RISE Working Paper 22/090
February 2022

The Role of Coherence in Strengthening Community Accountability for Remote Schools in Indonesia

Yue-Yi Hwa, Sharon Kanthy Lumbanraja, Usha Adelina Riyanto, and Dewi Susanti

Abstract

Incoherence in accountability relationships can hamper the quality of education. Such incoherence can be a particular challenge in resource-constrained, remote villages where teachers tend to have higher educational capital and social status than the parents and communities that they serve. We analyze quantitative and qualitative data from a randomized controlled trial of a social accountability mechanism (SAM) for schools in remote Indonesian villages. The intervention had three treatment arms, all of which included the SAM, which engaged village-level stakeholders in a consensus-building process that led to joint service agreements for supporting the learning process. Prior analyses have found that all three treatment arms significantly improved student learning, but the treatment arm combining the SAM with performance pay based on camera-monitored teacher attendance led to much larger gains than the SAM-only treatment or the treatment arm combining the SAM with teacher performance pay based on a community-evaluated scorecard. Drawing on a range of quantitative data sources across all treatment schools (process monitoring, survey, and service agreement indicators) and qualitative data from nine case study schools (interviews and focus group discussions), we show firstly that the student learning gains across all three treatment arms were accompanied by increases in the coherence of the accountability relationships between village-level stakeholders, and in the degree to which these relationships were oriented toward the purpose of cultivating learning. We further show that the treatment combining SAM with camera-monitored teacher performance pay led to greater improvements in the coherence of accountability relationships than the other treatment arms, because the cameras improved both the technical capacity and the social legitimacy of community members to hold teachers accountable. This coherence-focused, relational explanation for the relative effectiveness of the treatment arms has more explanatory power than alternative explanations that focus narrowly on information quality or incentive structure. Our analysis reinforces arguments for ensuring that accountability structures are coherent with the local context, including local social structures and power dynamics.

The Role of Coherence in Strengthening Community Accountability for Remote Schools in Indonesia

Yue-Yi Hwa
Blavatnik School of Government, University of Oxford

Sharon Kanthy Lumbanraja
The World Bank

Usha Adelina Riyanto
The World Bank

Dewi Susanti
The World Bank

Acknowledgements:

We acknowledge financial support for the original data collections from the Government of Australia's Department of Foreign Affairs and Trade and USAID through Trust Funds managed by the World Bank. The Research on Improving Systems of Education (RISE) Study in Indonesia, managed by SMERU Research Institute, also co-financed the endline surveys. We are grateful to Clare Leaver, Menno Pradhan, Jan Priebe, Christopher Bjork, Arya Gaduh, and Jason Silberstein for valuable feedback on an earlier draft. During the analysis and writing of this report, Yue-Yi Hwa was a Research Fellow for the RISE Programme and Sharon Kanthy Lumbanraja, Usha Adelina Riyanto, and Dewi Susanti were consultants with the World Bank. The paper has also been published as a World Bank report.

This is one of a series of working papers from "RISE"—the large-scale education systems research programme supported by funding from the United Kingdom's Foreign, Commonwealth and Development Office (FCDO), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:
Hwa, Y., Lumbanraja, S.K., Riyanto, U.A., and Susanti, D. 2022. The Role of Coherence in Strengthening Community Accountability for Remote Schools in Indonesia. RISE Working Paper Series. 22/090. https://doi.org/10.35489/BSG-RISE-WP_2022/090

1 Introduction

Despite substantial increases in education budgets and student enrolment over the last two decades, learning outcomes in Indonesia have stagnated (Beatty et al., 2018; de Ree et al., 2018). Furthermore, teacher performance tends to be unsatisfactory, with high rates of teacher absenteeism in rural and remote areas. To address these challenges, KIAT Guru (*Kinerja dan Akuntabilitas Guru*, or Teacher Performance and Accountability) was implemented as a randomized controlled trial from 2016 to 2018 in remote districts of East Nusa Tenggara and West Kalimantan provinces in Indonesia.

KIAT Guru tested three treatments, all of which included a social accountability mechanism (SAM). Under the SAM, each community formulated a multi-stakeholder service agreement for improving local school quality, then conducted monthly evaluations of teachers and school leaders using a community scorecard. The treatment arms differed in whether and how they combined SAM with individual-level performance pay, which took the form of deductions from a government-funded allowance that would effectively double a teacher's base salary if awarded in full. In "SAM-only," no performance pay was implemented, so allowance-eligible teachers continued to receive the allowance in full. In "SAM+Score," SAM was combined with performance pay based on the community scorecard. In "SAM+Cam," SAM was combined with performance pay based on camera-monitored teacher attendance.

