.091		0.098		0.066		0.079		0.105	
	(0.038)**		(0.051)*		(0.038)*		(0.042)*		(0.048)**	
Covariate: [...]	-0.152	-0.212	0.000	-0.008	0.109	0.108	0.155	0.095	0.062	0.194
	(0.021)***	(0.027)***	(0.024)	(0.018)	(0.027)***	(0.031)***	(0.033)***	(0.032)***	(0.064)	(0.068)***
... × SAM	0.034	0.011	0.000	0.006	0.012	0.050	0.031	0.055	-0.029	-0.112
	(0.029)	(0.037)	(0.035)	(0.033)	(0.035)	(0.035)	(0.047)	(0.044)	(0.073)	(0.061)*
... × SAM+Cam	0.012	0.020	0.009	0.002	0.071	0.056	0.064	0.045	0.016	-0.137
	(0.029)	(0.037)	(0.038)	(0.030)	(0.034)**	(0.039)	(0.044)	(0.042)	(0.077)	(0.075)*
... × SAM+Score	0.018		0.002		0.023		-0.008		-0.045	
	(0.030)		(0.032)		(0.035)		(0.043)		(0.072)	
Observations	31022	15297	31022	15297	24700	13655	24700	13655	21448	14773
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: †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 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	0.019	-0.019	0.039	0.005	0.074	0.048
	(0.034)	(0.037)	(0.039)	(0.042)	(0.045)*	(0.044)
SAM+Cam	0.050	-0.015	0.085	-0.023	0.083	0.008
	(0.033)	(0.036)	(0.038)**	(0.040)	(0.049)*	(0.046)
SAM+Score	-0.017		-0.021		0.067	
	(0.036)		(0.040)		(0.046)	
Non-TSA-recipient	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	-0.880	-0.108	1.218	0.302	0.974	0.066
	(0.817)	(0.828)	(1.001)	(1.007)	(0.449)***	(0.493)	(0.550)*	(0.601)
SAM+Cam	0.190	0.543	-0.140	0.036	1.305	0.073	1.063	-0.029
	(0.812)	(0.827)	(0.987)	(0.997)	(0.446)***	(0.491)	(0.543)*	(0.595)
SAM+Score	-0.116		0.182		0.553		0.621	
	(0.822)		(1.005)		(0.451)		(0.552)	
Non-TSA-recipient			-2.492	-2.549			-1.696	-0.997
			(1.303)*	(1.323)*			(0.718)**	(0.792)
$\ldots \times$ SAM			0.907	1.422			0.631	0.628
			(1.688)	(1.702)			(0.929)	(1.019)
$\ldots \times$ SAM+Cam			0.859	1.276			0.638	0.177
			(1.731)	(1.767)			(0.952)	(1.057)
$\ldots \times$ 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: Weighted Presence Indicators and Teacher Efforts

	Presence	Teach in Class	Self-reported Hours					Number of Parent Meetings
			Prepare	Teach	Assess	Additional Intra-Cur. Lessons	Extra-Curricular Activities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A. Indicator Weights							
Indicator Weight								
... × SAM	-0.110	-0.031	4.182	-2.556	-2.407	5.152	0.868	0.247
	(0.247)	(0.401)	(5.785)	(2.907)	(1.358)*	(1.360)***	(0.962)	(2.916)
... × SAM+Cam	-0.640	-0.083	-12.123	-6.650	0.034	4.035	-0.066	-0.836
	(0.233)***	(0.537)	(7.857)	(5.039)	(1.226)	(1.321)***	(1.087)	(2.639)
... × SAM+Score	-0.056	-0.176	-17.374	-0.362	0.340	1.985	0.131	9.965
	(0.222)	(0.360)	(7.056)**	(5.451)	(2.150)	(1.091)*	(1.174)	(2.705)***
Observations	1140	1140	1150	1150	1150	1150	1150	3354
	Panel B. Indicator Scores							
Indicator Score								
... × SAM	-0.001	0.006	-0.005	-0.024	-0.025	0.045	0.021	-0.007
	(0.004)	(0.003)**	(0.057)	(0.039)	(0.013)*	(0.014)***	(0.012)*	(0.024)
... × SAM+Cam	-0.005	0.002	-0.066	-0.049	-0.006	0.039	-0.004	0.010
	(0.003)*	(0.006)	(0.064)	(0.056)	(0.015)	(0.014)***	(0.010)	(0.025)
... × SAM+Score	-0.002	-0.000	-0.157	-0.014	-0.010	0.040	0.009	0.102
	(0.003)	(0.003)	(0.051)***	(0.048)	(0.022)	(0.011)***	(0.009)	(0.028)***
Observations	1024	1024	1029	1029	1029	1029	1029	2993
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The outcome variables for column 1–2 are from the Teacher Absence Survey (TAS); columns 3–7 are from the teacher survey; and column 8 is from the parent survey. The regressors in Panel A are the interactions of the treatment group indicators with the weights assigned to the scorecard indicator related to the various teacher effort categories. The regressors in Panel B are the interactions of the treatment group indicators with the scores given to indicators in January 2018 (before the outcome variables were collected) for the various teacher effort categories. All regressions include treatment group dummy variables. Teacher regressions (columns 1–7) control for teacher age and sex. Standard errors clustered at the school level. */**/*** denotes 10/5/1 percent significance levels.

