Scores, Camera, Action?
Incentivizing Teachers in Remote Areas

Arya Gaduh* — *University of Arkansas*

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

Jan Priebe ‡ — *GIGA Institute Hamburg, University of Göttingen*

Dewi Susanti § — *World Bank, Jakarta*

October 2019

Abstract

Poor teacher accountability leads to poor education quality, especially in remotely-located schools that are costly to supervise. This paper reports the impacts of three interventions that linked community-based monitoring to a government allowance for teachers working in remote areas in Indonesia. In all treatments, the project helped communities to formulate a joint commitment between schools and community members to improve education. Teacher-specific scorecards were developed based on this commitment and performance was evaluated and disseminated by a newly-formed user committee. Treatment 2 and 3 added to this a pay for performance scheme that relied on the community reports. In Treatment 2, the remote area allowance was made dependent on teacher presence, which was monitored with a camera with a time stamp. In Treatment 3, the overall score on the scorecard determined the allowance. We find improvements in learning outcomes across all treatments; however, the strongest impacts of between 0.17-0.20 standard deviation (s.d.) were observed for Treatment 2. In this treatment, teachers increased teaching hours and parents increased investments in their children's education. We show evidence that bargaining and the community's propensity to punish free-riders may have a role in affecting treatment effectiveness.

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

Keywords: Teacher incentives, community-based monitoring, performance pay, remote-area policy

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

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

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

§The World Bank, Indonesia Stock Exchange (IDX) Tower 2 L12 & L15, Jalan Jend Sudirman, Senayan, DKI Jakarta 12190, Indonesia. Email: dsusanti@worldbank.org.

Acknowledgments *A large number of people contributed to the design, implementation, data collection, data analysis, and policy recommendation of this research project. For design, we are grateful to Amanda Beatty, Christopher Bjork, Jishnu Das, Deon Filmer, Scott Guggenheim, Rema Hanna, Nur Hidayat, Gunawan, Marliyanti, Karthik Muralidharan, Setiawan Cahyo Nugroho, Lant Pritchett, Jurist Tan, Robert Wrobel, Deny Purwo Sambodo, Halsey Rogers, Dewi Sudharta, and Daniel Suryadharma. For excellent research assistantships, we thank Usha Adelina, Emilie Berkhout, Kurniawati, Sharon Kanthy Lumbanraja, and Indah Ayu Prameswari. Survey data collection was led by Dedy Junaedi, Lulus Kusbudiharjo, Anas Sutisna, and Mulyana. Implementation by BaKTI was led by Muhammad Yusran Laitupa, Setiawan Cahyo Nugroho, Tri Yuni Rinawati, and Caroline Tupamahu. Research and implementation supports were provided by the World Bank under the leaderships of Nina Bhatt and Kevin Tomlinson, with inputs from Gregorius Kelik Endarso, Tazeen Fasih, Yulia Herawati, Lily Hoo, Megha Kapoor, Camilla Holmemo, Javier Luque, Cristobal Ridao-Cano, Audrey Sacks, Chatarina Ayu Widiarti, Noah Bunce Yarrow, and Fazlania Zain. We are grateful to Andrew Brownback, Robert Garlick, Alejandro Ome, and audiences at the 2019 briq/IZA Workshop on Behavioral Economics of Education, the 2019 RISE Seminar, the 2019 Pacific Development Conference, the 2019 Midwest International Economic Development Conference, the 2019 DIAL Development Conference, the 2019 Annual International Conference of the Research Group on Development, and the 2019 NEUDC conference for helpful comments and suggestions.*

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

1 Introduction

Service delivery in remote areas in developing countries is often of extremely low quality. The high cost of delivering services combined with the unwillingness of teachers and doctors to locate to remote areas make the challenge of providing quality services a daunting one. Many governments have responded to the challenge by pouring in additional resources, for example by giving special allowances for teachers willing to work in remote schools. However, the little evidence on the effectiveness of this approach is not promising. A hardship allowance in Gambia, equal to 30 to 40 percent of base salary, had no impact on student learning outcomes (Pugatch and Schroeder, 2018). An absenteeism survey in Indonesia found that remote-area teachers receiving a similar allowance were absent from school 32 percent of the time, more than other teachers in the same schools which did not receive the allowance (SMERU, 2010, Table 12). These findings are in line with a broader literature showing that unconditional grants are not a cost-effective way to improve education in developing countries (de Ree et al., 2018; Mbiti et al., 2019).

A key challenge to improve service delivery in remote areas is finding effective ways to hold service providers accountable. Administrators lack information on the quality of services delivered and travel costs make supervision visits prohibitively expensive. One potential solution is to rely on alternative ways of gathering information. The Indian NGO Seva Mandir did so by asking teachers they employed to take pictures of themselves at school at the beginning and end of the work day, which were then transmitted by mobile phones. Salary payments were made dependent on the presence as recorded by the cameras. The NGO World Vision adopted a similar approach in rural public school in Uganda, but relied on head master reports on teacher presence. Teachers were paid bonuses based on reported presence. Despite evidence that incentivizing service providers based on this type of information improves performance (Duflo et al., 2012; Cilliers et al., 2018), governments have been hesitant to link such reports to actual incentives (Banerjee et al., 2008). Instead, most programs rely on some form of social accountability mechanism in the hope that that increased transparency and community pressure will provide sufficient incentive for changing behavior. The evidence of effectiveness is mixed and suggests that while these processes can be implemented successfully, the impact on service delivery is limited because communities are in a relatively weak bargaining position to demand real change (Joshi, 2013).

In this paper, we report the results of a randomized controlled trial of three interventions that combine social accountability and pay-for-performance mechanisms to improve teacher accountability and learning outcomes in primary schools in remote villages of Indonesia. We worked with the Indonesian government to implement three interventions that combine different ideas from the pay for performance literature and community monitoring. We test their effectiveness in a large sample of mostly public schools in Indonesia. The pay for performance components incentives the teachers special allowance (TSA), a government financed supplemental allowance in the amount of up to one times the teacher's base salary for teachers assigned to a remote area.

The common component of these interventions is the social accountability mechanism (SAM) with two main elements. First, facilitators help communities formulate a joint agreement between teachers, village leadership, and parents to improve education quality. Teacher commitments are then formalized through a set of locally-defined service standards that include teacher-specific scorecards, which always include the teacher presence indicator. Second, SAM also facilitates the formation of a user committee

(UC), whose task is to monitor and evaluate each teacher based on his/her scorecard on a monthly basis. The scores are reported to the district government and UCs are encouraged to meet every month to discuss teacher evaluations at the school and publicize the results.

The three interventions vary in the way the teacher scorecards translate into teacher incentives. Treatment 1 relies entirely on the aforementioned SAM. The monthly meetings provide informal pressure if the performance falls short of an agreed commitment to improve service delivery. Administrators are also informed of scores, and could act upon them. Treatments 2 and 3 add to this a pay-for-performance mechanism (PPM). In these treatments, poor performance as captured in the scorecard leads to cuts to the TSA. In Treatment 2, the cut is based on the teacher presence indicator only, which is objectively verified using a tamper-proof smartphone camera provided to the schools. Meanwhile, in Treatment 3, the cut is based on the overall score on the scorecard; however, these schools did not receive the tamper-proof camera. Teachers with maximum scores, and teachers in control group and Treatment 1 always receive their full TSA.

The study was done in 270 schools in 2 districts in East Nusa Tenggara and 3 districts in West Kalimantan provinces of Indonesia from October 2016 to March 2018. To implement it, we worked closely with the National Team for the Acceleration of Poverty Reduction (TNP2K), the Indonesian Ministry of Education and Culture (MOEC), and the participating district governments. Districts were selected from the central government's list of disadvantaged regions, while taking into account cost and implementation considerations. Within each districts, we included schools that satisfied our remoteness criteria and had a minimum number of teachers receiving the remote-area allowance. We then use stratified-random assignment to assign schools to the control and treatment groups.

This experiment allows us to address some research questions that have not been addressed in earlier studies on community participation and incentive pay for teachers in developing countries. To our knowledge, it is the first study that evaluates the impact of tying pay for performance to community monitoring reports. Having treatment arms with and without teacher incentives allow us to test the incremental effect of pay for performance over a mechanism that relies on community monitoring only. Second, having both Treatments 2 and 3 allow us to compare an incentive scheme based on an objective, well-measured, but incomplete indicator of performance with one that is more comprehensive, but also more subjective in the way it measures performance. While the former is arguably fairer, it could lead to teachers shifting just focusing on the incentivized indicators (presence) while reducing effort on other activities that matter for learning (Holmstrom and Milgrom, 1991). Treatment 3 overcomes this problem to a large extent, but introduces uncertainty over how the evaluations will take place which, particularly in a low trust environment, increases the scope for bargaining and could make the incentive less effective (Baker et al., 1994). All experiments allow us to investigate how education managers change their behavior in light of new information and incentive structures.

The performance pay experiment is also unique in that: (i) it incentivizes a salary component that is part of a regular civil servant pay;[1] and (ii) it provides *negative* incentives for under-performance instead of bonuses for good performance. The former increases the scope for a possible scale up in the public sector, which is often problematic when the government only gets involved after the initial experiment

[1]While the large share of the affected teachers in our experiment are civil servants, non-civil servant teachers (including those teaching in private schools) also received TSA if eligible.

(Banerjee et al., 2008; Bold et al., 2018). It also signals to teachers that the government is involved, which may result in them taking the performance evaluation more seriously. With respect to the latter, if teachers exhibit loss aversion, the threat of cutting an existing allowance may be more effective than providing additional allowances to improve performance (Fryer Jr. et al., 2018).

Our results show that all treatments increased learning outcomes, measured by assessments of Indonesian and mathematics. However, while we find that all treatments led to positive improvements in learning, Treatment 2 shows the most promise. Treatments 1 and 3 led to similar improvements in learning. In comparison, the effect sizes in Treatment 2 approximately doubled. Treatment 2 led to improvements in Indonesian and mathematics by 0.17 and 0.20 standard deviation (s.d.). Overall, these impacts do not differ by gender, but are stronger for students in earlier grades and those who were better performing at baseline.

The treatments increased teacher presence and led to other behavioral changes. Across all treatments, we found that teachers were more likely to be in school when they were supposed to. We also find that teachers were more likely to teach (instead of doing administrative or other work) in class. In Treatments 2 and 3, many of these behavioral changes were driven by teachers who received the remote area allowance. In Treatment 3 however, we found that non-recipients decreased effort in response to the intervention.[2]

We also find that treatments led to significant changes in parental behavior and perception of the school. Treatments increase parental investments in their children's education. Parents in treated communities increased education expenditures and were more likely to find additional support (such as a tutor) for their children. Moreover, parents in treated communities also interacted more with teachers, and were more satisfied with education service delivery and their children's schools in general. These improvements tend to be stronger in Treatment 2 where the scorecard and pay-for-performance mechanism are supplanted with the camera.

Can the teachers responses be explained by increased top-down supervision or informal pressure from the SAM? Our evidence suggests that both mechanisms are at play in delivering results. Across the board, principals increase their supervision of teachers. Supervision from district officials and school inspectors also increased in Treatment 2. Communities only make sparse use of the tools they have been provided to hold teachers accountable. Teacher rating scores are generally very high at around 95 percent of the maximum score.[3] The average salary cut for teachers who received the remote area allowance was 5 percent. This indicates that the scores function as a bargaining tool that do not always translate into penalties.

Finally, we explore whether local norms and bargaining play a role in affecting outcomes. We find that the scores in Treatment 3 are somewhat higher than in Treatments 1 and 2, while independently observed outcomes do not corroborate this. Furthermore, we also find that user committees report a

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

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

higher incidence of pressure from teachers to user committees to improve scores in Treatment 3, indicating that subjective evaluations resulted in more bargaining. Furthermore, we find punishment norms significantly influence the effectiveness of the pay-for-performance treatments. Using a lab-in-the-field experiment in a subset of schools to estimate local punishment norms, we find larger student learning gains in schools with a higher propensity to punish free riders.

The cost-effectiveness of our interventions are comparable to interventions that adopted similar approaches. Treatment 2, which was the most successful among our interventions, improved learning outcomes by 0.2 standard deviation (s.d.) at the cost of USD 44 (in current 2017 dollar) per student. When converted to current 2011 dollar for comparability, this implies a cost of USD 22 per 0.1 s.d. learning improvement for Treatment 2. This cost is somewhere in the middle of the distribution of the cost-effectiveness of the various interventions reported in JPAL (2019).

The paper contributes to the empirical evidence on how community monitoring and teacher performance pay incentives can be used to increase education quality in rural areas of a developing country (Glewwe and Muralidharan, 2016, Section 4.4). We show that a facilitated process of setting standards and monitoring improves learning outcomes in the short run, through a combination of increased parental and teacher effort. This is promising, considering the generally weak track record of improving education through increased community participation. We provide external validity to idea of linking pay to teacher presence monitored using tamper proof cameras (Duflo et al., 2012) and show it can be implemented in in public schools using allowances paid by the government. Our paper also contributes to the literature on personnel economics of the state in developing countries (Finan et al., 2017). We show that when the institutional capacity for monitoring is weak, and clients can monitor service delivery, a very simple contract based on monitoring presence only works better than a more comprehensive, less well specified one. This is an important question that arises in many labor contracts (Baker et al., 1994) and this paper compares these two approaches in one experiment.

The rest of the paper is organized as follows. The next section discusses the context and the experimental design, including how the interventions were implemented in the field and how communities respond to the interventions. Section 3 describes the data collection and empirical strategy. The following two sections discusses the impact of the treatments on student learning outcome (Section 4), and teacher behavior and parental engagements in their children's education (Section 5). Section 6 provides further insight the political economy aspects of implementing a pay-for-performance mechanism that relies on community monitoring and reports. Section 8 concludes.

2 Experimental Design

2.1 Context: Teacher Accountability and Community Participation

With almost universal access to basic education in 2017, the Indonesian government has shifted its attention from access to equity and quality improvement. Disparity among rural and urban locations persist in education service delivery and outcomes. Two thirds of schools in remote areas are lacking teachers, while two thirds of urban schools have too many teachers (World Bank, 2013). Around 50 percent of population age 15 and above in rural areas have not or just completed an elementary education, compared to 35 percent in urban areas. Recent international assessments show that Indonesian student

learning outcomes remain at the bottom rank of participating countries (World Bank, 2013; OECD, ed, 2014). Student learning outcomes in remote areas are lagging significantly behind urban areas (ACDP, 2014; Stern and Nordstrum, 2014).