After one year of implementation, student learning outcomes improved significantly in all three KIAT Guru treatments compared to the control group (see Section 2.3 for detailed results). The SAM+Cam treatment group achieved the best all-around results (Gaduh et al., 2020; World Bank, 2020). A follow-up impact evaluation conducted one year after the endline showed that the gains persisted in SAM+Cam (Gaduh et al., 2021). Additionally, a prior qualitative study of KIAT Guru concluded that SAM+Cam was the most effective treatment because it empowered local community members to hold school leaders and teachers accountable. The reputational scorecard mechanism and a complementary financial disincentive (unlike SAM-only) was seen as unbiased and technically straightforward (World Bank, 2020). In contrast, the subjective and more technically demanding scorecard-based performance pay mechanism of SAM+Score was seen as less appropriate given the capacities and educational status of local community members relative to the teachers they were evaluating (ibid).

In this new study, we analyze field data from the nine schools in the qualitative case studies, along with three rounds of survey and process monitoring data from 203 treatment schools. This allows us to further investigate some factors that may have driven the learning gains from the project, building on the preceding KIAT Guru studies. We frame the analysis around the concept of *coherence*, or the degree of alignment between the various components of a specific education system. We draw on the RISE education systems framework, which focuses on coherence between stakeholders and design elements in accountability relationships for education service delivery, conceptualized as principal-agent relationships (Pritchett, 2015; Spivack, 2021). Given the persistently low learning levels in Indonesia despite improvements in input-oriented aspects of the education system, we are especially interested in whether these relationships are oriented toward learning as compared to other goals. Accordingly, we consider three aspects of coherence: the degree to which accountability relationships are *coherent for learning*; the degree to which there is *coherence between stakeholders* in village-level accountability relationships; and the degree to which there is *coherence between design elements* in the voice & choice accountability relationship between the village community (principals) and teachers/school leaders (agents).

We pose a two-part research question: *To what extent does the concept of coherence explain (a) the gains in student learning across all three KIAT Guru treatment groups and (b) the differences in effectiveness between the treatment groups?* Part (b) of the research question warrants particular interest because the greater effectiveness of SAM+Cam cannot be entirely explained by mainstream literature on incentives in job performance. The fact that the reductive, attendance-focused performance pay indicator of SAM+Cam was more effective than the multi-criteria indicator of SAM+Score runs counter to the argument that single-indicator performance pay schemes are ill-suited for multidimensional jobs such as teaching—particularly when the incentivized indicator is an input (e.g., teacher behavior) rather than a desired outcome (e.g., student growth). This argument has been advanced in both the economics literature, as in Holmstrom and

Milgrom's (1991) model of multitask principal-agent relationships, and educational research, as in Murnane and Cohen's (1986) analysis of teacher merit pay schemes in the U.S. Andrabi and Brown (forthcoming) offer a recent empirical example of this dynamic, where a performance pay scheme in private schools in Pakistan yielded similar test score gains whether teacher pay was contingent solely on student test scores or on broader appraisals by the school leader. However, the narrow test-score-based treatment weakened the non-incentivized domain of student socioemotional development.

Yet incentives based on single input indicators have, in some contexts, led to improvements in more complex outcomes. A notable example is a camera-monitored teacher performance pay intervention in India evaluated in Duflo, Hanna, and Ryan (2012), which led to significant gains in both teacher attendance and student learning. However, as we show below, it is unlikely that the teacher attendance gains in SAM+Cam were substantial enough to be the sole driver of the gains in student learning.[1] Hence one goal of this paper is to explain the puzzle of why the narrow incentive structures of SAM+Cam proved to be more effective than the broad-based incentives in SAM+Score.