Table 10: 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 11: 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 12: User Committee Reports of Pressure from School

	Intimidated	Pressure to Increase Score	Threats for Low Score
	(1)	(2)	(3)
SAM+Cam	0.021	-0.005	0.071
	(0.040)	(0.056)	(0.044)
SAM+Score	0.066	0.119	0.165
	(0.041)	(0.056)**	(0.044)***
Constant	0.107	0.035	-0.038
	(0.083)	(0.114)	(0.090)
Control group mean	0.030	0.075	0.000
Observations	201	201	201
School-level controls	Yes	Yes	Yes
Strata FE	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 13: 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 14: 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 15: 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	0.204	0.484	0.286	0.240	0.024	0.056	0.049	-0.019
	(0.150)	(0.157)***	(0.130)**	(0.142)*	(0.151)	(0.158)	(0.054)	(0.060)
SAM+Cam	0.309	0.502	0.380	0.469	0.202	0.321	0.049	-0.040
	(0.148)**	(0.156)***	(0.129)***	(0.141)***	(0.149)	(0.157)**	(0.053)	(0.059)
SAM+Score	0.433		0.527		0.538		0.052	
	(0.151)***		(0.131)***		(0.152)***		(0.054)	
Non-TSA-recipient	-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 16: 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.

Figures

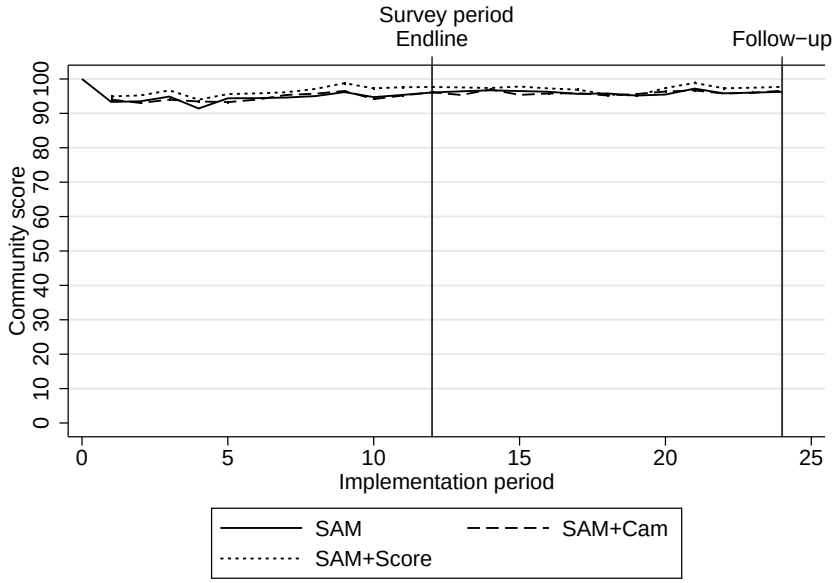
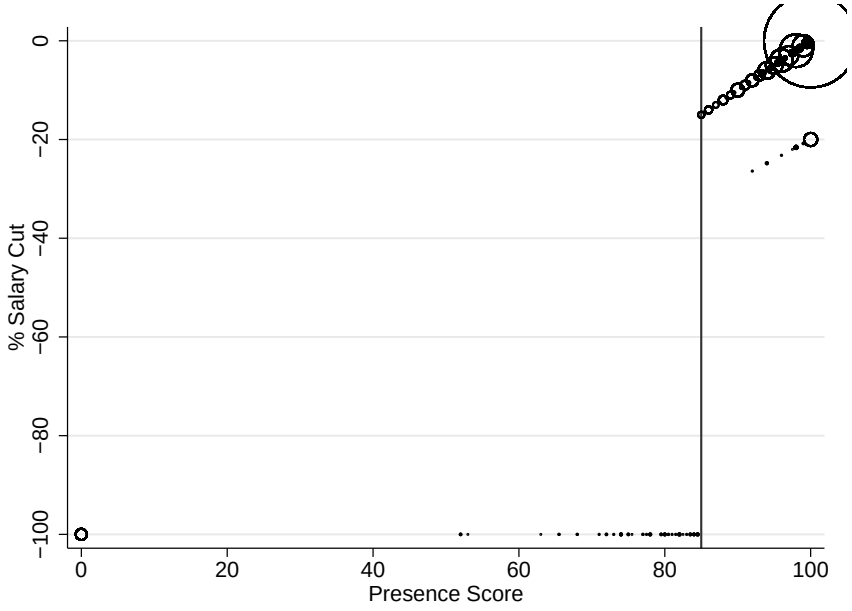