Until recently, the government approach to improve quality focused on improving teacher welfare. Twenty percent of national and district government budgets are allocated for education and half of this is allocated to pay close to three million teachers' salaries and allowances. Since 2005, the government enacted the Teacher Law that provides a certification allowance (*Tunjangan Profesi Guru*) of up to their base salary for teachers who took the administrative steps to get themselves certified. Teachers whose school is located in special areas, including remote areas, receive an additional teacher's special allowance (*Tunjangan Khusus Guru*) of up to their base salary. However, as none of these allowances are determined by teacher performance, they hardly lead to quality improvement: Recipients of the special allowance were more likely to be absent relative to non-recipients in the same school (SMERU, 2010) and the professional allowance policy did not improve student learning (de Ree et al., 2018).

Teacher accountability is a key challenge to improve public education quality, particularly in more remote parts of Indonesia. Consider, for example, the problem of teacher absence. The rate of teacher absenteeism in Indonesia has declined over the past decade, but it remains high in remote areas (19.3 percent) compared to the national rate (9.4 percent) (Usman et al., 2004; ACDP, 2014). High absenteeism rates negatively affects quality, as it increases student absenteeism, drop-out rates, and lowers student learning outcomes (Usman et al., 2004; UNICEF, 2012; Hasan et al., eds, 2013; Suryahadi and Sambodho, 2013). Weak capacity to enforce quality standards, both at the top (government) and the bottom (community) contribute to the lack of improvement.

Indonesia has a wide-ranging experience with community-driven development (CDD) programs. Developed following the Asian Economic Crisis and the fall of the Suharto regime, 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). 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 in can, in part, be attributed to the long history Indonesia has in mobilizing community contributions for rural development programs (see p.71, Mansuri and Rao, 2012). Recent studies have investigated how CDD programs could be harnessed to increase use of health and education services (Olken et al., 2014).

2.2 Intervention Design

The Teacher Performance and Accountability interventions (hereafter referred to by its Indonesian abbreviation, *KIAT Guru*) aim to empower communities to hold teachers accountable. Its design was informed by international evidence on key elements necessary to ensure that a community-based approach can improve service performance. These elements include: (i) having a standard to which the service providers will be accounted for; (ii) improving communities' access to information, including their basic rights to services; (iii) giving communities the means to influence and voice concerns to service providers; and (iv) providing routes to sanction poorly performing service providers (Joshi, 2013; Ringold et al., 2012). There is also some evidence that locally-defined and agreed-upon service standards are more effective

than nationally-defined service standards in improving performance (World Bank, 2014, p.48).

This study follows up on an earlier study that tested different ways to strengthen school committees in rural Central Java. Pradhan et al. (2014) showed the importance of involving local leadership and ensuring that community involvement leads to concrete actions that improve education. It underlined the difficulty of inducing increased efforts of teachers if there are no incentives attached to community action. A pathway analysis suggested that the positive effects on learning in this study were mostly a result of increased inputs of the community and not teacher effort.

The final design for KIAT Guru was informed by an operational pilot conducted in 31 schools in very remote villages in Keerom, Kaimana, and Ketapang districts of Indonesia, from June 2014 to December 2015. The operational pilot tested the implementation of key processes (e.g., facilitation of community meetings, pay-for-performance mechanisms), the legal and administrative regulations, process-monitoring instruments, and the survey instruments. Key lessons learned from the operational pilot set the parameter for the implementation of the study, particularly on district and village selections.[4]

2.2.1 Experimental Treatments

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

Social Accountability Mechanism (SAM). All treatments include a facilitator-driven set of meetings to establish the service standards (i.e., the service agreement) and the monitoring institution (i.e., the user committee). The first of these meetings was an orientation meeting, attended by student, parents, community members, and school management (including teachers) to inform them about the pilot and their rights to participate in education service delivery. Subsequently, three separate meetings with representatives of students and alumni, parents and community members, and teachers gathered inputs from each stakeholder on how to improve learning environment at school and at home, and what needed to be done by various education stakeholders. Afterward, all stakeholders came together to formulate the service agreement. The service agreement lists a set of actions to improve the learning environment that parents, community leaders, teachers, and the school principal would commit to.

The service agreement became the basis for the the teacher- (and principal-)specific community scorecard. Between 5 and 8 indicators that the principal and teachers committed to in the service agreement were made part of the scorecard. Although meeting participants were free to choose the included indicators, the scorecard must always include the presence indicator. Once the indicators were chosen, participants then assigned a weight to each indicator that reflected (their belief of) its importance to improve learning. These weights must add up to 100. In a separate meeting, the UC would then define the

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

service standards that guide how each indicator would be scored. A scorecard would therefore consist of a set of indicators, each was accompanied with a weight and a scoring guideline.

To monitor and evaluate teacher compliance to his/her scorecard, a user committee (UC) was established. The UC must have a minimum of nine members with a majority of them being female. It should include three community/ religious leaders, while the rest are parents representing each grade level. The facilitation manual was cognizant of other village and school committees and encouraged overlapping memberships. However, as implemented, only a small percentage held memberships in other committees.

In addition to the UC, the facilitator also recruited a village cadre who would be prepared to take over the role of a facilitator. The village cadre organized monthly village meetings and facilitated the meetings. Seventy five percent of the village cadres were appointed and introduced by the first village meeting. They co-organized and co-facilitated meetings with the facilitators.

Both the village cadre and the user committee members were formally appointed through Village Head decrees, and they were recognized in the district- and national-level regulations as people whose roles were to organize meetings and monitor and evaluate teachers respectively. They received capacity development training from the pilot at the district or sub-district levels, and on-the-job mentoring by the facilitators. Their training included information on how to gather evidence to evaluate teacher service performance in three ways: conducting unannounced visits to the school, interviewing students or teachers, and auditing administrative documents. No teachers attended these trainings, except for one conducted between February to April 2018, when the pilot facilitator handed over the project to the stakeholders and provided capacity development to strengthen cooperation amongst stakeholders and sustain implementation.

Throughout implementation, the UC conducted monthly meetings to review the implementation of the service agreement and evaluate the scorecard. In these meetings, stakeholders were to present their view about the progress for SA indicators and discuss potential improvements. The UC were to present their monthly evaluation of the scorecard and allow each teacher an opportunity to respond. Once the score for each teacher was finalized, the meeting ended with everyone signing off on the evaluation results. These evaluation results were then posted or announced in another village meeting and sent to the district government.

After a few months of implementation, a village-wide meeting was held to evaluate the SA, the scorecard, and the UC membership. Prior to this evaluation meeting, the village cadre and members of the UC who had undergone training administered an adaptive Diagnostic Student Learning Assessment (hereafter, the diagnostic test). The diagnostic test identifies students' skills in basic literacy and numeracy along a learning continuum of the national curriculum. The diagnostic test was administered to a random sample of six students per grade level. A total of 5,967 students were tested by 897 UC members and the village cadres. Results from the diagnostic test were shared at the beginning of this evaluation meeting.

Pay-for-Performance Mechanism (PPM). To understand the PPM component, we first describe the incentive structures in our sample schools. More than 90 percent of the schools in our sample are public schools with three types of teacher status: permanent, contract, and school-contracted teachers. Permanent teachers are tenured civil servants (PNS) hired by the central government, while contract teach-

ers are hired either by district or provincial governments under annual contracts. Meanwhile school-contracted teachers are hired by the schools with a temporary employment status. The monthly pay range is highest for permanent teachers (between around USD 108 and USD 408 depending on seniority), followed by contract teachers (between around USD 73 to 146), and school-contracted teachers (between USD 22 and 51). Private schools have all three types of teachers. Some permanent and contract teachers had similar administrative status: they were similarly employed by the government as but assigned to the private schools.

The PPM component of our treatments is tied to the teacher's special allowance (TSA), equal to up to one time the base salary, for which both the permanent and contract (but not school-contracted) teachers were eligible. Until 2016, there was a national quota for the TSA allocation based on proposals from the districts. By 2017, when the pilot just began, the government made use of a national index that identified very remote and disadvantaged villages to allocate TSAs. Private and public school teachers with either permanent or contract status who had registered with MOEC and were assigned to very remote villages automatically received TSA. Its value ranged from USD 103 and up to one times the teacher's base salary per month. Note, however, that certified teachers also received certification allowance of a similar amount (on top of their base salary) that would not be affected by our PPM.

Our treatments vary in how performance evaluations affect the amount of TSA allowance received. There was no PPM component in Treatment 1, and therefore eligible teachers always received their full TSA amount. Treatments 2 and 3 differ in the indicators (and tools) that were used to link performance with (the cuts to) the TSA allowance. Across treatments, non TSA teachers were evaluated the same way as TSA teachers, but the evaluation did not affect their salary.

In Treatment 2, teacher presence is the only determinant of the amount of TSA received by eligible teachers. Teachers in Treatment 2 schools are provided with a tamper-proof smartphone camera to provide proof of their presence. They take pictures at the beginning and end of a school day and record the times on a manually-entered teacher attendance form. At the end of each month, members of the UC verify both entries and any letters provided by teachers to account for their absences. There are four types of possible entries, and each determines the total amount cut from their TSA. The entry type (daily percent cut) is as follows: full presence (0), partial presence (up to 1.5), excused absence (2), and unexcused absence (5). Once tallied at the end of the month, teachers whose total cut exceeded 15 percent will lose their monthly TSA. To accommodate the use of the smartphone camera, the facilitators held an additional training to use it during the monthly community meeting. Moreover, Treatment 2 schools added verification of the camera reports to its monthly meeting agenda.

In Treatment 3, the scores used to determine the amount of TSA received were based on the scorecard. Three things distinguish Treatment 3 from Treatment 2. First, even though the scorecards are monitored in all treatment groups, its score only affected the amount of teacher's remote-area allowance in Treatment 3 schools. Second, unlike in Treatment 2, there was no cut-off score below which a teacher would not receive the allowance: If a teacher received a score of 79 for that month, that she would receive 79 percent of her TSA allowance. Finally, recall that presence is a required indicator in all scorecards. However, without the camera, the UC would have needed to proactively monitor presence following the steps suggested during the SAM training to gauge teacher presence.

The TSA allowances in all treatment groups were paid on a quarterly basis. TSA for civil servant

teachers were paid by the district governments, while TSA for the non-civil servants were paid directly by the Ministry of Education and Culture. All payments were made through direct transfers into the teacher's bank account.

Table 1: Summary of the Treatments

	Control	Treatment 1	Treatment 2	Treatment 3
SAM: Scorecards and user committee	No	Yes	Yes	Yes
PPM: Presence indicator	No	No	Yes	Yes
PPM: Indicators other than presence	No	No	No	Yes
Tamper-proof camera	No	No	Yes	No
Number of schools	67	68	68	67

A Summary of the Treatments. Table 1 summarizes how our treatments are organized. *Treatment 1* facilitated the development of the service agreement, scorecard, and user committee but did not link any of the evaluation results to the TSA. As such, teachers in Treatment 1 schools receive the full amount of their TSA. *Treatment 2* similarly implemented the community empowerment intervention, but introduced a pay-for-performance scheme where cuts to the TSA are only determined by teacher absence. Cameras are used in Treatment 2 to objectively verify teacher presence. Finally, *Treatment 3* implemented a different pay-for-performance scheme: instead of relying solely on teacher presence, cuts to the TSA depended on the wide array of indicators listed in the scorecard (which would always include teacher presence). Moreover, cameras are not used in Treatment 3.

2.2.2 District and School Selection

We work in willing districts with significant problems of teacher absenteeism in remote, disadvantaged districts. Based on lessons learned from the operational pilot, we exclude districts with very weak governance and with transitory communities (i.e. fishing and the bush communities). To ensure manageable implementation costs, we excluded districts with very high transportation costs.[5] We also exclude conflict-prone areas, and districts that were part of many other education pilots. Finally, we limit the districts to those that had at least 40 primary schools in rural areas that fulfill our definition of eligible schools described below. Our final list included three districts in West Kalimantan (Ketapang, Sintang, and Landak) and two districts in East Nusa Tenggara (Manggarai Barat and Manggarai Timur).

Schools need to satisfy four eligibility requirements to participate in the study. First, each school must have a minimum of 70 registered students. Second, since the PPM interventions link evaluations to the remote-area allowances, at least 3 of its teachers must receive the remote-area allowance in 2017. Third, schools must satisfy a remoteness criterion of being located in a village that was at least one-hour drive away from the district capital. Our data suggest that on average, participating schools are located around 40 km (and about a two-hour travel time) from the subdistrict education office (i.e., *Unit Pelaksana*

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

Teknis Dinas Pendidikan). 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.[6]

2.2.3 Treatment Assignment and Compliance

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

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

Moreover, a few weeks before the intervention started, the Ministry of Education and Culture changed its mechanism to define eligible TSA locations. It used a national index instead of district head recommendations to determine eligibility, where all registered teachers working in these villages would automatically be eligible. This change took away the TSA eligibility of three villages. These affected schools were all part of the control group.

2.3 Details on the Implementation

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

2.3.1 The Social Accountability Intervention

The set of eight meetings started in November 2016 and completed in June 2017. Details on these meetings were retrospectively collected in 166 schools during monitoring visits. On average, meetings in this phase took 3.3 hours, with an average of 38 days to complete the seven set of meetings in each school

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

spanning from 6 to 155 days. While each facilitator was assigned to between four and six schools, the initial set of meetings in 57 percent (95 of 166) schools were facilitated by two or more facilitators due to personnel safety reasons and different strategies taken to encounter various logistical and geographical challenges. The formulation of the service agreement and teacher-specific scorecards took the longest time, with forty percent of the schools took between three to seven hours, and in the rest of the schools the outputs could only be achieved over 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. Initially, the second-most common indicator (after the requisite teacher-presence indicator) was a safe environment free of physical and verbal abuse — an indicator whose importance was emphasized during the socialization process. Other indicators were on improving learning (e.g., teachers were to conduct various ways to teach and enhance understanding, improve reading, writing and counting, provide additional lessons, provide feedback to students), motivating students, introducing students to social and cultural norms, communicating with parents, and improving teacher behaviors and conducts. Appendix Figure A.1 shows an example of the scorecard.

During the evaluation meeting where UC members can revise these indicators conducted around August 2017, we find an increase in indicators that focused on the student learning process from 33 to 48 percent.[7] At the same time, we find the committees were most likely to drop the corporal punishment indicator that teachers felt was too difficult to implement.[8] Due to geographical challenge and time constraint, the scorecard revision meetings were only facilitated in 173 out of 203 schools. In the rest of the schools, stakeholders were trained to conduct the meetings and were expected to implement them independently.

Figure 2 shows the evolution of the mean scores over time between August 2017 and June 2018. Average scores are generally high, in the range from 94 to 98 on a 100 point scale. The scores given for Treatment 3 are slightly higher than those given in Treatments 1 and 2. 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 throughout the duration of implementation. About 45 percent of UC members were female and around 31 percent had more than a secondary school education. Meanwhile, 26 percent of the village cadres were female, with the majority having a high-school degree or higher. During implementation, we observe variations in how 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. Focus group discussions with the facilitators identified that these differences were influenced by cultural norms, initial resistance from teachers to have their performance being evaluated so openly, and other village-specific idiosyncracies.