We hypothesize that (a) as KIAT Guru progressed, all three treatments improved the degree to which stakeholder relationships were coherent for student learning, which led to learning gains; and (b) there were greater increases in the overall coherence of accountability relationships under the most effective treatment, SAM+Cam. To preview the analysis, we find evidence to support both hypotheses. Coherence for learning improved across all three treatment groups. The greater increase in coherence in SAM+Cam was driven by the tamper-proof cameras for monitoring teacher attendance, which empowered local community members not only with additional technical capacity but also additional authority to evaluate teachers. Thus, the cameras helped to redress a power imbalance between teachers and school leaders, on one hand, and children, families, and community members, on the other. As we show in our consideration of alternative hypotheses, this emphasis on overall coherence and power dynamics in accountability relationships has, in this context, more explanatory power than hypotheses that focus narrowly on the quality of information or on financial incentives.

Using coherence as the main analytical construct in this paper is appropriate given the design of the KIAT Guru project itself. KIAT Guru aimed to strengthen accountability relationships between agents, i.e., teachers and school leaders as service providers, and both sets of principals whom they serve, i.e., village communities as service recipients and the government as an employer. KIAT Guru expanded on a prior study by Pradhan et al. (2014) in rural public primary schools in Indonesia, which found that strengthening the relationship between school stakeholders and other village actors—through facilitated joint planning meetings between school committees and village councils—had a significant positive effect on student test scores. In contrast, treatment arms in the same study that focused on the school committees alone, by providing either block grants or training, did not have a significant effect on student learning. Accordingly, KIAT Guru expanded the joint planning meetings in Pradhan et al. (2014) into a social accountability mechanism that involved service agreements between a wider set of stakeholders (parents, teachers, school leaders, and village leaders), with the additional incorporation of performance pay.

Thus, together with the prior analyses of KIAT Guru, this paper contributes to the growing body of research on interventions that aim to improve education service delivery in developing countries by empowering parents and local communities. These include interventions that target parents' awareness of how to improve educational quality (e.g., Barrera-Osorio et al., 2021, on providing training sessions to parent associations in Mexico),[2] interventions that improve parents' access to information on student and/or school performance (e.g., Andrabi, Das, & Khwaja, 2017, on school report cards in Pakistan), and

[1] However, it is important to note that teacher attendance was measured during one unannounced visit per month for all 30 months of the intervention in Duflo, Hanna, and Ryan (2012); whereas KIAT Guru only conducted unannounced teacher absence surveys at two points (baseline and endline), yielding a noisier measure of teacher attendance.

[2] Unlike the other interventions mentioned in this paragraph, the parent training sessions in Barrera-Osorio et al. (2021) did not improve student learning. However, they had the positive effect of reducing disciplinary actions in schools.

interventions that improve parent-teacher communication (e.g., Islam, 2019, on parent-teacher meetings to discuss student progress in Bangladesh).

We also contribute to research on interventions that test different financial incentive structures for teacher performance in low-resource settings. The design has similarities to Duflo, Hanna, and Ryan's (2012) evaluation of camera-monitored teacher performance pay in India, Cilliers et al.'s (2017) evaluation of teacher performance pay based on school leaders' reports of their attendance in Uganda,[3] and Andrabi and Brown's (forthcoming) evaluation of teacher performance pay based on student test scores and on appraisal by school leaders. In examining both the service provider–community relationship and the service provider–government relationship, we find that the structure of monitoring and incentives affected teachers' responses to KIAT Guru treatments. However, the differential effects between treatment groups were not due to financial incentives in isolation, but rather mediated by wider power dynamics in the village-level educational ecosystem.

We explore these power dynamics, and coherence in KIAT Guru more broadly, in the rest of this paper. In Section 2, we give an overview of the KIAT Guru project. In Sections 3 and 4, we lay out our conceptual framework, data sources, and analytical approach. In Section 5, we report our findings, focusing on three aspects of coherence. Section 6 is a brief discussion, and Section 7 concludes.