Figure 2: Average Scorecard Ratings Across Treatments



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

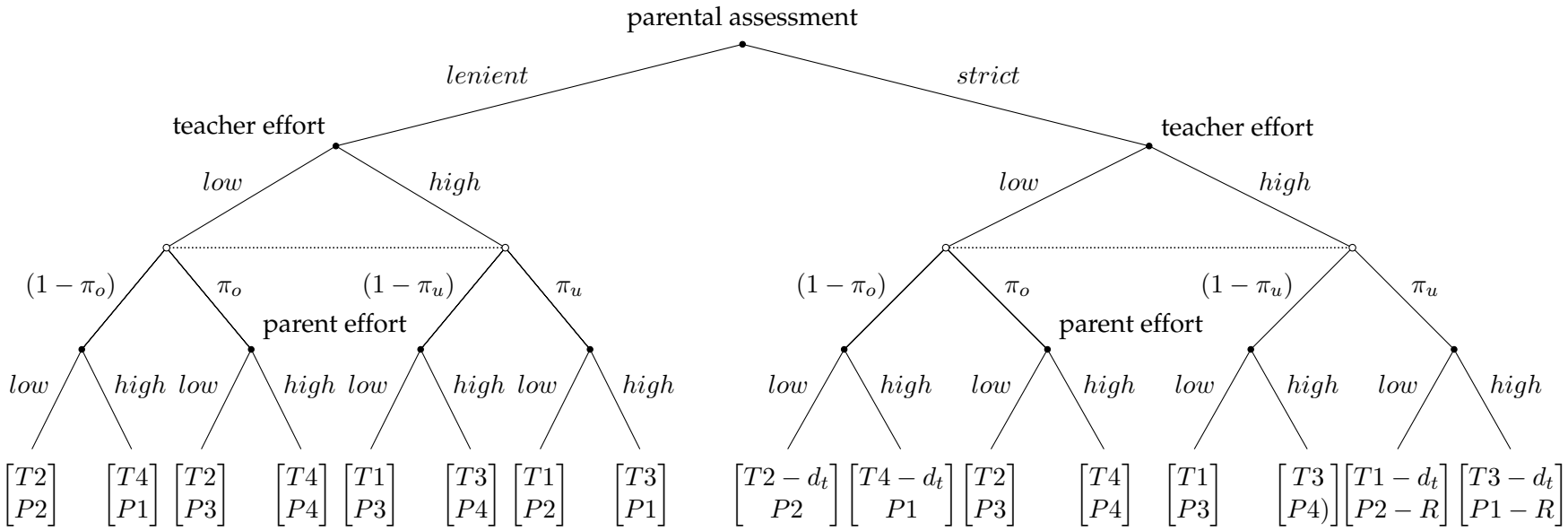


Figure 4: The Strictness of Parental Assessment, Imprecise Measurement, and Teacher Retaliation