[7] 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.

[8] 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 provide trainings or information on strategies to conduct positive discipline for children, the stakeholders were at a loss.

By the end of 2017, meeting facilitation was fully managed by the village cadres. In 2017, 83 percent of the treatment schools received funding from village heads to provide operational costs for monthly meetings and incentives for the village cadres and UC members. The amount and allocation of funding provided by village heads ranged widely.[9]

2.3.2 Pay-for-performance

Two issues affected the early implementation of the pay-for-performance mechanism. First, administrative holdups delayed the implementation of the incentive payments for approximately 15 percent of the 135 Treatment 2 and 3 schools. Out of 135 schools, 115 had their first evaluation meeting between April and May 2017, and received their first incentive payments in July 2017. The remaining 20 schools affected by the holdup held their first meeting in August 2017. By October 2017, all 135 schools have received their incentive payments. Second, due to the end-of-year budgetary account closure, TSA's for the second half of November and December 2017 were paid in full irrespective of the scorecard.

We find clear evidence that the scorecard did determine cuts to the allowance as stipulated by these treatments. TSA teachers in Treatment 3 received an average pay cut of around 6.5 percent, whereas teachers in Treatment 2 who received less than a full score received a cut of 16 percent. Furthermore, we find strong evidence of compliance of the pay-for-performance rule for Treatment 2. In Treatment 2, 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. Figure 3 plots the payment cut as a function of the presence score and finds that in 87 percent of the case, the payment schedule was applied correctly.[10]

3 Data and Empirical Strategy

3.1 Data Collection

Student Learning Assessments. To evaluate student outcomes, the research team developed its own student learning assessments (SLA) instruments. The instruments assess basic functional literacy (in Indonesian) and numeracy competencies along the learning continuum standards set in the 2006 national curriculum. Designed based on frameworks and findings from other assessment tools (ASER Centre, 2014; Uwezo, 2012; Gove and Wetterberg, 2011; Platas et al., 2014), the developed tools consist of (i) a diagnostic test which aims to quickly capture students' competencies in literacy and numeracy; and (ii) an evaluation test which maps students' more specific abilities along the literacy and numeracy learning continuum.

The diagnostic test results is an advocacy tool used to implement SAM and is not used for the impact evaluation (see Section 2.2.1). The evaluation test was fielded in all schools and is utilized for impact evaluation. Separate test booklets were developed for each elementary grade level with multiple-choice

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

[10]Interestingly, in 11.4 percent of the cases, a higher cut was applied than specified by the rule. The pattern of the data suggests that in these cases, the district applied a fixed cut for everyone and applied the incentive rule on top of this. There is no clear regional or inter temporal pattern on when this rule was applied.

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

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. The design and methodology of the KIAT Guru TAS were mainly adapted from Analytical and Capacity Development Partnership (2014) study in Indonesia, with additional input from instrument used in UNICEF (2012) study in Papua and West Papua. The instruments were pre-tested and utilized to gauge the rate of teacher absence from school and classrooms. Information on student absence from school are also collected. In its implementation, the research team implemented the TAS on the day of arrival for the the baseline and endline surveys, which were unannounced. Information on student absence from school are also collected.

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

Data Collection Timeline. An independent survey team collected the baseline and endline survey data, while project facilitators and project implementation team collected the monitoring data. Figure 1 shows the implementation timeline. The study started roughly the same time with the school academic year in July 2016. The baseline survey was conducted in October and November 2016 for 213 schools and completed in February 2017 for the remaining 57 schools. The endline survey was conducted in February until mid-April 2018, soon after the facilitators handed over facilitation to village cadres at the end of 2017.[11]

A qualitative research was also conducted by another group of researchers in nine schools in three districts. They conducted visits to these schools prior to the start of implementation in November 2016,

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

in September 2017 after a few monthly meetings have been implemented, and in March 2018 after project facilitators had left.

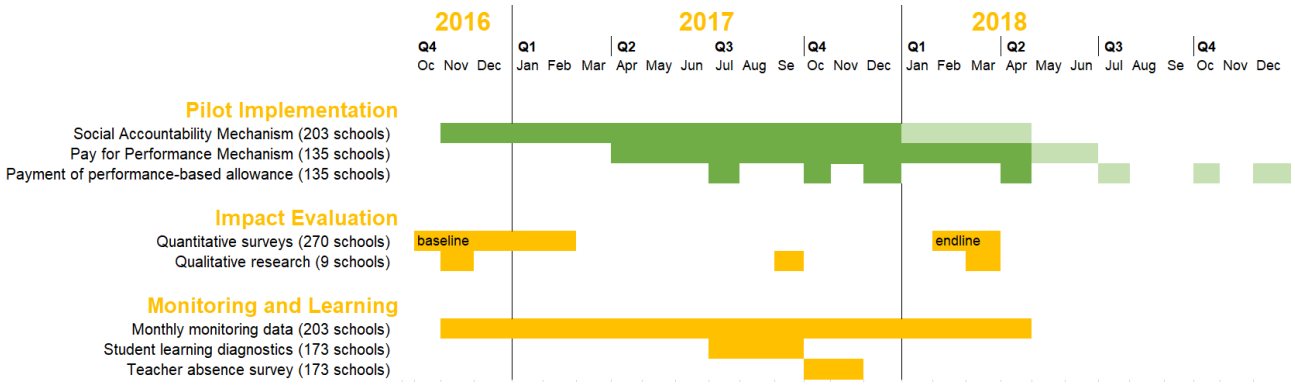


Figure 1: Implementation Timeline

3.2 Empirical Specification

We estimate the treatment effect based on the following regression model:

$$Y_{ijt}^{k} = \alpha_k + \delta Y_{ijt-1}^{k} + \sum_{R} \gamma^r T_j^{kr} + X_{ijt}^{k}\beta + \varepsilon_{ijt}^{k} \tag{1}$$

where Y_{ijt} = the outcome variable for individual i in school j at time $t \in \{0, 1\}$, α_k = the strata fixed effects, and X = control variables. T_j^r is the dummy variable for treatment regime r, and γ^r is the ITT estimate of interest for treatment r. The control variables depend on the outcome of interest, which we elaborate below for each outcome. Our baseline results include estimates with and without the control variables. Standard errors are clustered at the school level. In addition, for robustness, we implement the randomization inference procedure to calculate the *p-values* of the sharp null of no effect for each individual treatment, holding other treatments' assignments constant. Bjork et al. (2018) provides the pre-analysis plan for this study.

3.3 Baseline Summary Statistics and Covariate Balance

Table 2 presents the summary statistics of student, teacher, and parent baseline characteristics for the control and treatment groups. We observe poor literacy and numeracy among the students in the 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 is quite balanced across gender with 53 percent male. More than 80 percent of students have parents with only a primary education or less.

Teacher accountability is a serious problem in these schools. Our baseline teacher absence survey recorded almost 20 percent absenteeism rate. When present in school, only three quarters of the teachers were observed to be doing work. Furthermore, almost 40 percent of teachers were observed doing non-teaching activities in their classes.[12] The teacher population is balanced across gender and most teachers

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

have more than high school education. About half of the teachers are civil servants and close to one-fifth received the TSA in 2017.

Parents did not seem to be aware of these problems. At baseline, about 90 percent of parents believed that the quality of their children's school is either good or very good. Furthermore, only slightly more than one in five parent respondents reported teacher absence as one of the three main problems afflicting education in their community. At home, children received some form of tutoring for about 2.5 hours a week and less than half-a-percent of parents paid a tutor for their children.

Appendix Tables A.1–A.3 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. In our preferred specification, we include these covariates as control variables.

3.4 Differential Attrition and Entry

We use data on students who participated in the learning assessments and the administrative list of teachers working in each school to examine their differential attrition. Table 3 presents the results. Eight percent of the students in the control group could not be traced at endline. Columns 1–2 suggest that, controlling for individual characteristics, students are less likely to drop out of the student sample in Treatment 3. However, the lower attrition rate in Treatment 3 was not driven by better- or worse-performing students at baseline. Meanwhile, percent of teachers interviewed at baseline could not be traced at the endline, while 16 percent of teachers interviewed in endline could not be matched to the baseline. We find little evidence of differential attrition and entry of teachers across treatments.[13]

4 Impact on Student Learning Outcomes

Table 4 presents the results for the individual-level regressions of student learning outcomes. Results in odd-numbered columns do not include the control variables. The specification with the control variables includes sex, age dummies, both parents' education, the school-level mean baseline scores, and a private school dummy variable. In addition, we include a set dummy variables for missing baseline outcomes in all specifications and dummy variables for missing controls in the specification with controls. Columns 1–4 present the results for the learning outcomes from regressions based on the raw scores, while Columns 5–8 present the results based on the standardized scores.

We find that student learning outcomes improved in all treatments, but the effects were much more pronounced in Treatment 2. We discuss results based on the specification with the control variables (Columns 6 and 8). The SAM-only treatment (Treatment 1) improved Indonesian and mathematics outcomes by 0.09 and 0.07 standard deviation (s.d.) respectively. Linking monetary incentives to the overall scorecard rating (Treatment 3) yielded similar learning impacts of around 0.11 and 0.09 s.d. for Indone-

[13] In Appendix Tables A.4 and A.5, we explore the possibility of selective attrition and entry among students and teachers. The caveat on these results is that even if we find selective attrition/entry, they could be considered part of the treatment effect. We do not find evidence of selective attrition and entry among students. We find married teachers are more likely to drop out of the treatment schools.

sian and mathematics. However, having monetary incentives tied to the objective measures provided by the tamper-proof camera yielded impacts that were three halves and twice as large (0.17 and 0.20 for Indonesian and mathematics). The p-values from the randomization inference procedure are consistent with those from the regressions.

Heterogeneity Analysis. Table 5 presents our heterogeneity analysis by gender, baseline grades, and students' initial SLA scores. To estimate the heterogeneous impacts of the treatments, we estimate the following regression:

$$Y_{ijt}^{k} = \alpha_k + \delta Y_{ijt-1}^{k} + \gamma_h Z_{ij0} + \sum_{R} \gamma^r T_j^{kr} + \sum_{R} \gamma_h^r (T_j^{kr} \times Z_{ij0}) + X_{ijt}^{k}\beta + \varepsilon_{ijt}^{k} \quad (2)$$

where Z_{ij0} is the baseline variable we use for the heterogeneity analysis, γ_h^r is the differential impact for the subsample of individuals defined by Z, and the other variables are the same as in Equation 1. Columns 1 and 2 suggest that there is no evidence of heterogenous impacts of the treatments by gender. Columns 3 and 4 show that positive effects of Treatment 2 are more salient for early grade (Grades 1-3) students. Appendix Figure A.3, which plots the learning-outcome impacts by grade at baseline, supports the conclusion of stronger impacts on lower-grade students. Finally, columns 5 and 6 show that above-median-performing students — to wit, students with better baseline scores than their cohort in their class — benefit more from Treatment 2.

We find limited evidence for the role of TSA teachers on these learning outcomes. Table 6 presents the heterogeneous treatment impact by the TSA status of the students' teachers.[14] In this analysis, we focus on the panel sample of students. Columns 1–2 (3–4) separately examine the heterogenous impact of having a TSA teacher during the baseline (endline) academic year. For math, having a TSA teacher seems to amplify the treatment effects; however, for Indonesian, the results are more mixed. Columns 5–6 suggest that having a TSA teacher in both periods amplifies the treatment effects. Nonetheless, in all cases, the heterogeneous effects by their teachers' TSA status are noisily estimated and not statistically significant.

5 Teacher and Parent Responses

As students were not directly targeted in any of the interventions, we expect the impacts on student learning outcomes primarily to arise indirectly from changes in teacher efforts, and parental engagement and education investments. This section examines the impact of the interventions on teacher and parental behavior.

Teacher Presence and In-School Activities. Table 7 shows how the interventions affect independently recorded teacher presence and in-school activities. Columns 1–3 show that overall, the treatments had no impact on the likelihood that a teacher was present at school, but Treatments 1 and 2 had positive impacts on whether teachers were observed to be working when in school. Interestingly, the effect of Treatment 3 is negative, albeit non-significant. We also do not observe changes in the likelihood that teachers perform academic activities (hereafter, "teaching") when they were observed inside the classroom.

[14]The TSA status is based on the administrative data of TSA recipients by the 2017 academic year.

Columns 5–6 show that teachers' TSA status importantly affect their response to the treatments. The positive impacts of Treatment 2 the likelihood of working when in school or teaching when inside the classroom are exclusively driven by the TSA teachers. Importantly, the overall negative impacts of Treatment 3 on both of those outcomes were driven by the non-TSA teachers.

Teachers' Time Allocation. Table 8 reports how teachers adjust their allocation of time across different teaching and non-teaching activities as the results of the interventions. Column 1 suggests that the interventions did not affect the total time allocated to the school-related activities. However, teachers in Treatments 1 and 2 spent more time teaching intra-curricular materials, at the expense of assessment-related activities, such as grading homework and quizzes. This is suggestive evidence of some shift away from tasks that are more difficult to observe (by the user committee) to those that are easier to observe, and thus also easier to score (Holmstrom and Milgrom, 1991).[15] Importantly, Column 5 suggests that the treatments did not lead teachers to reallocate time toward non-school activities, such as private tutoring or other economic activities (e.g., teaching in other schools, farming, or other paid activities).[16] We see little evidence of time reallocation in Treatment 3 schools.

Parental Engagement in Education. Parents in treatment communities invest more money and time in their children's education. Table 9 reports parental investments in their child's education in the previous academic year. Education expenditures increased by about Rp 29,000 (approximately US$2) for Treatment 2 compared to the control-group average of Rp 324,154 (US$23), constituting an increase of 8.6 percent. For the other treatments the point estimates are smaller and not significantly different from those observed in the the control or Treatment 2. Across all treatments, parents report that their children receive more support in doing their homework by around 0.35 hours more from a base of about 2.5 hours per week. In Treatment 2 schools, the share of children with a paid tutor increased by 1.5 percentage point (p.p.) from a low base of 0.3 percent.

Parents also interacted more frequently with teachers in the previous academic year as the result of the interventions. We find consistent reports of increased interactions from both teachers (columns 4–5) and parents (columns 6–7). Although there is no overall increase in the number of meetings, Column 5 suggests that teachers and parents are now more likely to meet in a formal setting. Column 6 also shows that parents reported an increase of around 1 to 1.5 meetings (from a base of 1.5 meetings) to discuss learning-related issues, with a point estimate that is largest for Treatment 2.