2 Overview of the KIAT Guru project

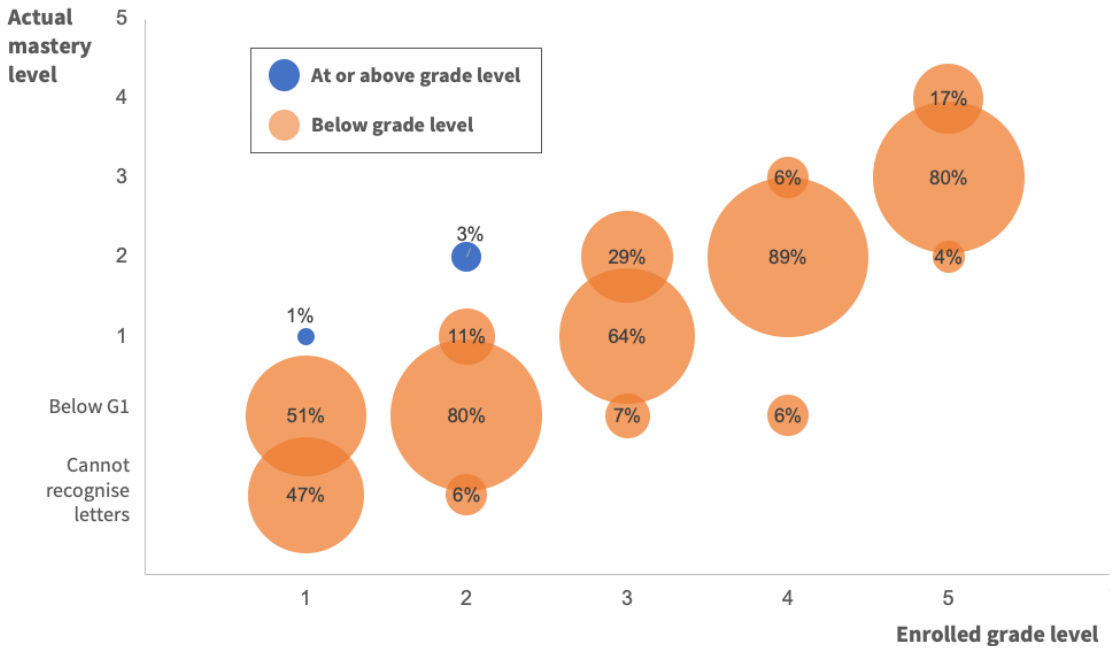
2.1 Context of the intervention

The KIAT Guru project was motivated by high teacher absenteeism (ACDP, 2016) and low student learning outcomes in remote areas of Indonesia (Stern & Nordstrum, 2014; BPS, 2017). The National Team for the Acceleration of Poverty Reduction (*Tim Nasional Percepatan Penanggulangan Kemiskinan*) under the Indonesian Vice President's Office collaborated with the Ministry of Education and Culture to initiate the project in 2016 with support from the World Bank. Five districts from the government's list of disadvantaged regions participated in the project: Ketapang, Landak, and Sintang in the province of West Kalimantan; and Manggarai Barat and Manggarai Timur in the province of East Nusa Tenggara. The government formally established a national committee along with local committees in each of the five districts to coordinate implementation throughout the project's timeline.

The project used two separate instruments to measure student learning in Indonesian (language) and mathematics. The first instrument was an evaluation test to measure learning outcomes at baseline, endline, and the one-year follow-up (Lumbanraja, Prameswari & Susanti, 2021). In the evaluation test conducted at baseline (between late 2016 and early 2017), 25 percent of classrooms had no teacher, with students being two grade levels below the curriculum standard on average, as shown in Figure 1. Yet, parents surveyed at baseline expressed satisfaction with the quality of education and learning outcomes, suggesting either low expectations or inadequate information (World Bank, 2019). To address the lack of easily available data on student learning outcomes, the KIAT Guru research team developed a second student learning instrument: a rapid diagnostic test to give community members a gauge of student literacy and numeracy (Lumbanraja & Prameswari, 2021). This diagnostic test was similar to learning assessments used in community advocacy in comparable contexts (ASER, 2014; PAL Network, 2018), and was administered regularly throughout the intervention.

[3] However, local monitoring of teacher performance in KIAT Guru's SAM+Score treatment involved not only official reports but also monitoring by community members.

Figure 1. Distribution of students' Indonesian literacy levels against their enrolled grade levels in 270 remote primary schools (KIAT Guru baseline study)



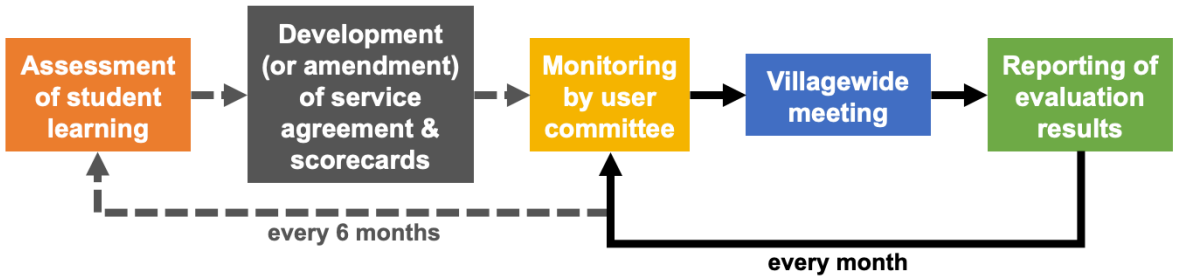
Source: Authors, using data from Table 52 in World Bank (2019).