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 Implementation Details

I The Social Accountability Intervention

The set of seven meetings to set up the service agreement and the user committee (UC) was 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. The meetings to formulate the service agreement and teacher-specific scorecards took the longest time. In 40 percent of the schools, this process took a single meeting of between three to seven hours; in the remaining schools, it required two or more meetings. 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. Appendix Figure B.1 shows an example of 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. Following a UC meeting to revise these indicators, the share of indicators oriented toward the student learning process increased from 33 to 48 percent.[1] At the same time, we find the UCs were most likely to drop the corporal punishment indicator because teachers felt that it was too difficult to implement.[2] The meetings to revise the scorecards took place in all treatment schools between July 2017 and January 2018.

Did the treatments affect the indicator choices? To answer this question, we assigned the revised indicators into 12 categories and coded whether each had a frequency (e.g., twice a week) or time requirement (e.g., for fifteen minutes) specified. Panel A of Table B.1 presents the regression of the weight (i.e., maximum score in a scorecard) assigned to the indicator class and the treatment. The coefficients on the constant in Panel A of Table B.1 indicate the average weights assigned to the indicators classes. We find that after the presence indicator, indicators related to pedagogy (0.24), assessment (0.14), and additional intra-curricular classes (0.10) received the highest weight. However, Panel A suggests that there is no correlation between the weight assigned to the indicator class and the treatment.

Panel B of Table B.1 presents a similar regression, but instead of regressing the indicator weight, we regress the interaction of the indicator weight and a dummy variable of whether the indicator included a time or frequency requirement. Here, we find that the indicators in the two SAM+ treatments are less likely to specify time or frequency requirements. This result suggests that the incentive treatments reduced the specificity in the teachers' commitment to improve learning.

We also investigate whether the weight assigned to an indicator in the service agreement varies by the baseline performance of the teacher. To this end, we matched the indicator classes to performance measures included in the baseline survey. Appendix Table B.2 describes how we mapped the indicators to the teacher effort variables. Table B.3 shows little indication of a systematic relationship between the indicators chosen and their associated teacher effort variable at baseline.

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

[1] 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.

[2] 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.

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 B.1: A Sample of the Community Scorecard

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 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.[3] In 2017, 83 percent of the treatment schools received some funding from their village government to provide operational costs for monthly meetings and incentives for the village cadres and UC members. By 2019, all treatment schools received operational funding from their village government.

II 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 first had their TSA incentivized in April 2017. The remaining 22 schools were affected by the holdup and held their first meeting between June and July 2017. By October 2017, all 135 schools had their TSA incentivized. Second, due to the end-of-year budgetary account closure, TSAs for the second half of November and December were paid in full irrespective of the scorecard.

We find clear evidence that the scorecard determined 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).

[3]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.

Table B.1: Weighted Indicators By Treatment

	Present	Teach in Class	Prepare	Assess	Additional Intra-cur	Extra-curricular	Pedagogy	Others	Parent Engagement	Student Presence	Freq/ time specified
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Panel A. Without frequency/time specified										
SAM+Cam	-0.010	-0.011	-0.008	0.002	0.014	-0.011	0.028	0.001	0.007	-0.008	
	(0.012)	(0.010)	(0.005)	(0.015)	(0.015)	(0.010)	(0.021)	(0.003)	(0.009)	(0.004)*	
SAM+Score	-0.015	0.009	-0.004	-0.012	0.009	-0.016	-0.004	0.007	0.002	-0.004	
	(0.013)	(0.012)	(0.005)	(0.014)	(0.015)	(0.009)*	(0.019)	(0.003)**	(0.009)	(0.004)	
Constant	0.294	0.039	0.014	0.144	0.098	0.046	0.243	0.002	0.032	0.012	
	(0.009)***	(0.007)***	(0.004)***	(0.010)***	(0.011)***	(0.007)***	(0.014)***	(0.002)	(0.006)***	(0.003)***	
	Panel B. With frequency/time specified										
SAM+Cam	-0.007	-0.007	-0.006	-0.018	-0.028	-0.002	-0.008	-0.000	-0.005	-0.007	-0.069
	(0.011)	(0.004)*	(0.003)*	(0.012)	(0.012)**	(0.008)	(0.013)	(0.002)	(0.007)	(0.003)**	(0.026)***
SAM+Score	-0.018	0.001	-0.006	-0.009	-0.004	-0.004	-0.026	0.001	-0.002	-0.004	-0.060
	(0.011)*	(0.004)	(0.003)*	(0.011)	(0.013)	(0.007)	(0.013)**	(0.002)	(0.008)	(0.003)	(0.024)**
Constant	0.272	0.009	0.007	0.082	0.063	0.025	0.091	0.002	0.023	0.009	0.577
	(0.007)***	(0.003)***	(0.003)**	(0.008)***	(0.009)***	(0.005)***	(0.009)***	(0.001)	(0.005)***	(0.002)***	(0.018)***
Observations	1287	1287	1287	1287	1287	1287	1287	1287	1287	1287	1287
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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