6 The Political Economy of Implementation

The interventions introduced new tools to manage education to communities, parents and school administrators. In this section, we assess how they responded to the intervention in ways that go beyond providing inputs into the education production function. For principals and administrators, we are particularly interested in whether they responded to the information that was generated through the community monitoring. For all stakeholders, we are interested in whether the program affected their job

[15] In Appendix Table A.6, we show that even though intra-curricular teaching during school hours only increased for Treatment 2, it increased for after-school teaching in all treatments.

[16] Appendix Table A.7 breaks down the off-school activities and find no significant effect for private tutoring.

satisfaction and satisfaction with education service delivery. Positive responses on both aspects increase the likelihood that the interventions could be sustained with less project supervision. Finally, we explore how pre-existing norms can influence the success of these interventions.

6.1 School Management and Stakeholder Engagement

The new ways to monitor and enforce standards that these interventions introduced may differentially affect how schools are managed and this may have contributed to outcomes. 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 subdistrict or district education office), so that they could act on the information. Moreover, school principals might also feel the need to engage both their teachers and parents more. Table 10 presents the impacts of the interventions on how schools interact with other stakeholders and how teachers are managed.

Engagement with Education Officials. Columns 1–2 show that among the interventions, Treatment 2 was the most successful in increasing the engagements of supervising officials. Column 1 shows that Treatment 2 increased the number of meetings with the subdistrict education official by 0.9 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 0.6 from a base of 1.4. The pattern was similar for Treatment 1, albeit statistically insignificant, while Treatment 3 did not increase the number meetings with officials.

Principal Monitoring and Evaluation. Our findings from columns 3–5 suggest that the interventions led to increased teacher monitoring and evaluation by school principals across the board. Column 3 shows that all treatments increased teachers' likelihood of receiving in-class observation by the school principal. Furthermore, Column 5 shows that all treatments also increased their likelihood of receiving a routine performance evaluation from the school principal by between 9 p.p. (Treatment 1) and 15 p.p. (Treatment 2) from a base of 45 percent.

6.2 Teacher and Parent Satisfactions

Broad-based stakeholder support is important, especially for performance pay policies like our interventions. Ex ante, the impact of this type of interventions on satisfactions are ambiguous. On the one hand, subjecting teachers to routine evaluations and tying them with their pay can create dissatisfaction among teachers. On the other hand, if these evaluations are viewed as fair (or fairer than previous methods), they may improve overall satisfaction. Moreover, by increasing interactions between teachers and the community, the interventions may improve the relations between teachers and parents.

Teacher Satisfaction. In a context where teachers typically received their allowances unconditionally, pay-for-performance interventions introduced here can be considered unfair, leading teachers to feel unappreciated. Table 11 suggests that this concern may be unfounded. Columns 1–2 show that all treatments led to an overall increase in teacher satisfaction of the appreciation received from various stakeholders. Columns 5–6 suggest that teacher satisfaction of outside appreciation are not significantly different between TSA and non-TSA teachers.

At the same time, TSA-teachers who are affected by the interventions reported less satisfactions over their salary and, for Treatment 2, over their current job. Column 7 shows that TSA-receiving teachers in all schools are generally more satisfied with their salary. However, TSA teachers in treated schools are less satisfied with their salary. Interestingly, Column 8 suggests that interventions with a pay-for-performance component (Treatments 2 and 3) increased the share of teachers satisfied in their current job among *the non-TSA* teachers, but not among the TSA teachers.

Parent Assessments of Their Children's School. We also show in Table 12 that across the board, parents' assessment of the school quality are positively affected by these interventions. At baseline, there was already a high degree of satisfaction, even among control schools: 91 percent of parents rated their children's school as either good or very good. The interventions increased this by about 5 p.p. The interventions also affected their perception of whether teacher absenteeism is a main education problem in their community: The fraction of parents who reported this concern fell by between 6 p.p. (Treatment 1) and 8 p.p. (Treatment 3) from a base of 27 percent.

6.3 Norms and Bargaining

Our interventions facilitated all treated communities to establish a locally-agreed standard that can be used as a guide to evaluate teacher performance. In addition, the performance-pay mechanism in Treatments 2 and 3 empowered communities to hold their teachers accountable. Finally, the tamper-proof camera in Treatment 2 provide an extra tool to allow these communities to conduct their evaluation of teacher presencebased on more objective evidence.

Empowerment, however, is unlikely to affect outcomes if there is no willingness to exercise the power when required. This could arise for two reasons. First, different societies may exhibit different willingness to punish violations to an agreed standard (Ensminger and Henrich, eds, 2014). Societies that are unwilling to punish will not be able to effectively use performance-pay tools to induce accountability among teachers. Second, if evaluations are based on standards that are negotiable, teachers could try to influence these evaluations instead of improving their performance. In our experiment, the tamper-proof camera and the focus on the presence indicator in Treatment 2 provided boundaries on the negotiability of the standards. We show that both of these factors are important determinants of the effectiveness of the interventions.

To test the role of norms, we conducted a lab-in-the-field experiment at baseline to measure the different communities' willingness to punish. 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.[17] We only conducted this experiment in 182 schools that were randomly selected from the 270 participating schools. Appendix D provides detailed description of this lab-in-the-field experiment and how we construct the school-level willingness-to-punish measure. Using this continuous measure, we then categorized schools into those with above-/below-median punishment norm.

Table 13 presents the heterogenous impact of our interventions by the baseline punishment norm.

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

For the analysis, we limit the sample to TSA teachers and students who had a TSA teacher at baseline or endline who would have been directly affected by the performance pay scheme.[18] Columns 1–3 show that punishment norms strongly predicts the effectiveness of the treatment on TSA teachers' presence (and school-related activities) in Treatment 2, but not in the other treatments. Nonetheless, columns 4–5 suggest that the ability to increase teacher presence does not always translate into marginal improvements in their students' learning outcomes.

While the absence of a heterogenous impact by punishment norm for Treatment 1 (which lacks a performance-pay component) makes sense, its absence for Treatment 3 is puzzling. We argue that the more subjective evaluation standards in Treatment 3 — which opens up a room for teachers to negotiate for a higher score — may explain a part of this puzzle. A qualitative study in our treated schools suggest that teachers in Treatment 3 often questioned the validity of these evaluations, which were often conducted by parents who were less educated than these teachers. The relatively higher stature of teachers in the community put them in a position to pressure user committee members to improve their score. Indeed, we find corroborating evidence from our survey of user committee members: Table 14 shows that user committee members in Treatment 3 schools are more likely to be pressured to increase the evaluation scores and received threats for a low score than those in the other treated schools.

7 Cost Effectiveness

The investment cost of implementation for project facilitators was at USD 5,058 per school or USD 40 per student, which includes all costs made over the period of this study.[19] The cost was USD 506 per school or USD 4 per student higher for Group 2 schools, to cover for the purchase of mobile phones and the maintenance of the application. After one year of intervention, Group 2 improved learning outcomes by 0.2 standard deviation, at USD 44 per student. This means it costs USD 22 per student per 0.1 standard deviation increase. Details on the cost calculation can be found in Appendix Section E.

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

[18] Appendix Table A.8 show the results for non-TSA teachers and students who was never taught by a TSA teacher during the intervention period.

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

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

teacher incentive intervention in Kenya costs USD 16 per 0.1 standard deviation increase, while in India it costs USD 1 per 0.1 standard deviation increase.

8 Conclusion

We present results from a set of interventions to improve education quality in public schools in remote areas of Indonesia through a combination of community monitoring and pay for performance tied to a remote area allowance. While we find that all treatment lead to learning improvements, the treatment which combines community monitoring with a a simple pay-for-performance scheme based on absence aided by a tamper-proof camera worked best in improving learning outcomes. At the same time, the treatment which relied on social accountability only (without teacher incentives) failed to increase teacher effort, suggesting that a strategic approach, as conceptualized by Fox (2015), that integrates social accountability with measures to increase public sector responsiveness outperforms a tactical approach that relies of information alone to generate collective action and influence public sector performance. We demonstrate that such a strategic approach can be implemented using a government financed teacher allowance.

Our comparison of two different pay for performance mechanisms shows that a simple contract based on monitoring presence only works better than a more comprehensive, less well specified one. This is an important question that arises in many labor contracts (Baker et al., 1994; Khan et al., 2016). While we do find some evidence of teachers shifting focus to tasks which are easy to observe, the overall impact on learning is the highest for the contract that puts the strongest incentive on teacher presence. This simple contract also leads to less conflict at the community level. The study suggests that a simple transparent rule, even though it only targets an incomplete measure of performance, can work better than a comprehensive evaluation which is more prone to subjectivity. This finding is of relevance to many developing countries where governments face difficulty in holding teachers accountable.

References

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

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

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

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.

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

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

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

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.

Chu-Chang, Mae, 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.

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

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

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

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

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

Finan, F., B.A. Olken, and R. Pande, "The Personnel Economics of the Developing State," in "Handbook of Economic Field Experiments," Vol. 2, Elsevier, 2017, pp. 467–514.

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

Fryer Jr., Roland G., Steven Levitt, John List, and Sally Sadoff, "Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment," Unpublished manuscript April 2018.

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.

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 develop-

ment, Washington, D.C: World Bank, 2013.

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.

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.

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.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani, "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1627–1673.

OECD, ed., *What students know and can do: student performance in mathematics, reading and science* number Organisation for Economic Co-operation and Development, Programme for International Student Assessment ; Vol. I. In 'PISA 2012 results.', rev. ed., febr. 2014 ed., Paris: OECD, 2014. OCLC: 931534176.

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.

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.

Pugatch, Todd and Elizabeth Schroeder, "Teacher pay and student performance: evidence from the Gambian hardship allowance," *Journal of Development Effectiveness*, April 2018, *10* (2), 249–276.

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.

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

Stern, Jonathan and Lee Nordstrum, "Indonesia 2014: The National Early Grade Reading Assessment (EGRA) and Snapshot of School Management Effectiveness (SSME) Survey - Report of Findings," Technical Report, RTI International June 2014.

Suryahadi, Asep and Prio Sambodho, "Assessment of Policies to Improve Teacher Quality and Reduce Teacher Absenteeism," SMERU Working Paper, SMERU 2013.

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

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

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

World Bank, "Indonesia - Spending More or Spending Better : Improving Education Financing in In-

donesia," Technical Report, World Bank, Jakarta 2013.

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

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

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

9 Tables

Table 2: Baseline Summary Statistics

	Mean	Standard deviation	N
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			
Male	0.51	0.50	2329
Age	37.49	10.41	2326
Married	0.84	0.37	2331
Share with [...] education			
...less than high school	0.01	0.08	2329
...high school	0.28	0.45	2329
...more than high school	0.71	0.45	2329
Share with [...] status			
...civil servant	0.50	0.50	1952
...certified	0.21	0.40	1952
...TSA-receiving	0.18	0.38	2331
Share of teachers observed to be:			
...present in school	0.81	0.40	1952
...working when in school	0.74	0.44	1952
...teaching when in class	0.61	0.49	1952
(Self-reported) hours spent monthly:			
...preparing lessons	17.51	16.65	2021
...teaching curricular materials	64.84	22.05	2021
...assessing student work	12.85	11.61	2021
...teaching extra-curricular materials	4.22	6.04	2021
...on off-own-school employment	17.12	32.64	2048
Panel C. Parent characteristics			
Mother is the respondent	0.47	0.50	4218
Respondent's age	39.70	8.59	4218
Education expenditure (Rp.)	367,010	233,031	4338
Accompanied learning hours in the previous week	2.48	2.94	4338
Paid tutor	0.004	0.06	4338
Number of meetings with teacher on:			
...learning	1.76	4.38	4338
...other issues	1.06	3.02	4338
Share of parents who believe:			
School quality is good or very good	0.90	0.30	4338
Teacher absence is a main problem	0.22	0.42	4338

Table 3: Impact on Attrition and Entry of Students and Teachers

	Student			Teacher	
	Attrition		Entry	Attrition	Entry
	(1)	(2)	(3)	(4)	(5)
Treatment 1	-0.008	-0.010	0.007	-0.015	-0.004
	(0.006)	(0.008)	(0.012)	(0.016)	(0.020)
Treatment 2	-0.009	-0.012	-0.000	0.005	-0.001
	(0.006)	(0.008)	(0.009)	(0.018)	(0.019)
Treatment 3	-0.014	-0.017	0.007	-0.004	0.031
	(0.005)***	(0.007)**	(0.009)	(0.018)	(0.022)
Treatment 1 × Above-median student		0.003			
		(0.008)			
Treatment 2 × Above-median student		0.006			
		(0.007)			
Treatment 3 × Above-median student		0.006			
		(0.007)			
Control group mean	0.08	0.08	0.20	0.13	0.16
Test of equality (P-val)					
Treatment 1 v. 2	0.954	0.836	0.497	0.240	0.859
Treatment 2 v. 3	0.193	0.322	0.369	0.642	0.116
Treatment 1 v. 3	0.209	0.296	0.987	0.478	0.094
R2	0.495	0.493	0.550	0.196	0.313
Observations	26613	26613	31022	2292	2331
Strata FE	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes

Notes: For *students*: Individual control variables are sex, age dummies, both parents education, and dummy variables for individuals with missing controls. Above-median students are those whose average standardized scores of both subjects are above their class median. For *teachers*: Individual control variables are sex, age, age squared, marriage status, civil servant status, teacher certification status, and dummy variables for individuals with missing controls. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 4: Impact on Student Learning Outcomes

	Raw Score				Standardized Score			
	Bahasa Indonesia		Mathematics		Bahasa Indonesia		Mathematics	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment 1	1.297	1.710	1.159	1.506	0.070	0.085	0.055	0.071
	(0.917)	(0.731)**	(0.925)	(0.826)*	(0.044)	(0.035)**	(0.044)	(0.040)*
Treatment 2	3.842	3.641	3.821	4.066	0.185	0.174	0.187	0.200
	(0.987)***	(0.738)***	(1.018)***	(0.849)***	(0.047)***	(0.035)***	(0.049)***	(0.040)***
Treatment 3	1.534	2.213	1.273	1.818	0.086	0.110	0.066	0.093
	(0.868)*	(0.679)***	(0.913)	(0.797)**	(0.041)**	(0.032)***	(0.044)	(0.038)**
Control group mean		47.13		47.03		0.00		0.00
Test of equality (P-val)								
Treatment 1 v. 2	0.013	0.017	0.011	0.005	0.018	0.019	0.008	0.003
Treatment 2 v. 3	0.017	0.061	0.016	0.014	0.028	0.070	0.017	0.014
Treatment 1 v. 3	0.797	0.512	0.905	0.719	0.713	0.497	0.806	0.602
Randomization Inference								
(P-value, N = 1000)								
Treatment 1	0.193	0.025	0.244	0.083	0.142	0.025	0.202	0.093
Treatment 2	0.002	0.000	0.003	0.000	0.000	0.000	0.002	0.000
Treatment 3	0.119	0.006	0.198	0.031	0.060	0.002	0.149	0.018
R2	0.879	0.888	0.881	0.886	0.290	0.334	0.276	0.301
Observations	31022	31022	31022	31022	31022	31022	31022	31022
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	No	Yes	No	Yes	No	Yes	No	Yes