2.2 The social accountability mechanism

Under KIAT Guru, all three treatment groups implemented a social accountability mechanism (SAM), which operated in a cycle as summarized in Figure 2KIAT Guru treatment groups and outcomes. The initial cycle of SAM implementation began with KIAT Guru project facilitators presenting village-level student learning data from the baseline evaluation test, benchmarked against national curriculum targets, in a community meeting to trigger a series of discussions on how learning environments at schools and at home can better support learning. Modeled after community participation in monitoring frontline service providers (Bjorkman & Svensson, 2009), the meetings were initially conducted with separate stakeholder groups—upper-grade level students and alumni, parents and community representatives, and school leaders and teachers—prior to having intergroup meetings to develop a schoolwide service agreement and teacher-specific scorecard. At the end of this series of meetings, community representatives elected a gender-balanced user committee (UC), consisting of at least nine members (six parents, representing each grade level, and three community leaders).

The UCs monitored teachers and school leaders regularly and presented their findings at a monthly village-wide meeting, where each teacher and school leader was assigned scorecard ratings, which were reported to village and district governments. After each six-month period of implementation, the UCs and a locally appointed village cadre conducted another round of diagnostic learning assessments, which were followed by a round of amendments to the service agreements and scorecards, if the community deemed such amendments necessary.

Figure 2. The social accountability mechanism in KIAT Guru



The primary source of frontline support for KIAT Guru came from project facilitators, each of whom initiated implementation in several schools. They facilitated the initial stakeholder discussions that led to service agreements and scorecards, and raised awareness of relevant issue among community representatives in "socialization" sessions. These sessions offered school-specific information on student learning outcomes along with general information on children's rights to quality education and local communities' rights to participate in and monitor education service delivery. They also trained UCs in collecting data to monitor progress on service agreements and scorecard indicators (through reviewing administrative data, conducting unannounced spot checks in school, and interviewing students and parents). Additionally, the facilitators identified, mentored, and coached a village cadre to take over their role in facilitating meetings once project implementation was eventually handed over to the communities (Marliyanti, Adelina & Susanti, 2022).

Besides this on-the-ground support from project facilitators, KIAT Guru was also supported through official channels. Each village government issued decrees to formalize the appointment of the UC and the village cadre, thus legitimizing their roles to participate in, monitor, and evaluate teacher scorecards. The district and national governments also issued special regulations to support the project. Alongside the commissioning decrees was an expectation that the village government would also allocate some funds to support the UC, the village cadre, and the school, such as money to cover the costs of office supplies or light refreshments for UC members while on duty.

2.3 KIAT Guru treatment groups and outcomes

Alongside the SAM that was implemented in all three treatment groups, KIAT Guru tested three different models of teacher accountability and incentives, as summarized in Table 1. The first group implemented a SAM-only treatment, which relied on social rewards and sanctions for teachers' scorecard performance. The other treatment groups added a performance-based pay element, where unsatisfactory teacher performance could result in cuts to the government-funded teacher special allowance, enjoyed by roughly 35 percent of teachers across the treatment groups.[4] The second group, SAM+Score, tied the teacher special allowance to the monthly teacher scorecard as evaluated by the UC, such that a teacher who scored 95 percent in a given month would receive 95 percent their allowance for that month. The third group, SAM+Cam, tied the allowance to teachers' presence in school. SAM+Cam schools were provided with a smartphone with an application called KIAT Kamera, and teachers were required to record their daily presence in school using KIAT Kamera. Every month, the UCs would compare KIAT Kamera data with administrative attendance records to verify whether teachers' absences had been formally excused by the school leader.[5] The UC's verification reports were then used to calculate

[4] This special allowance (*Tunjangan Khusus Guru*) is allocated by the central government for civil servant teachers appointed to schools in special areas (including remote areas). It is equal to their base salary. In addition, teachers who are registered as civil servants also receive an allowance of IDR 1.5 million (USD 105). Non-civil-servant and non-registered teachers, most of whom are school-contracted teachers, do not receive this allowance, even if they serve in remote schools. For more on the different employment statuses (e.g., registered civil servants vs. contract teachers) in Indonesia, see Huang et al. (2020).