Table B.2: Mapping from Indicator to Teacher Effort Variables

Indicator		Teacher Effort	
Variable Name	Description	Variable Name	Data Source
Present	Teacher is present in school during office hours	Teacher presence	TAS
Teach in class	Teacher is in class and teaching	Teacher teach in class	TAS
		Hours spent teaching	Teacher survey
Prepare	Teacher prepares/provides/completes a lesson plan	Preparation hours	Teacher survey
Assess	Teacher gives/corrects/grades in-school or take-home assignments	Assessment hours	Teacher survey
Additional Intra-curricular	Teacher provides additional lessons/ enrichment/tutoring of intra-curricular materials	Hours for additional lessons	Teacher survey
Extra-curricular	Extra-curricular activities (e.g., going to the library, eagle scouts, art)	Extracurricular hours	Teacher survey
Pedagogy	Implement certain teaching methods or use teaching props; ensure that children understand the materials	-	-
Parent Engagement	Teacher conducts regular meetings with parents; communicates with parents if students were absent; communicate student progress through a student book	Number of teacher-parent meetings	Parent survey
Student Presence	Student is present	-	-

Notes: TAS = Teacher absence survey. All teacher hours variables are self reported.

Table B.3: Weighted Indicators and Baseline Teacher Efforts

	Present in School (1)	Teach in Class (2)	Prep (3)	Assess-ment (4)	Extra Intra-cur. (5)	Extra-cur. (6)	Parent engagement (7)
Baseline Effort	0.012	0.002	0.000	-0.001	0.003	-0.001	-0.001
	(0.012)	(0.008)	(0.000)	(0.002)	(0.002)*	(0.001)	(0.001)*
... × SAM+Cam	-0.024	0.003	0.000	-0.001	-0.003	0.004	0.001
	(0.017)	(0.010)	(0.000)	(0.002)	(0.002)	(0.002)	(0.001)
... × SAM+Score	-0.021	-0.014	-0.001	-0.000	-0.001	0.000	0.001
	(0.017)	(0.011)	(0.001)	(0.002)	(0.003)	(0.002)	(0.001)*
Constant	0.267	0.020	0.061	0.080	0.069	-0.036	0.032
	(0.024)***	(0.019)	(0.029)**	(0.039)**	(0.025)***	(0.055)	(0.006)***
... × SAM+Cam	0.007	-0.011	-0.008	0.013	0.023	-0.014	0.008
	(0.020)	(0.012)	(0.006)	(0.018)	(0.015)	(0.011)	(0.009)
... × SAM+Score	-0.001	0.019	-0.001	-0.006	0.023	-0.018	0.003
	(0.021)	(0.015)	(0.007)	(0.017)	(0.015)	(0.010)*	(0.009)
Observations	952	952	874	874	874	874	2695
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Baseline variables on whether the teacher was present or teaching in class (columns 1 and 2) were obtained from the TAS. Baseline hours that teacher spent to prepare, assess, provide additional intra-curricular lessons, and implement extracurricular activities were calculated from the self-reported teacher hours from the teacher survey (columns 3–6). The number of meetings between parents and teachers (column 6) was calculated from the maxima number of meetings on various topics from the parent survey data. Except for column 6, all regressions controlled for teacher age, sex, highest level of education, and a dummy for missing the education variable. Standard errors clustered at the school level. */**/*** denotes significant at the 10/5/1 percent significance levels.