Notes: Control variables include sex, age dummies, both parents' education, baseline outcome, dummy variables for individuals with missing controls, school-level mean scores, and the school's public/private status. 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 5: Heterogeneous Impact by Gender, Baseline Grade Level, and Baseline Learning Outcomes

	Male		[. . .] at Baseline			
			Grades 4–5		Above-Median Score	
	Indonesian	Math	Indonesian	Math	Indonesian	Math
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment 1	0.070	0.055	0.084	0.095	0.101	0.046
	(0.040)*	(0.043)	(0.045)*	(0.052)*	(0.037)***	(0.045)
Treatment 2	0.174	0.191	0.192	0.232	0.116	0.147
	(0.039)***	(0.044)***	(0.043)***	(0.054)***	(0.039)***	(0.045)***
Treatment 3	0.116	0.073	0.105	0.101	0.074	0.063
	(0.038)***	(0.041)*	(0.043)**	(0.052)*	(0.037)**	(0.042)
Subgroup: [. . .]	-0.220	-0.090	0.113	0.074	0.197	0.149
	(0.024)***	(0.023)***	(0.038)***	(0.042)*	(0.026)***	(0.028)***
. . . × Treatment 1	0.029	0.030	-0.007	-0.068	-0.036	0.047
	(0.031)	(0.036)	(0.053)	(0.060)	(0.037)	(0.038)
. . . × Treatment 2	0.000	0.017	-0.109	-0.123	0.066	0.068
	(0.032)	(0.032)	(0.048)**	(0.060)**	(0.036)*	(0.037)*
. . . × Treatment 3	-0.011	0.038	-0.062	-0.071	0.016	0.020
	(0.033)	(0.035)	(0.051)	(0.060)	(0.038)	(0.038)
Observations	31022	31022	24719	24719	24700	24700
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Control variables include sex, age dummies, both parents' education, baseline outcome, dummy variables for individuals with missing controls, school-level mean scores, and the school's public/private status. Above-median students are those whose average standardized scores of both subjects are above their class median. Columns 3–6 excluded first-grade students at endline. In addition to excluding first-grade students at endline, Columns 5–6 also dropped students with missing baseline learning outcomes. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 6: Heterogeneous Impact by TSA Status of Endline and Baseline Teachers

	Indonesian (1)	Math (2)	Indonesian (3)	Math (4)	Indonesian (5)	Math (6)
Treatment 1	0.052	0.047	0.084	0.023	0.049	0.033
	(0.046)	(0.056)	(0.044)*	(0.056)	(0.058)	(0.073)
Treatment 2	0.106	0.157	0.155	0.147	0.108	0.140
	(0.052)**	(0.056)***	(0.049)***	(0.059)**	(0.066)	(0.078)*
Treatment 3	0.060	0.014	0.056	0.023	0.032	0.022
	(0.045)	(0.050)	(0.044)	(0.053)	(0.054)	(0.068)
TSA teacher at baseline						
... × Treatment 1	0.056	0.038				
	(0.060)	(0.066)				
... × Treatment 2	0.073	0.046				
	(0.058)	(0.068)				
... × Treatment 3	0.037	0.102				
	(0.054)	(0.063)				
TSA teacher at endline						
... × Treatment 1			-0.002	0.076		
			(0.056)	(0.069)		
... × Treatment 2			-0.013	0.057		
			(0.055)	(0.073)		
... × Treatment 3			0.042	0.084		
			(0.051)	(0.060)		
TSA teacher only at baseline						
... × Treatment 1					0.099	-0.009
					(0.086)	(0.081)
... × Treatment 2					0.126	0.027
					(0.089)	(0.088)
... × Treatment 3					0.078	0.014
					(0.079)	(0.087)
TSA teacher only at endline						
... × Treatment 1					0.012	0.032
					(0.081)	(0.091)
... × Treatment 2					0.003	0.037
					(0.081)	(0.095)
... × Treatment 3					0.067	-0.018
					(0.076)	(0.085)
TSA teacher at base- and endline						
... × Treatment 1					0.043	0.082
					(0.077)	(0.097)
... × Treatment 2					0.048	0.079
					(0.076)	(0.101)
... × Treatment 3					0.065	0.133
					(0.066)	(0.085)
Observations	24719	24719	24719	24719	24719	24719
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Student controls	Yes	Yes	Yes	Yes	Yes	Yes
Interacted level variables	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Control variables include sex, age dummies, both parents' education, baseline outcome, dummy variables for individuals with missing controls, school-level mean scores, and the school's public/private status. "New students" are students with no student test scores at baseline, including the whole first-graders at endline. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 7: Impact on Teacher Behaviors in School

	Present at school (1)	Working at school (2)	Teaching in class (3)	Present at school (4)	Working at school (5)	Teaching in class (6)
Treatment 1	0.020	0.050	0.040	0.035	0.056	0.017
	(0.026)	(0.029)*	(0.031)	(0.042)	(0.044)	(0.049)
Treatment 2	0.042	0.057	0.035	0.035	0.009	-0.054
	(0.026)	(0.029)*	(0.030)	(0.043)	(0.048)	(0.051)
Treatment 3	-0.028	-0.034	-0.025	-0.071	-0.109	-0.130
	(0.029)	(0.032)	(0.033)	(0.048)	(0.052)**	(0.055)**
TSA-receiving teacher				0.051	0.020	-0.006
				(0.037)	(0.039)	(0.044)
... × Treatment 1				-0.027	-0.010	0.038
				(0.053)	(0.055)	(0.058)
... × Treatment 2				0.006	0.075	0.141
				(0.051)	(0.056)	(0.066)**
... × Treatment 3				0.063	0.116	0.166
				(0.053)	(0.055)**	(0.062)***
Control group mean	0.82	0.77	0.61	0.82	0.77	0.61
Test of equality (P-val)						
Treatment 1 v. 2	0.392	0.786	0.848	1.000	0.319	0.166
Treatment 2 v. 3	0.009	0.003	0.046	0.030	0.032	0.181
Treatment 1 v. 3	0.071	0.005	0.031	0.025	0.001	0.007
R2	0.845	0.797	0.660	0.846	0.798	0.663
Observations	1954	1954	1954	1954	1954	1954
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Results are based on the data from the teacher absence survey (TAS) instrument. Each of columns capture whether the teacher was observed: (1) to be at school; (2) working (instead of, e.g., taking a break); (3) teaching and performing other learning-related activities when observed in class. Regressions are estimated using linear probability. Baseline school-level controls include: the total number of teachers, number of civil servant teachers, number of certified teachers, number of students, and a private school dummy. Baseline individual-level teacher controls include: age, age-squared, gender, marital status, civil servant and certification statuses, and lagged dependent variable. In addition, we include dummy variables for missing individual- and school-level controls and lagged variables. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 8: Impact on Teachers' Self-Reported Time Allocation (Hours per Month)

	For Own School				Work Outside Own School	For Own School				Work Outside Own School
	Overall academic	Curricular teaching	Assessment	Other Academic		Overall academic	Curricular teaching	Assessment	Other Academic	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment 1	-0.011	3.226	-1.378	-1.886	-4.237	4.500	5.858	-0.823	-0.617	-2.098
	(2.851)	(1.867)*	(0.745)*	(1.322)	(2.051)**	(3.867)	(2.551)**	(1.072)	(1.786)	(2.536)
Treatment 2	2.159	4.268	-1.707	-0.377	-2.466	2.612	4.423	-1.252	-0.641	-2.881
	(2.831)	(1.713)**	(0.807)**	(1.419)	(2.185)	(3.728)	(2.242)**	(1.248)	(1.791)	(2.815)
Treatment 3	1.140	1.623	-0.433	-0.069	-3.731	1.838	2.189	-0.460	0.038	-3.503
	(2.955)	(1.763)	(0.877)	(1.478)	(1.995)*	(3.614)	(2.235)	(1.149)	(1.875)	(2.512)
TSA-receiving teacher						6.980	5.099	0.686	1.228	3.685
						(3.655)*	(2.356)**	(1.259)	(1.690)	(3.773)
... × Treatment 1						-9.283	-5.517	-1.136	-2.535	-4.342
						(4.957)*	(3.184)*	(1.465)	(2.399)	(3.994)
... × Treatment 2						-1.573	-0.814	-0.938	0.372	0.456
						(4.504)	(2.836)	(1.634)	(2.190)	(4.097)
... × Treatment 3						-1.904	-1.516	0.001	-0.295	-0.651
						(4.184)	(2.802)	(1.477)	(2.107)	(4.092)
Control group mean	99.00	62.54	13.90	22.56	19.66	99.00	62.54	13.90	22.56	19.66
Test of equality (P-val)										
Treatment 1 v. 2	0.370	0.518	0.633	0.193	0.357	0.617	0.578	0.699	0.988	0.769
Treatment 2 v. 3	0.675	0.081	0.122	0.808	0.511	0.816	0.310	0.495	0.686	0.814
Treatment 1 v. 3	0.644	0.314	0.210	0.134	0.751	0.457	0.134	0.721	0.698	0.540
R2	0.904	0.909	0.587	0.633	0.400	0.904	0.909	0.587	0.634	0.401
Observations	2021	2021	2021	2021	2048	2021	2021	2021	2021	2048
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: "Overall academic" (columns 1 and 6) include teaching preparation, curricular teaching, extra-curricular activities, and assessments both during and after school hours. "Other academic" (columns 4 and 9) include preparation and extra-curricular activities. Baseline school-level controls include: the total number of teachers, number of civil servant teachers, number of certified teachers, number of students, and a private school dummy. Baseline individual-level teacher controls include: age, age-squared, gender, marital status, civil servant and certification statuses, and lagged dependent variable. In addition, we include dummy variables for missing individual- and school-level controls and lagged variables. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 9: Impact on Parental Investment in Education

	Education Expenditures	Accompanied Learning Hours	Paid Tutor	Parent/Teacher Meetings			
				Teacher Survey		Parent Survey	
				All meetings	Formal	Discuss [...]	
						Learning	Others
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment 1	9591.1	0.313	-0.0002	0.0173	0.224	0.818	0.808
	(13223.4)	(0.193)	(0.004)	(0.361)	(0.0698)***	(0.285)***	(0.270)***
Treatment 2	29397.8	0.335	0.016	-0.163	0.254	1.109	0.942
	(13518.5)**	(0.192)*	(0.006)***	(0.368)	(0.0705)***	(0.289)***	(0.275)***
Treatment 3	9426.4	0.306	0.005	-0.0947	0.175	0.948	1.157
	(13998.7)	(0.194)	(0.004)	(0.393)	(0.0655)***	(0.341)***	(0.330)***
Control group mean	324153.7	2.458	0.003	2.543	0.358	1.499	0.734
Test of equality (P-val)							
Treatment 1 v. 2	0.138	0.902	0.010	0.431	0.696	0.364	0.709
Treatment 2 v. 3	0.155	0.876	0.066	0.784	0.269	0.658	0.572
Treatment 1 v. 3	0.990	0.970	0.232	0.685	0.477	0.705	0.348
R2	0.733	0.427	0.061	0.288	0.268	0.251	0.202
Observations	5386	5394	5401	2021	2021	5401	5401
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Education expenditures are for the previous school year. "Accompanied learning hours" is the number of hours the child received support for education work from someone in the household in the past week. All regressions control for the baseline value of the dependent variable, in interaction with a dummy variable indicating if the baseline value is missing. Individual controls include the child's sex and age, the education level of both parents, a dummy indicating whether the mother is the respondent, and the age and age-squared of the respondent. Columns 4–5 are the number of meetings with parents as reported by teachers. Estimates in these columns include baseline school-level and teacher-level control variables. Baseline school-level controls include: the total number of teachers, number of civil servant teachers, number of certified teachers, number of students, and a private school dummy. Individual-level teacher controls include: age, age-squared, gender, marital status, civil servant and certification statuses, and lagged dependent variable. Columns 6–7 are the number of meetings where parents discuss learning and other topics with teachers as reported by parents. Individual controls include the child's sex and age, the education level of both parents, the age and age-squared of the respondent, and dummy variables of whether the respondent is a user committee member and whether s/he is the mother of a child in the school. It also includes dummy variables for each control variable indicating whether it is missing. The regressions also control for whether the school is a public or private school. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 10: Official Engagement and Principal Supervision

	Official Supervision		Principal Supervision		
	Meeting	Inspector Supervision	In-class Observation	Evaluation	
				Ever	Routine
	(1)	(2)	(3)	(4)	(5)
Treatment 1	0.316	0.250	0.082	0.048	0.089
	(0.466)	(0.321)	(0.038)**	(0.034)	(0.043)**
Treatment 2	0.880	0.620	0.080	0.096	0.145
	(0.461)*	(0.320)*	(0.034)**	(0.031)***	(0.041)***
Treatment 3	-0.060	0.380	0.092	0.071	0.122
	(0.470)	(0.325)	(0.036)**	(0.031)**	(0.042)***
Control group mean	2.24	1.42	0.67	0.76	0.45
Test of equality (P-val)					
Treatment 1 v. 2	0.216	0.244	0.942	0.137	0.163
Treatment 2 v. 3	0.042	0.452	0.696	0.391	0.521
Treatment 1 v. 3	0.412	0.684	0.772	0.478	0.404
R2	0.663	0.643	0.758	0.822	0.567
Observations	270	270	2021	2021	2021
Strata FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes

Notes: All outcomes are recorded with respect to the current academic year. Columns 1–2 are reported by the school principal. Column 1 is the number of meetings with subdistrict education officials. Column 2 is the number of monitoring visits by school inspectors and/or for private schools, a representative of the private foundation. The school-level controls include: the the total number of teachers, number of civil servant teachers, number of certified teachers, number of students, and a private school dummy and distance to the subdistrict UPTD office. Outcomes in columns 3–7 come from the teacher survey. Column 3 captures whether teachers are observed or monitored when teaching and columns 4–5 are whether teachers receive formal evaluation. The regressions also control for whether the school is a public or private school. All regressions include dummies for missing control variables. Standard errors for estimates in columns 3–9 are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 11: Impact on Teachers' Subjective Assessments