[5] "School leader" here refers to the school principal. School leaders were also subject to the interventions; they also underwent the same evaluations described for the teachers. Their evaluations were sent to and verified by district-level school supervisors.

deductions to a 100-point presence score for each teacher, with daily deductions of 1.5 points for partial attendance (i.e., arriving late and/or leaving early), 2 points for an excused absence, and 5 points for an unexcused absence. If a teacher's monthly presence score fell below 85, they would not receive their allowance. Otherwise, the amount matched their presence score, such that a teacher who received a presence score of 90 in a given month would receive 90 percent of their allowance for that month.

Table 1. Summary of KIAT Guru treatment groups

	Control	SAM-only	SAM+Score	SAM+Cam
Social accountability mechanism with user committee and scorecards		✓	✓	✓
Performance pay based on:				
Teacher presence in school			✓	✓
Other aspects of teacher practice			✓	
Tamper-proof cameras for monitoring presence in school				✓
Number of schools	67	68	68	67

Source: adapted from Table 1 in Gaduh et al. (2020).

Thus, across the treatment groups, the UCs had different capacities for sanctioning underperforming teachers. In SAM-only, they could penalize teachers reputationally by giving them a low score. In SAM+Score, the UC arguably had the most discretionary power, as the scores they chose to award to each teacher would result in both reputational and financial sanctions. In SAM+Cam, the UC's discretionary judgements affected the scores that were associated with social sanctions, but they had less room for discretion with the camera-validated teacher presence scores that determined the financial sanctions.[6]

The impact evaluation of KIAT Guru found that all three treatments improved learning outcomes, with average effect sizes across Indonesian (language) and mathematics of 0.08SD for SAM-only, 0.11SD for SAM+Score, and 0.20SD for SAM+Cam (see Table 3 in Gaduh et al., 2021; see also Gaduh et al., 2020). This analysis rejected equality between SAM+Cam and the other two treatment arms ($p<0.02$). An additional round of data collection for SAM-only and SAM+Cam one year after the endline found that learning outcome gains persisted in SAM+Cam with an effect size of 0.13SD on average across the two subjects, whereas learning outcomes in SAM-only were no longer significantly different from the control group (ibid).[7]

In addition to the quantitative impact evaluation, a second study drew on three rounds of detailed qualitative fieldwork in nine schools, with three schools from each of the three treatment groups. The qualitative study concluded that SAM+Cam was the most effective treatment because it combined enforcement (in the form of a performance-based pay cut) with an evaluation metric that was viewed as objective rather than contingent on user committee members' interpretations (World Bank, 2020).

In this reanalysis, we revisit both the qualitative data and the three rounds of quantitative survey data, complemented with process monitoring data to build a deeper understanding of the processes underlying the differential success of these treatment groups.

[6] For more on the theoretical underpinnings of the three KIAT Guru treatments, refer to the impact evaluation papers (Gaduh et al., 2020, 2021) and the qualitative fieldwork report (World Bank, 2020).
[7] Given budget constraints as well as the government's stated interest in scaling up SAM+Cam based on the results of the initial impact evaluation, follow-up data were not collected for SAM+Score (Gaduh et al., 2021).

3 Conceptual framework

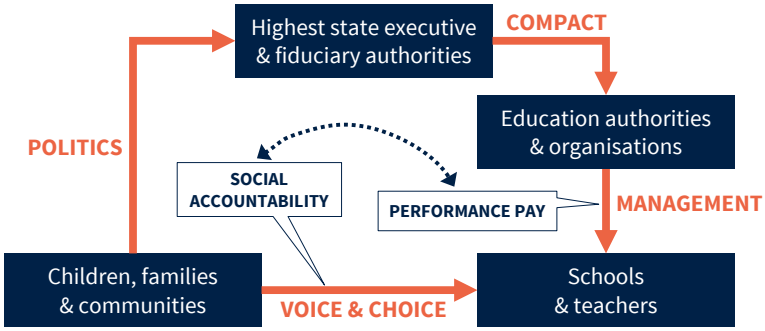
The main analytical construct in this paper is *coherence* in relationships within education systems. We define coherence as "the degree of alignment between the various components of a specific education system". Given the many interacting components within complex education systems, we focus our analysis of coherence on those aspects emphasized within the RISE education systems framework (Pritchett, 2015; Spivack, 2021; see also World Bank, 2003). This framework identifies four accountability *relationships* within education systems, and five main *design elements* within each relationship.