C 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, and simple utility functions that capture how teachers and parents weigh the benefit from student learning against the utility cost of putting in effort. We derive the parameter restrictions on the model and provide an interpretation.

Assume a linear production of learning:

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

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:

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

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

Table 6a indicates the following ordering of the payoffs:

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

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

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

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

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

which indicates that the cost of effort for teachers is larger than the marginal utility gains of learning even if parents put in effort, while for parents the utility cost of effort is within the bounds of the marginal utility gains of effort with and without teacher effort.

D 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:

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

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

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

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

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

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 D.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 D.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 D.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

E 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.[4]

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.

[4] 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.

F Efficiency Analysis

We provide the cost breakdown to implement *KIAT Guru* ini Table F.1. The table includes all direct costs to set up and maintain the SAM intervention in 203 schools with an average of 132 students per school. Panel A presents the one-time cost to set up the SAM institutions in the villages in the first year. The project spent USD 1,026,759 to cover the facilitator salaries, transportation costs, trainings, and other costs associated with the initial set of meetings. The annualized cost to train 41 facilitators was USD 143,889. The facilitators were hired for 15 months for a total cost of USD 501,483. Each facilitator visited a school with an average of 11 visits with an average transport cost of USD 112 per visit. In addition, village-level meetings to set up the service agreement and user committee at the village level (which included a set of seven meetings) cost USD 633 per school (for a total of USD 128,499). The one-time cost to set up SAM was USD 40 per student. However, there are additional costs for the SAM+Cam intervention to purchase of one smart phone per school and salaries for two personnel to develop and maintain the application.[5] When annualized, this cost came down to USD 33,422 — or equal to an additional USD 4 per student for the SAM+Cam treatment.

Panel B presents the annualized marginal cost to maintain the interventions. The average annual cost to sustain SAM was USD 2,182 per school (USD 17 per student), with an additional USD 492 per school (USD 4 per student) for SAM+Cam schools. This cost covers an annual refresher training, monthly meetings, and evaluation meeting. For each school, the annual costs for the refresher training and monthly meetings were USD 709 and USD 834 respectively. In addition, at the end of each semester, the UC and school providers reviewed the content of the service agreement and the community scorecard indicators, and reappointed the UC members in an evaluation meeting. The annual per school cost of the evaluation meetings was USD 639.

Table F.1: Implementation Costs of *KIAT Guru*

	Panel A. One-time Investment Cost		
	Total Cost	Cost per School	Cost per Student
	(1)	(2)	(3)
Training	143,889	709	6
Salary	501,483	2,470	19
Transport	252,888	1,246	10
Initial Meetings	128,499	633	5
Total	1,026,759	5,058	40
SAM+Cam Additional Cost	33,422	492	4
	Panel B. Annual Maintenance Cost		
Refresher Training	143,927	709	5
Monthly Meetings	169,302	834	7
Evaluation Meetings	129,717	639	5
Total	442,946	2,182	17
SAM+Cam Additional Cost	33,422	492	4

[5]The total additional cost for SAM+Cam was USD 86,341 over the 31-month implementation period (between November 2016 and May 2019).

G Appendix Tables and Figures

I Tables

Table G.1: 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 G.2: 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)
Received TSA in 2017	0.59	0.65	0.63	0.61	0.06	0.04	0.02	-0.02	-0.04	-0.02
	(0.49)	(0.48)	(0.48)	(0.49)	(0.10)	(0.31)	(0.64)	(0.58)	(0.24)	(0.56)
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 (when scheduled)	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 G.3: Impact on Student Learning Outcomes: No Individual Controls