	Teacher Satisfaction of			Share Teachers Satisfied with Current Job	Teacher Satisfaction of			Share Teachers Satisfied with Current Job
	Appreciation from		Salary		Appreciation from		Salary	
	District	Village			District	Village		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment 1	0.112	0.152	0.137	0.026	0.103	0.180	0.245	0.023
	(0.058)*	(0.054)***	(0.052)***	(0.020)	(0.081)	(0.075)**	(0.073)***	(0.028)
Treatment 2	0.245	0.200	0.216	0.064	0.301	0.230	0.322	0.117
	(0.057)***	(0.053)***	(0.052)***	(0.020)***	(0.081)***	(0.075)***	(0.073)***	(0.028)***
Treatment 3	0.279	0.260	0.310	0.055	0.322	0.219	0.367	0.080
	(0.058)***	(0.054)***	(0.052)***	(0.020)***	(0.079)***	(0.074)***	(0.072)***	(0.027)***
TSA-receiving teacher					0.183	0.007	0.307	0.043
					(0.090)**	(0.084)	(0.081)***	(0.031)
... × Treatment 1					-0.000	-0.055	-0.237	0.000
					(0.114)	(0.106)	(0.103)**	(0.040)
... × Treatment 2					-0.125	-0.057	-0.235	-0.106
					(0.115)	(0.107)	(0.104)**	(0.040)***
... × Treatment 3					-0.101	0.080	-0.139	-0.054
					(0.114)	(0.106)	(0.103)	(0.039)
Control group mean	-0.08	-0.01	-0.06	0.85	-0.08	-0.01	-0.06	0.85
Control-group unstandardized baseline mean (7-point scale)	4.5	5.0	4.1		4.5	5.0	4.1	
Test of equality (P-val)								
Treatment 1 v. 2	0.019	0.360	0.128	0.056	0.016	0.513	0.299	0.001
Treatment 2 v. 3	0.552	0.254	0.064	0.630	0.801	0.886	0.536	0.176
Treatment 1 v. 3	0.003	0.040	0.001	0.147	0.007	0.598	0.093	0.042
R2	0.217	0.159	0.358	0.900	0.220	0.160	0.364	0.900
Observations	2021	2021	2021	2021	2021	2021	2021	2021
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Columns 1–3 and 5–7 regress standardized satisfaction outcomes from the teacher survey (from a 7-point Likert scale). Columns 4 and 8 outcome is the share of teachers who are either "satisfied" or "very satisfied" with their job in the school. Controls include school-level and teacher-level baseline variables. Baseline school-level controls include: the total number of teachers, number of civil servant teachers, number of certified teachers, number of students, and a private school dummy. Individual-level teacher controls include: age, age-squared, gender, marital status, civil servant and certification statuses, and lagged dependent variable. All regressions also include dummies for missing control variables. Standard errors for estimates are clustered at the school level. */**/*** denotes significance at the 10/5/1 percent level.

Table 12: Impact on Parents' Subjective Assessments

	Share of Parents Agreeing ...	
	School is Good/ Very Good (1)	Teacher Absence a Problem (2)
Treatment 1	0.048	-0.056
	(0.018)***	(0.030)*
Treatment 2	0.051	-0.071
	(0.018)***	(0.031)**
Treatment 3	0.051	-0.080
	(0.018)***	(0.031)**
Control group mean	0.91	0.27
Test of equality (P-val)		
Treatment 1 v. 2	0.811	0.570
Treatment 2 v. 3	0.989	0.723
Treatment 1 v. 3	0.809	0.348
R2	0.950	0.288
Observations	5285	5401
Strata FE	Yes	Yes
Basic Controls	Yes	Yes

Notes: Column 1 outcome is the share of parent respondents who considered the school "good" or "very good", while column 2 is the share of parents who considered teacher absence one of the most important education problems in the community. Individual controls include the child's sex and age, the education level of both parents, the age and age-squared of the respondent, and dummy variables of whether the respondent is a user committee member and whether s/he is the mother of a child in the school. All regressions also control for whether the school is a public or private school and include dummies for missing control variables. Standard errors are clustered at the school level. */**/*** denotes significance at the 10/5/1 percent level.

Table 13: Heterogenous Impact by School-Level Punishment Norms

	Teacher Behavior			Student Learning	
	Present at school	Working at school	Teaching in class	Indonesian	Mathematics
	(1)	(2)	(3)	(4)	(5)
Treatment 1	0.018	0.079	0.043	0.038	0.080
	(0.045)	(0.049)	(0.067)	(0.070)	(0.080)
Treatment 2	-0.006	0.043	0.059	0.097	0.196
	(0.047)	(0.050)	(0.062)	(0.068)	(0.085)**
Treatment 3	0.023	0.049	0.032	0.007	0.074
	(0.057)	(0.060)	(0.072)	(0.063)	(0.088)
Above-Median Punishment	-0.061	-0.065	-0.092	-0.175	-0.104
Norm	(0.069)	(0.068)	(0.074)	(0.075)**	(0.088)
... × Treatment 1	0.049	0.018	0.053	0.067	0.015
	(0.087)	(0.086)	(0.097)	(0.108)	(0.115)
... × Treatment 2	0.185	0.169	0.219	0.193	0.034
	(0.090)**	(0.094)*	(0.092)**	(0.107)*	(0.121)
... × Treatment 3	-0.010	-0.036	0.034	0.230	0.122
	(0.094)	(0.090)	(0.096)	(0.102)**	(0.113)
Control group mean	0.84	0.78	0.57	-0.01	-0.05
R2	0.865	0.814	0.674	0.349	0.306
Observations	827	827	827	17190	17190
Strata FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes

Notes: Control variables are identical to those used for baseline learning outcome and teacher behavior estimates. 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. The analytical samples include TSA teachers (columns 1–3) and students who have had a TSA teacher at baseline, endline, or both (columns 4–5). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table 14: User Committee Reports of Pressure from School

	Intimidated	Pressure to Increase Score	Threats for Low Score
	(1)	(2)	(3)
Treatment 2	0.020	-0.0004	0.075
	(0.040)	(0.055)	(0.044)*
Treatment 3	0.060	0.121	0.166
	(0.040)	(0.055)**	(0.044)***
Constant	0.028	0.074	-0.001
	(0.028)	(0.039)*	(0.031)
T2 v. T3 equality (p-val)	0.320	0.030	0.040
Observations	202	202	202
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?" */**/*** denotes 10/5/1 percent significance levels

10 Figures

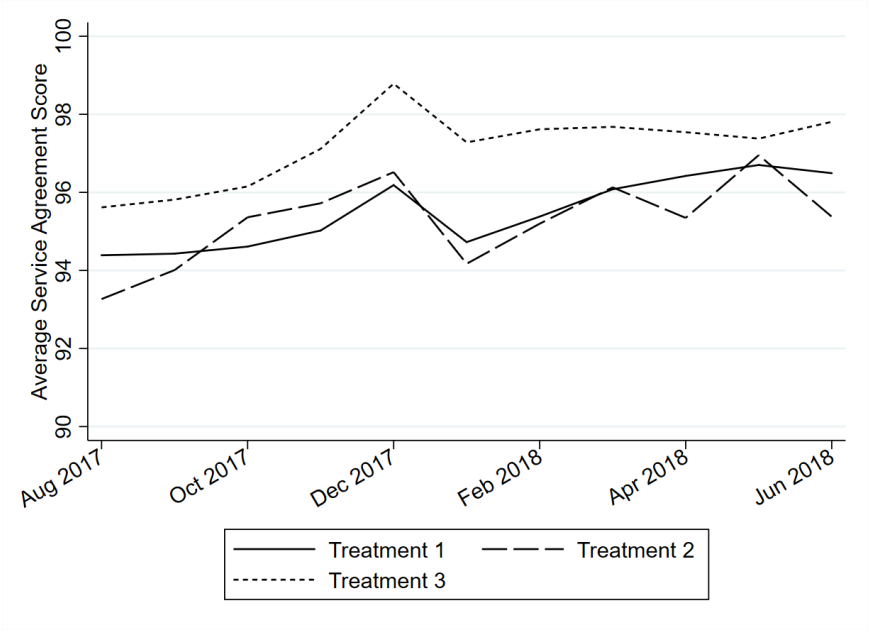
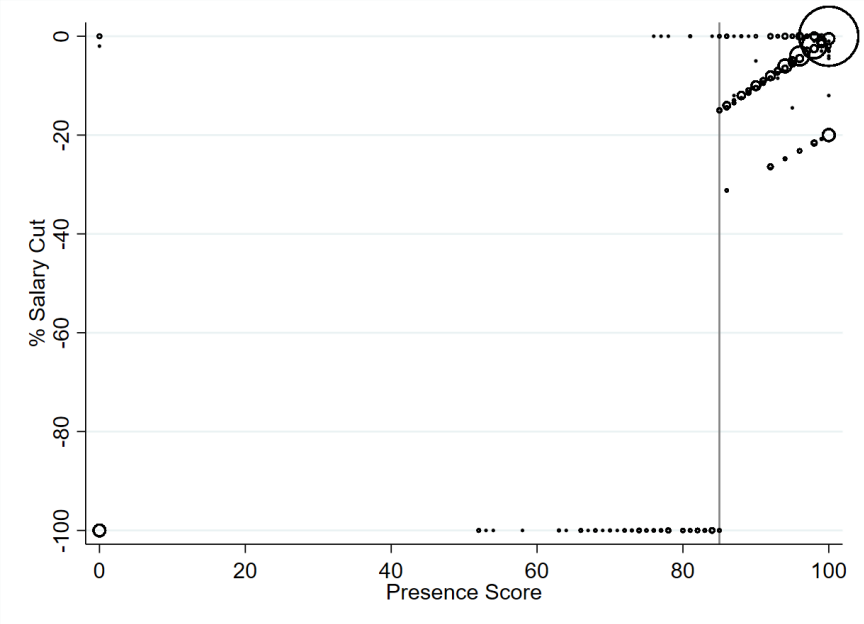


Figure 2: Average Scorecard Ratings Across Treatments



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 June 2018, excluding December 2017 when salaries were not cut.

Figure 3: Compliance of the 85 Percent Rule in Treatment 2

Appendix

A Additional Tables and Figures

I Tables

Table A.1: Balance Tables: Student Characteristics

	Mean (μ) (standard errors)				Differences = $\mu_{[...]} - \mu_{Control}$ (p-value)			Differences between $\mu_{[...]}$ and $\mu_{[...]}$ (p-value)		
	Control	Treatment 1	Treatment 2	Treatment 3	Treatment 1	Treatment 2	Treatment 3	Tr. 2 - Tr. 1	Tr. 3 - Tr. 1	Tr. 3 - Tr. 2
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)

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

Table A.2: Balance Tables: Teacher Characteristics

	Mean (μ) (standard errors)				Differences = $\mu_{[...]} - \mu_{Control}$ (p-value)			Differences between $\mu_{[...]}$ and $\mu_{[...]}$ (p-value)		
	Control	Treatment 1	Treatment 2	Treatment 3	Treatment 1	Treatment 2	Treatment 3	Tr. 2 - Tr. 1	Tr. 3 - Tr. 1	Tr. 3 - Tr. 2
Male	0.53	0.52	0.49	0.50	-0.01	-0.04	-0.03	-0.03	-0.02	0.01
	(0.50)	(0.50)	(0.50)	(0.50)	(0.63)	(0.12)	(0.23)	(0.25)	(0.46)	(0.68)
Age	37.39	37.65	37.55	37.37	0.26	0.16	-0.02	-0.10	-0.28	-0.18
	(10.69)	(10.33)	(10.31)	(10.35)	(0.68)	(0.81)	(0.97)	(0.88)	(0.66)	(0.78)
Married	0.84	0.84	0.83	0.84	-0.00	-0.02	-0.01	-0.01	-0.00	0.01
	(0.36)	(0.37)	(0.38)	(0.37)	(0.86)	(0.42)	(0.77)	(0.53)	(0.91)	(0.59)
Share with [...] education:										
...less than high school	0.01	0.01	0.00	0.01	-0.01	-0.01	-0.01	-0.00	-0.00	0.00
	(0.11)	(0.07)	(0.06)	(0.07)	(0.39)	(0.27)	(0.37)	(0.65)	(0.96)	(0.68)
...high school	0.30	0.29	0.26	0.27	-0.01	-0.04	-0.03	-0.03	-0.02	0.01
	(0.46)	(0.46)	(0.44)	(0.44)	(0.84)	(0.25)	(0.38)	(0.31)	(0.47)	(0.76)
...more than high school	0.69	0.70	0.74	0.73	0.01	0.05	0.04	0.03	0.02	-0.01
	(0.46)	(0.46)	(0.44)	(0.45)	(0.69)	(0.16)	(0.28)	(0.29)	(0.48)	(0.73)
Share with [...] status:										
...civil servant	0.49	0.49	0.49	0.51	-0.01	0.00	0.02	0.01	0.03	0.02
	(0.50)	(0.50)	(0.50)	(0.50)	(0.78)	(0.97)	(0.55)	(0.75)	(0.40)	(0.57)
...certified	0.20	0.19	0.21	0.23	-0.01	0.01	0.03	0.02	0.04	0.02
	(0.40)	(0.39)	(0.41)	(0.42)	(0.64)	(0.74)	(0.30)	(0.44)	(0.15)	(0.48)
...TSA-receiving	0.16	0.19	0.19	0.18	0.03	0.03	0.02	0.00	-0.01	-0.01
	(0.37)	(0.39)	(0.39)	(0.38)	(0.35)	(0.36)	(0.56)	(0.98)	(0.73)	(0.73)
Share of teachers observed to be:										
...present in school	0.79	0.78	0.81	0.84	-0.01	0.02	0.04	0.03	0.05*	0.03
	(0.41)	(0.41)	(0.39)	(0.37)	(0.74)	(0.61)	(0.19)	(0.39)	(0.09)	(0.41)
...working when in school	0.74	0.73	0.75	0.74	-0.01	0.02	0.00	0.02	0.01	-0.02
	(0.44)	(0.44)	(0.43)	(0.44)	(0.81)	(0.63)	(0.97)	(0.46)	(0.80)	(0.69)
...teaching when in class	0.61	0.62	0.62	0.61	0.01	0.01	0.00	0.01	-0.01	-0.01
	(0.49)	(0.49)	(0.49)	(0.49)	(0.81)	(0.69)	(0.99)	(0.87)	(0.82)	(0.70)
(Self-reported) hours spent monthly:										
...preparing lessons	17.83	16.42	17.47	18.26	-1.40	-0.36	0.43	1.05	1.84	0.79
	(18.32)	(16.21)	(16.09)	(15.94)	(0.38)	(0.82)	(0.78)	(0.46)	(0.18)	(0.55)
...teaching curricular materials	62.54	65.27	67.10	64.41	2.74	4.57**	1.87	1.83	-0.86	-2.69
	(22.64)	(23.11)	(21.54)	(20.74)	(0.21)	(0.04)	(0.35)	(0.37)	(0.64)	(0.15)
...assessing student work	13.90	12.26	11.88	13.35	-1.64*	-2.02**	-0.55	-0.38	1.08	1.47
	(13.25)	(10.08)	(10.90)	(11.91)	(0.09)	(0.04)	(0.61)	(0.62)	(0.23)	(0.10)
...teaching extra-curricular materials	4.73	3.78	4.13	4.25	-0.96*	-0.60	-0.48	0.36	0.48	0.12
	(7.08)	(5.63)	(5.42)	(5.88)	(0.09)	(0.27)	(0.40)	(0.47)	(0.37)	(0.82)
...on off-own-school employment	19.66	15.78	17.30	15.81	-3.87	-2.36	-3.84	1.51	0.03	-1.48
	(35.84)	(30.78)	(35.67)	(27.70)	(0.20)	(0.48)	(0.17)	(0.61)	(0.99)	(0.59)