Each accountability relationship focuses on a set of principals—or actors who want a task to be completed—and a set of agents—or actors whom the principals engage to complete the task. The four relationships are:

- the *politics* relationship, between citizens (principals) and the highest authorities of the state such as the president/prime minister (agents);
- the *compact* relationship, between the highest authorities of the state (principals) and education authorities and organizations such as the education ministry (agents);
- the *management* relationship, between education authorities and organizations (principals) and frontline providers such as school leaders and teachers (agents); and
- the *voice & choice* relationship, between service recipients such as children, families, and communities (principals) and frontline providers (agents).

Figure 3 shows how these four relationships are interconnected in an education system, and how the two main KIAT Guru components map onto these relationships. The social accountability mechanism operates primarily within the voice & choice relationship, and the performance pay component operates primarily within the management relationship (because the financial incentive involves deductions from a government-allocated special allowance). Nonetheless, the two components are connected. Under the performance pay component, the user committee of local community members either generated (in SAM+Score) or verified (in SAM+Cam) the performance measure that determined the financial incentive. The fact that these community-based appraisals drew on attendance records administered by school leaders also helped to bring these village-level relationships into coherence with each other. Above the community-school level, the clearly specified service agreement indicators and scorecards also helped to align expectations—and improvement efforts—from schools to village and district governments.

Figure 3. How the components of KIAT Guru map onto the RISE education systems framework



Source: adapted from Pritchett (2015) and Spivack (2021).

In each of the four relationships, there are five design elements that can fundamentally affect the nature of the relationship and what the relationship accomplishes. The five design elements are:

- *delegation*, which is the goal that principal wants the agent to fulfill, such as when children, families, and communities (principals in the voice & choice relationship) want teachers (agents in the voice & choice relationship) to educate the children under their care;

- *finance*, which represents the resources that the principal provides to the agent to fulfil the delegated goal, such as the fees that families pay to local schools or the time they contribute to supporting learning-related initiatives;
- *information*, which refers to what the principal knows about the agent's performance, such as when parents receive report cards from the school;
- *motivation*, which is how the agents' wellbeing is affected by their performance on the delegated goal, such as when a teacher feels personally fulfilled or gains social recognition from the local community from helping students to improve their learning; and
- *support*, which refers to resources and processes that strengthen the accountability relationship, especially when they are provided to the less empowered actor in the relationship.

Using the framework of these design elements and accountability relationships, we examine three different aspects of coherence in accountability relationships. First, we look at the goal or purpose for which accountability relationships are coherent, looking specifically at *coherence for learning* in the village-level voice & choice relationship. This part of the analysis focuses on the design element of delegation, i.e., the goals that principals want agents to fulfil, with secondary attention to the design element of finance, i.e., the extent to which principals are willing to make costly contributions to the achievement of these goals. The overarching purpose or orientation of an accountability relationship is especially important in domains like education where there are multiple priorities that can inadvertently compete for resources and for stakeholders' attention. In many education systems that have high levels of enrolment but low levels of student learning, accountability relationships are oriented toward the expansion of access to schooling, but they are not coherent for cultivating student learning (Pritchett, 2015).

Second, we look at *coherence between stakeholders in accountability relationships*. In this part of the analysis, we depart from the RISE framework in that we do not focus on design elements within accountability relationships. In our study context, the village-level voice & choice relationship includes a heterogeneous mix of principals: students, parents, school committees, village heads, user committees, and village cadres, all of whom participated in field interviews and/or focus group discussions. The nature of the qualitative case study data is such that we can either aggregate across multiple stakeholder groups to summarize their perceptions of specific design elements, which we do subsequently, or explore more generalized perceptions of each disaggregated stakeholder group, which we do in this part of the analysis. Specifically, we look at how different stakeholder groups generally regard other groups (e.g., how parents regarded teachers), exploring whether KIAT Guru led to more positive views of each other's contributions to student learning. We look at the views of principals and agents within both the voice & choice and management relationships at the village level. This falls within the broader definition of coherence as alignment between the components of a specific education system.