	Mean Score		Mean Score (Grades 3–6)		Grade Repetition	
	2018 (1)	2019 (2)	2018 (3)	2019 (4)	2018 (5)	2019 (6)
SAM	0.066	-0.004	0.079	-0.001	0.013	-0.000
	(0.040)*	(0.049)	(0.037)**	(0.048)	(0.010)	(0.008)
SAM+Cam	0.190	0.135	0.157	0.142	0.005	0.013
	(0.044)***	(0.049)***	(0.039)***	(0.049)***	(0.010)	(0.008)
SAM+Score	0.085		0.071		0.012	
	(0.038)**		(0.035)**		(0.010)	
Control group mean					0.08	0.04
Control group raw-score:						
Mean	47.08	41.08	47.97	40.63		
Standard deviation	18.86	19.66	19.12	19.73		
Test of equality (P-val)						
SAM v. SAM+Cam	0.006	0.006	0.056	0.005	0.386	0.131
SAM+Cam v. SAM+Score	0.016		0.032		0.477	
SAM v. SAM+Score	0.652		0.827		0.881	
Randomization Inference (P-value, N = 10)						
SAM	0.100	0.900	0.100	0.900	0.400	1.000
SAM+Cam	0.000	0.000	0.000	0.000	0.700	0.300
SAM+Score	0.000		0.000		0.300	
R2	0.355	0.116	0.450	0.118	0.106	0.059
Observations	31022	15611	21448	15108	24719	13257
Individual controls	No	No	No	No	No	No
Strata FE	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 G.4: Impact on Student Learning Outcomes By Subject

	All Sample				Grades 3–6†			
	Indonesian		Mathematics		Indonesian		Mathematics	
	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	0.096	0.009	0.089	0.048
	(0.037)**	(0.027)	(0.040)*	(0.044)	(0.035)***	(0.027)	(0.042)**	(0.045)
SAM+Cam	0.190	0.096	0.202	0.176	0.150	0.096	0.182	0.183
	(0.036)***	(0.028)***	(0.041)***	(0.046)***	(0.035)***	(0.028)***	(0.041)***	(0.047)***
SAM+Score	0.122		0.094		0.100		0.082	
	(0.034)***		(0.038)**		(0.034)***		(0.038)**	
Control group raw-score:								
Mean	47.13	38.12	47.03	44.04	49.17	37.63	46.78	43.63
Standard deviation	21.80	26.95	20.41	19.69	22.23	27.38	20.31	19.47
Test of equality (P-val)								
SAM v. SAM+Cam	0.013	0.003	0.003	0.003	0.156	0.002	0.033	0.003
SAM+Cam v. SAM+Score	0.061		0.013		0.186		0.018	
SAM v. SAM+Score	0.447		0.602		0.924		0.863	
Observations	31022	15611	31022	15611	21448	15108	21448	15108
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. †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 G.5: 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 G.6: 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)
... × SAM	0.098	0.025	0.197	-0.070
	(0.086)	(0.133)	(0.096)**	(0.144)
... × SAM+Cam	0.290	-0.041	0.169	-0.051
	(0.096)***	(0.126)	(0.131)	(0.148)
... × 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

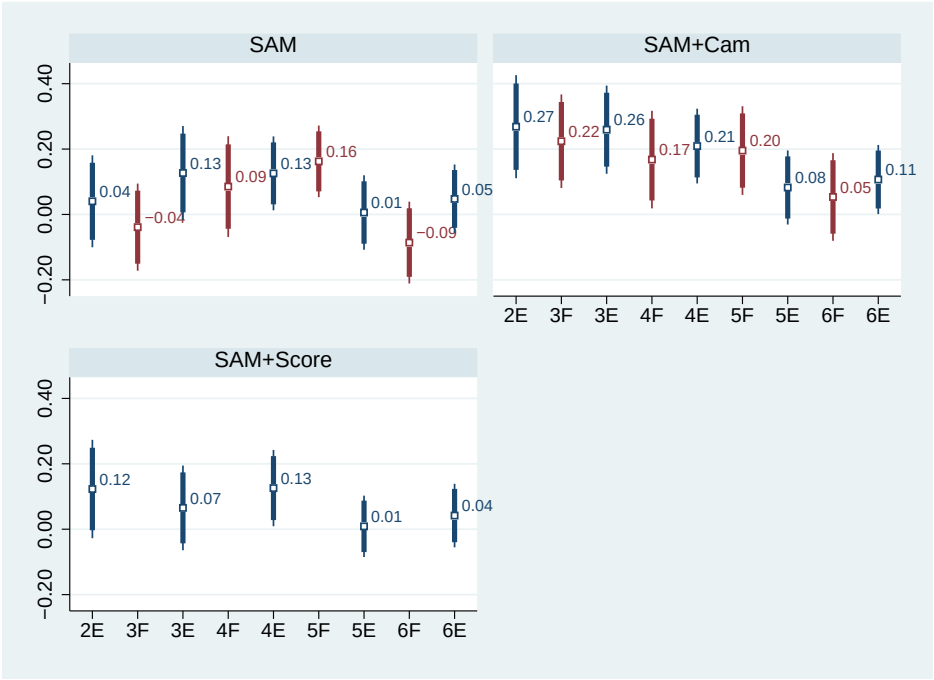
Table G.7: Impact on School Principal's Reported Excuse for Teacher Absences