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

Table A.3: Balance Tables: Parent Characteristics

	Mean (μ) (standard errors)				Differences = $\mu_{[...]} - \mu_{Control}$ (p-value)			Differences between $\mu_{[...]}$ and $\mu_{[...]}$ (p-value)		
	Control	Treatment 1	Treatment 2	Treatment 3	Treatment 1	Treatment 2	Treatment 3	Tr. 2 - Tr. 1	Tr. 3 - Tr. 1	Tr. 3 - Tr. 2
Mother is the respondent	0.46	0.47	0.46	0.51	0.01	0.01	0.06**	-0.01	0.05*	0.05*
	(0.50)	(0.50)	(0.50)	(0.50)	(0.70)	(0.85)	(0.04)	(0.85)	(0.09)	(0.06)
Respondent's age	39.68	39.14	39.37	39.12	-0.54	-0.30	-0.55	0.23	-0.02	-0.25
	(8.98)	(8.48)	(8.62)	(8.75)	(0.25)	(0.50)	(0.22)	(0.59)	(0.97)	(0.55)
Education expenditure (Rp.)	301,890	311,114	298,330	326,076	9,224	-3,561	24,186	-12,785	14,962	27,746
	(250,895)	(252,715)	(239,781)	(264,421)	(0.60)	(0.84)	(0.17)	(0.45)	(0.40)	(0.11)
Accompanied learning hours in the 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)
Paid tutor	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00
	(0.05)	(0.06)	(0.08)	(0.05)	(0.60)	(0.19)	(1.00)	(0.46)	(0.60)	(0.19)
Number of meetings with teacher on:										
...learning	1.88	1.88	1.75	1.82	0.00	-0.13	-0.06	-0.13	-0.06	0.07
	(12.49)	(4.51)	(3.89)	(5.27)	(1.00)	(0.78)	(0.91)	(0.62)	(0.84)	(0.81)
...other issues	0.87	1.10	1.04	1.19	0.23	0.17	0.32	-0.06	0.09	0.15
	(3.10)	(2.98)	(2.74)	(3.16)	(0.28)	(0.44)	(0.18)	(0.79)	(0.72)	(0.56)
Share of parents who believe:										
School quality is good or very good	0.89	0.88	0.92	0.91	-0.01	0.04	0.02	0.05**	0.03	-0.01
	(0.32)	(0.33)	(0.27)	(0.29)	(0.69)	(0.13)	(0.37)	(0.02)	(0.12)	(0.41)
Teacher absence is a main problem	0.23	0.22	0.24	0.20	-0.01	0.01	-0.03	0.02	-0.01	-0.03
	(0.42)	(0.41)	(0.42)	(0.40)	(0.66)	(0.89)	(0.40)	(0.51)	(0.60)	(0.27)

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

Table A.4: Selective Attrition and Entry of Students

	Attrition (1)	Entry (2)
Treatment 1	-0.023	0.013
	(0.024)	(0.054)
... × Above-median student	0.002	
	(0.008)	
... × Male	-0.003	0.001
	(0.008)	(0.012)
... × Age	0.001	-0.000
	(0.002)	(0.005)
... × Mother has post-primary education	0.002	-0.003
	(0.010)	(0.024)
... × Father has post-primary education	0.013	-0.021
	(0.010)	(0.022)
Treatment 2	-0.043	0.050
	(0.026)*	(0.053)
... × Above-median student	0.006	
	(0.008)	
... × Male	0.000	-0.022
	(0.009)	(0.013)*
... × Age	0.002	-0.005
	(0.002)	(0.005)
... × Mother has post-primary education	-0.003	0.015
	(0.010)	(0.023)
... × Father has post-primary education	0.015	-0.006
	(0.009)*	(0.020)
Treatment 3	-0.024	0.088
	(0.023)	(0.048)*
... × Above-median student	0.005	
	(0.007)	
... × Male	-0.006	-0.017
	(0.008)	(0.013)
... × Age	0.001	-0.008
	(0.002)	(0.004)*
... × Mother has post-primary education	0.001	-0.006
	(0.009)	(0.025)
... × Father has post-primary education	0.006	0.016
	(0.008)	(0.021)
Control group mean	0.08	0.20
R2	0.050	0.379
Observations	25483	30576
Strata FE	Yes	Yes
Individual Controls	Yes	Yes

Notes: Individual control variables are sex, age, both parents education, and dummy variables for individuals with missing controls. Above-median students are those whose average standardized scores of both subjects are above their class median. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table A.5: Selective Attrition and Entry of Teachers

	Attrition (1)	Entry (2)
Treatment 1	0.112	-0.296
	(0.358)	(0.401)
... × Male	0.057	-0.063
	(0.040)	(0.045)
... × Age	-0.013	0.014
	(0.020)	(0.021)
... × Age^2	0.000	-0.000
	(0.000)	(0.000)
... × Married	0.167	0.034
	(0.076)**	(0.074)
... × Civil servant	0.007	-0.020
	(0.052)	(0.064)
... × Certified	-0.003	-0.120
	(0.061)	(0.067)*
Treatment 2	0.400	-0.118
	(0.359)	(0.414)
... × Male	0.015	-0.102
	(0.038)	(0.043)**
... × Age	-0.030	0.007
	(0.019)	(0.022)
... × Age^2	0.000	-0.000
	(0.000)	(0.000)
... × Married	0.255	-0.013
	(0.075)***	(0.080)
... × Civil servant	-0.058	0.056
	(0.050)	(0.066)
... × Certified	-0.003	-0.142
	(0.067)	(0.076)*
Treatment 3	-0.130	0.419
	(0.415)	(0.425)
... × Male	0.012	-0.085
	(0.041)	(0.048)*
... × Age	-0.001	-0.021
	(0.023)	(0.022)
... × Age^2	-0.000	0.000
	(0.000)	(0.000)
... × Married	0.232	0.093
	(0.073)***	(0.076)
... × Civil servant	-0.001	0.006
	(0.051)	(0.069)
... × Certified	0.024	-0.089
	(0.063)	(0.073)
Control group mean	0.13	0.16
R2	0.209	0.326
Observations	2292	2331
Strata FE	Yes	Yes
Individual Controls	Yes	Yes

Notes: Individual control variables are sex, age, both parents education, and dummy variables for individuals with missing controls. Above-median students are those whose average standardized scores of both subjects are above their class median. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table A.6: Impact on Teachers' Self-Reported Time Allocation on Own-School Activities (Hours per Month)

	Prepara-tion	Curricular Teaching [...] School		Assess-ment	Extra-curricular	Prepara-tion	Curricular Teaching [...] School		Assess-ment	Extra-curricular
		During	After				During	After		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment 1	-0.924	1.324	1.858	-1.378	-0.957	-0.048	3.235	2.575	-0.823	-0.612
	(1.190)	(1.764)	(0.590)***	(0.745)*	(0.424)**	(1.595)	(2.395)	(0.876)***	(1.072)	(0.615)
Treatment 2	0.241	2.972	1.311	-1.707	-0.581	0.251	3.194	1.241	-1.252	-0.923
	(1.300)	(1.612)*	(0.519)**	(0.807)**	(0.436)	(1.667)	(2.130)	(0.658)*	(1.248)	(0.610)
Treatment 3	0.535	0.534	1.065	-0.433	-0.546	0.886	0.954	1.212	-0.460	-0.794
	(1.329)	(1.653)	(0.571)*	(0.877)	(0.460)	(1.676)	(2.080)	(0.727)*	(1.149)	(0.633)
TSA-receiving teacher						1.516	5.079	0.105	0.686	-0.335
						(1.547)	(2.358)**	(0.737)	(1.259)	(0.575)
... × Treatment 1						-1.815	-4.117	-1.393	-1.136	-0.622
						(2.123)	(2.993)	(1.168)	(1.465)	(0.759)
... × Treatment 2						-0.178	-0.932	0.113	-0.938	0.684
						(2.059)	(2.771)	(0.955)	(1.634)	(0.705)
... × Treatment 3						-0.813	-1.220	-0.299	0.001	0.528
						(1.962)	(2.707)	(0.960)	(1.477)	(0.746)
Control group mean	17.83	59.35	3.19	13.90	4.73	17.83	59.35	3.19	13.90	4.73
Test of equality (P-val)										
Treatment 1 v. 2	0.285	0.266	0.374	0.633	0.309	0.847	0.985	0.161	0.699	0.542
Treatment 2 v. 3	0.803	0.071	0.675	0.122	0.928	0.692	0.252	0.968	0.495	0.794
Treatment 1 v. 3	0.189	0.585	0.222	0.210	0.278	0.538	0.281	0.170	0.721	0.720
R2	0.572	0.917	0.245	0.587	0.422	0.572	0.918	0.246	0.587	0.423
Observations	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Baseline school-level controls include: the total number of teachers, number of civil servant teachers, number of certified teachers, number of students, and a private school dummy. Baseline individual-level teacher controls include: age, age-squared, gender, marital status, civil servant and certification statuses, and lagged dependent variable. In addition, we include dummy variables for missing individual- and school-level controls and lagged variables. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table A.7: Impact on Teachers' Self-Reported Time Allocation on Off-Own-School Activities (Hours per Month)

	Tutor	Agriculture	Other Work	Leisure	Tutor	Agriculture	Other Work	Leisure
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment 1	0.001	-2.458	-1.990	4.826	-0.092	-3.320	1.096	-1.738
	(0.099)	(1.546)	(1.383)	(3.381)	(0.141)	(2.207)	(1.605)	(4.379)
Treatment 2	0.047	-2.658	0.106	0.084	-0.066	-5.325	2.570	0.099
	(0.091)	(1.407)*	(1.662)	(3.395)	(0.139)	(2.157)**	(2.134)	(4.493)
Treatment 3	0.137	-1.574	-2.430	2.563	0.027	-4.876	1.046	1.862
	(0.104)	(1.356)	(1.470)*	(3.408)	(0.142)	(2.084)**	(1.667)	(4.433)
TSA-receiving teacher					-0.264	-4.618	8.885	-10.190
					(0.137)*	(2.496)*	(3.247)***	(5.651)*
... × Treatment 1					0.206	2.198	-6.838	13.490
					(0.160)	(2.994)	(3.071)**	(6.249)**
... × Treatment 2					0.242	5.585	-5.623	1.003
					(0.171)	(2.717)**	(3.527)	(6.338)
... × Treatment 3					0.242	6.990	-7.646	2.135
					(0.160)	(2.918)**	(3.323)**	(6.171)
Control group mean	0.09	13.74	6.19	360.98	0.09	13.74	6.19	360.98
Test of equality (P-val)								
Treatment 1 v. 2	0.647	0.880	0.119	0.113	0.867	0.293	0.470	0.664
Treatment 2 v. 3	0.363	0.349	0.072	0.414	0.525	0.797	0.469	0.673
Treatment 1 v. 3	0.186	0.458	0.668	0.435	0.408	0.395	0.975	0.395
R2	0.416	0.421	0.191	0.987	0.416	0.424	0.198	0.987
Observations	2021	2021	2021	2021	2021	2021	2021	2021
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Basic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

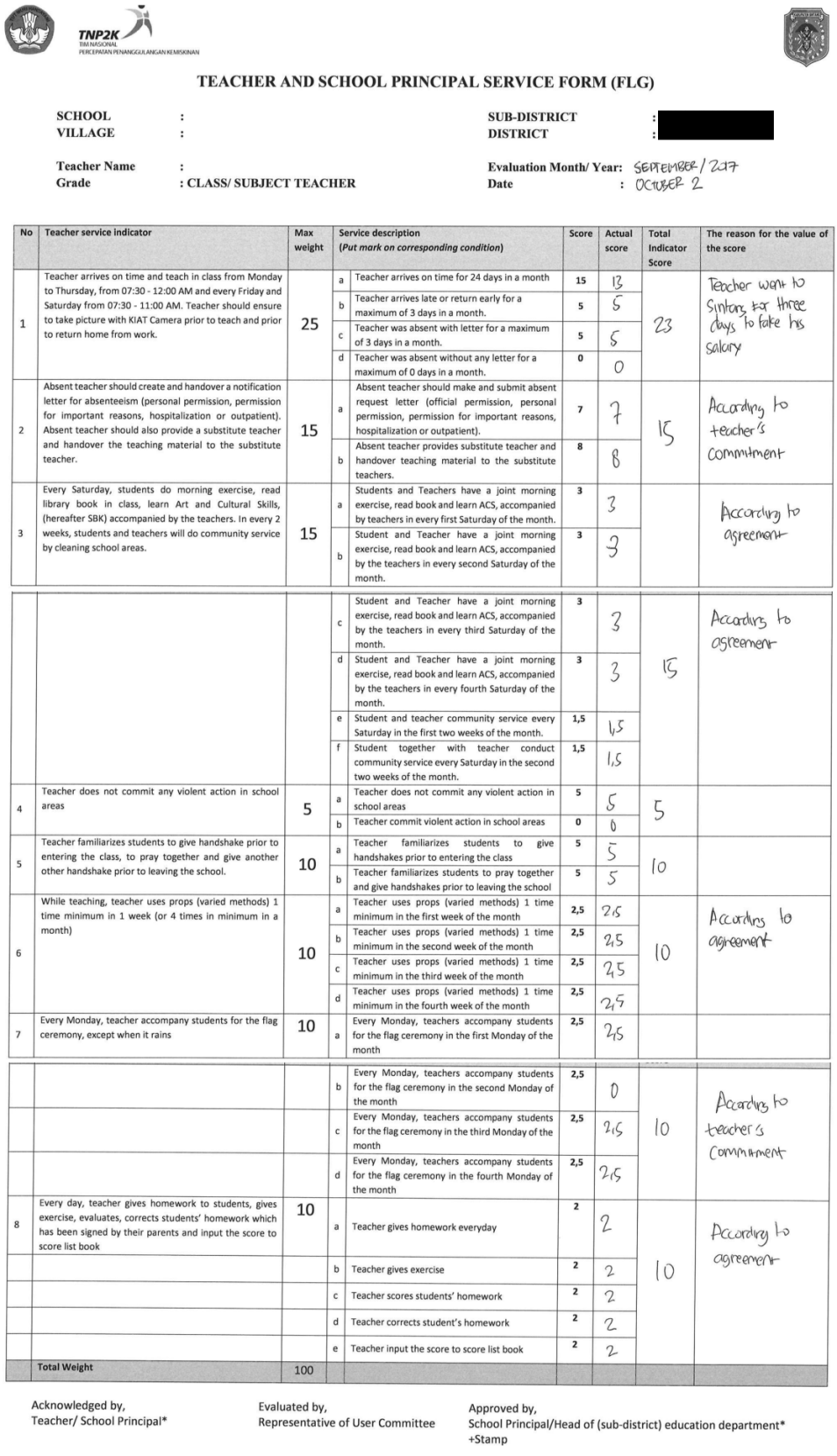
Notes: "Leisure" is calculated from subtracting the total on-school and off-school employment activities from 480 total available hours in a month. Baseline school-level controls include: the total number of teachers, number of civil servant teachers, number of certified teachers, number of students, and a private school dummy. Baseline individual-level teacher controls include: age, age-squared, gender, marital status, civil servant and certification statuses, and lagged dependent variable. In addition, we include dummy variables for missing individual- and school-level controls and lagged variables. Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

Table A.8: Heterogenous Impact by School-Level Punishment Norms for Non-TSA Teachers

	Teacher Behavior			Student Learning	
	Present at school	Working at school	Teaching in class	Indonesian	Mathematics
	(1)	(2)	(3)	(4)	(5)
Treatment 1	0.056	0.106	0.084	0.036	-0.218
	(0.064)	(0.068)	(0.074)	(0.109)	(0.107)**
Treatment 2	0.079	0.010	0.031	-0.101	-0.121
	(0.070)	(0.086)	(0.092)	(0.106)	(0.116)
Treatment 3	-0.059	-0.111	-0.247	-0.075	-0.215
	(0.075)	(0.090)	(0.086)***	(0.107)	(0.115)*
Above-Median Punishment	-0.015	-0.041	-0.051	-0.231	-0.330
Norm	(0.087)	(0.094)	(0.103)	(0.106)**	(0.130)**
... × Treatment 1	-0.031	-0.057	-0.057	0.297	0.581
	(0.100)	(0.111)	(0.132)	(0.158)*	(0.195)***
... × Treatment 2	0.031	0.113	-0.004	0.478	0.523
	(0.122)	(0.136)	(0.147)	(0.170)***	(0.172)***
... × Treatment 3	0.045	0.084	0.220	0.194	0.227
	(0.108)	(0.125)	(0.137)	(0.146)	(0.187)
Control group mean	0.81	0.78	0.68	0.19	0.21
R2	0.869	0.810	0.714	0.408	0.375
Observations	512	512	512	3832	3832
Strata FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes

Notes: Control variables are identical to those used for baseline learning outcome and teacher behavior estimates. 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. The analytical samples include non-TSA teachers (columns 1–3) and students who did not have a TSA teacher at both baseline and both (columns 4–5). Standard errors are clustered at the school level. */**/*** denotes 10/5/1 percent significance levels

II Figures



TNP2K
TIM NASIONAL PERCEPATAN PENANGGULANGAN KEMISKINAN

TEACHER AND SCHOOL PRINCIPAL SERVICE FORM (FLG)

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

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

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

Acknowledged by,
Teacher/ School Principal*

Evaluated by,
Representative of User Committee

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

Figure A.1: A Sample of the Community Scorecard

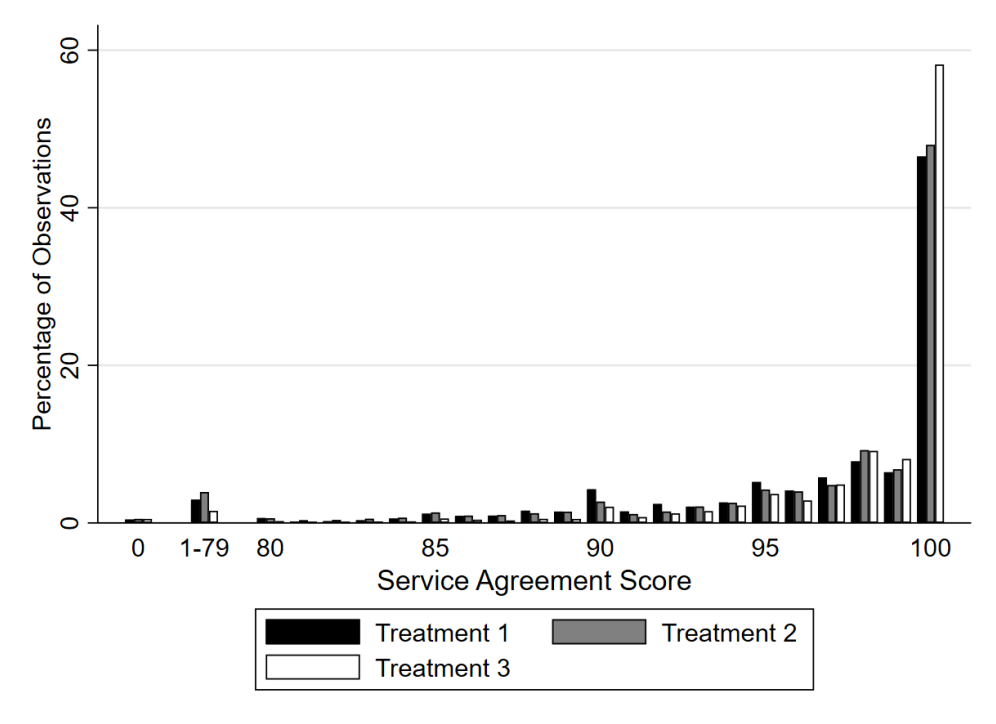
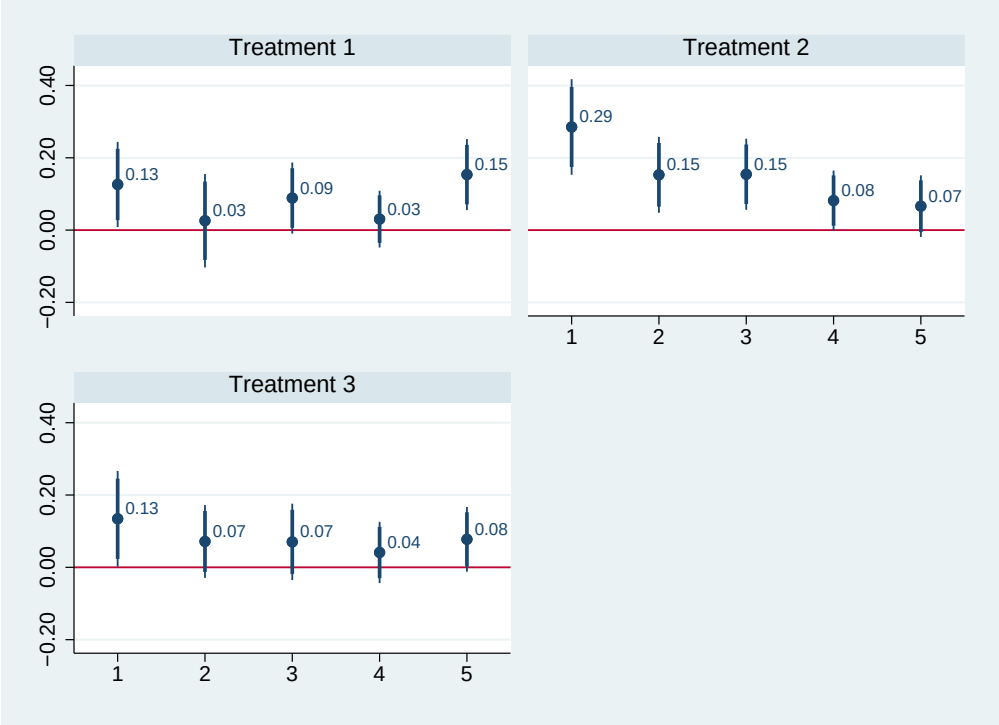
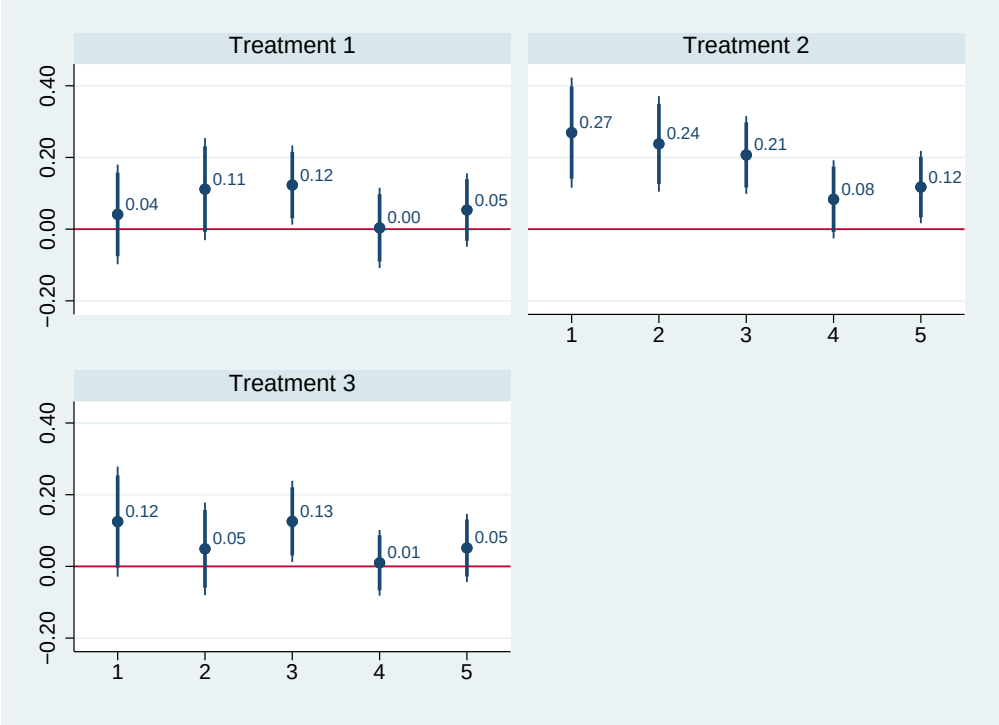


Figure A.2: The Distribution of the Service Agreement Scores by Treatment



(a) Indonesian



(b) Mathematics

Figure A.3: Impact on Standardized SLA Score by Baseline Grade Level

B Social Accountability Mechanism (SAM)

The SAM provides community members with an explicit role to monitor and evaluate teacher service performance and to ensure teacher accountability. The SAM is implemented by project facilitators, each is responsible for five to six villages. They initially conducted the meetings, but later built the capacity of Kader Desa (village cadres) to facilitate meetings and provided them with on-the-job mentoring. The SAM finances: (i) project facilitators; (ii) training and coaching of Kader Desa; (iii) capacity building for parents and community members through trainings and coaching by project facilitators and Kader Desa; (iv) project-facilitator-facilitated meetings between community members and teachers to develop Service Agreements (SA) and Community Scorecard; (v) meeting for community members in selecting User Committee (UC); (vi) monthly meetings to discuss UC's evaluation of teacher service quality and community members' evaluations on education service improvements; and (vii) transmitting the UC's evaluation and consolidated inputs to district education authorities.

The SA and the scorecard were developed through the following sequence of meetings led by project facilitators and Kader Desa: (i) with upper grade students and alumni representatives to discuss their wishes for ways to improve learning environments at school and at home; (ii) with parents and community representatives to present their children's wishes, and discuss what should be done by teachers, parents, and community representatives to improve learning environments at school and at home; (iii) with principal and teachers to cover the same topics; (iv) with principal, teachers, parents, and community members to agree upon a list of each of their promises to improve learning environments at school and at home; the SA listed the promises from each stakeholder; the scorecard is shortlisted from principal's and teachers' promises in the SA, with specific outputs and weights. Principal, teachers, parents, and community representatives were mostly free to choose the indicators to be included in the SA and the scorecard. However, facilitators were instructed to encourage indicators that relate to student learning and that are quantifiable. In addition, an indicator on teacher presence must be included in all SA and scorecards.

C 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.

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

To construct a measure of punishment norms, we employ a public goods game with punishment lab-in-the-field experiment (Fehr and Gächter, 2000). Budgetary constraints meant that we could only implement the experiment in 180 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: 38, 46, 42, and 44 participating schools were part of the Control, Treatment 1, Treatment 2, and Treatment 3 respectively.

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

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

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

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

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

E Efficiency Analysis

The one-time investment cost for the Project facilitators to conduct KIAT Guru approach was a total of USD 1,026,759, or at USD 5,058 per school and USD 40 per student (Table E.9). This cost covered trainings, salaries and transportation costs of 41 Project facilitators working in 203 schools, with an average of 132 students per school. Over three calendar years of implementation, the cost of training and

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

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

Table E.9: One-time Investment Cost to Introduce KIAT Guru

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

KIAT Guru continued after the endline survey. Below we report the cost of sustaining KIAT Guru. As these cost occurred after the endline survey, they have not been included in the cost benefit analysis in the main text. The average annual cost to sustain SAM was USD 2,182 per school or USD 17 per student (Table E.10), with additional USD 506 per school or USD 4 per student for Group 2 schools. This cost covers an annual refresher training, monthly meetings, and evaluation meeting. The annual cost of training per school at USD 709, or USD 6 per student. The annual cost to conduct monthly meetings was USD 834 per school, USD 7 per student. The average cost per village to conduct evaluation meetings at the end of every semester was USD 639 per school, USD 5 per student. In these evaluation meetings, User Committee and school providers reviewed the content of Service Agreement and Community Score Card indicators, and reappointed User Committee members. In 2017, 169 of 176 (96%) village governments provided the financial supports, and in 2018, all of them did, with an average of USD 674 (31% of total annual cost to sustain SAM). Given that these costs are only a very small fraction of Village Fund, the cost of maintaining KIAT Guru activities in the villages are completely sustainable. Group 2 schools had a total of USD 86,341 of additional cost to cover the purchase of one smart phone per school and salaries for two personnel to develop and maintain the application. On annual basis, this cost came down to USD 33,874, or at USD 506 per school, USD 4 per student.

Table E.10: Average Annual Cost per School for SAM

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