	Off-school assignments (No penalty)		Late arrival, early departure (0.5–1.5% penalty)†		Sick and personal leave (0–2% penalty)‡		Others	
	2018	2019	2018	2019	2018	2019	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SAM	0.100	0.199	0.019	-0.140	-0.068	-0.052	-0.050	0.001
	(0.074)	(0.088)**	(0.048)	(0.052)***	(0.070)	(0.069)	(0.095)	(0.088)
SAM+Cam	0.084	0.206	-0.032	-0.133	0.025	-0.144	-0.077	0.063
	(0.089)	(0.090)**	(0.053)	(0.053)**	(0.072)	(0.067)**	(0.101)	(0.092)
SAM+Score	0.112		0.004		0.041		-0.146	
	(0.073)		(0.056)		(0.065)		(0.096)	
Control group mean	0.20	0.28	0.09	0.15	0.21	0.22	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

(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 G.1: Impact on Mean Scores at Midline and Endline by Baseline Grade

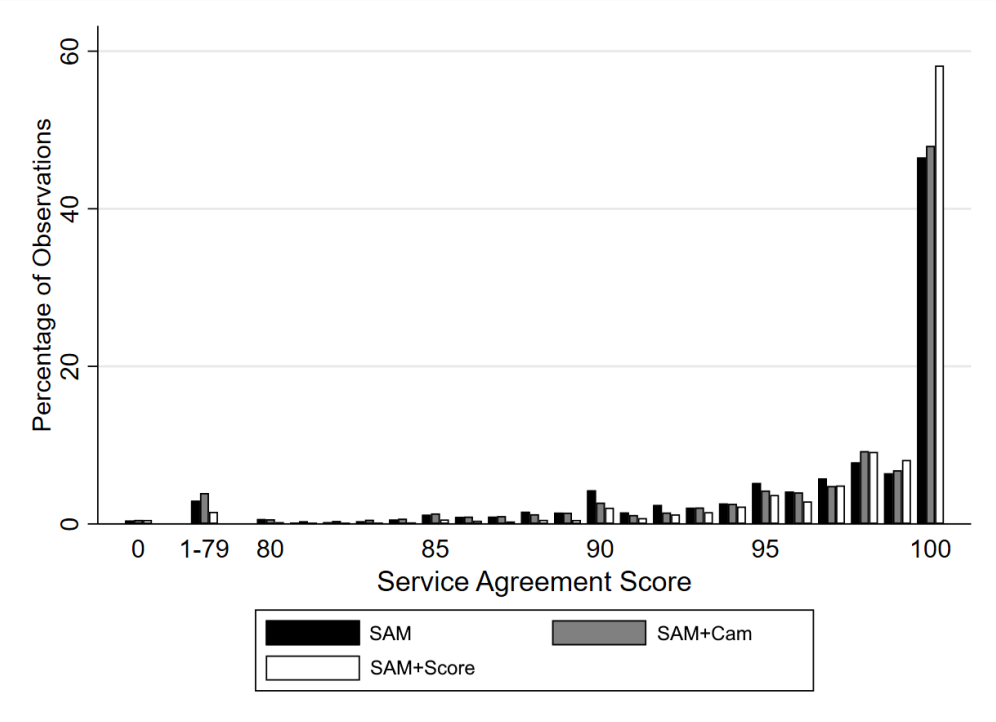


Figure G.2: The Distribution of the Service Agreement Scores by Treatment