Third, returning more closely to the RISE framework, we look at *coherence between design elements within a single accountability relationship* (Pritchett, 2015; Spivack, 2021). For example, if the education ministry delegates to teachers the goal of improving student learning levels, but does not provide teachers with the training and instructional materials needed for effective classroom learning, this would constitute an incoherence between delegation and information in the management relationship. In our analysis, we focus on coherence (or lack thereof) between design elements in the voice & choice relationship, in the context of the village-level principals and agents who participated in KIAT Guru.

As noted in the introduction, we ask a two-part research question and test two hypotheses within this conceptual framework. The question is: *To what extent does the concept of coherence explain (a) the gains in student learning across all three KIAT Guru treatment groups and (b) the differences in effectiveness between the treatment groups?* We hypothesize that: (a) as KIAT Guru progressed, all three treatments improved the degree to which stakeholder relationships were coherent for student learning, which led to learning gains; and (b) there were greater increases in the overall coherence of accountability relationships under the most effective treatment, SAM+Cam. In the analysis that follows, we weigh the evidence for both hypotheses within each of the three aspects of coherence outlined above.

3.1 A note on "support" in the context of power imbalances

Spivack (2021) defines support in the RISE education systems framework as "preparation and assistance that the principal provides to the agent to complete the task" (p. 9). Our definition of support overlaps with Spivack's in some instances. Spivak gives the example of the education ministry providing training to teachers in the management relationship, which falls within her definition of support as assistance from the principal (in this case, the ministry) to the agent (in this case, teachers). This also falls within our definition of support as resources and processes that strengthen the accountability relationship especially by empowering the weaker actor: without appropriate training, teachers are unlikely to fulfil their complex responsibilities, which would weaken the accountability relationship because of the wide gap between stated delegation and feasible performance.

However, we depart from Spivack's definition of support because the power dynamic in one of the most important relationships in the KIAT Guru project, i.e., the voice & choice relationship, was such that the principals (children, families, and communities in remote villages) were initially much less empowered than the agents (school leaders and teachers). Consequently, the relationship could only be coherent with adequate support to rebalance that dynamic and strengthen the accountability relationship. Such support for UCs included capacity building sessions, appointment via official decrees, and access to tamper-proof data, which strengthened both their technical capacity and perceived legitimacy, as we discuss below.

We are far from being the first to point out that hierarchical differences between teachers and other stakeholders can affect accountability structures. For example, Narwana (2015) observed that teachers in Haryana, India, were resistant to a community accountability initiative partly because they regarded themselves as higher-status actors who should not be subject to the judgement of lower-status, less-educated villagers. Broekman (2015) found that teachers in Indonesia questioned even their fellow teachers' authority to conduct peer appraisals, on the basis that such authority should be held exclusively by those higher in the hierarchy, such as school leaders or supervisors.

More generally, Fox (2015) argues in a review of social accountability interventions in developing countries that "voice" from citizens needs to be complemented with "teeth" from the state (i.e., positive and negative incentives) because anti-accountability forces are likely to hold more power than pro-accountability forces, whether within the state or in society—or, in the language of the RISE framework, the voice & choice relationship needs to be coherent with the management relationship. Similarly, in a review of 48 studies of interventions that aimed to use information to improve rural service delivery, Kosec and Wantchekon (2020) conclude that informational interventions will only yield gains if three conditions hold: the information must be relevant, the recipients of the information must have incentives to act on it, and they must also have the power to act on it. They further note that the rural poor—a category encompassing the communities in KIAT Guru—often lack the power to respond to the information they receive (see also Banerjee & Duflo, 2008, on the relative powerlessness of village education committees in Uttar Pradesh, India). Given this need to redress the disempowerment of rural villagers in accountability for service delivery, we redefine the design element of support to incorporate these power dynamics.

4 Data and methods

4.1 Data sources

Process monitoring data. Throughout the implementation of KIAT Guru, process monitoring data from the 203 treatment schools were collected by project facilitators and project implementation teams. Process monitoring data includes some indicators that were routinely collected (every month) and some that were occasional. Routinely collected data include monthly teacher scorecard ratings and data from the teacher attendance verification process in SAM+Cam schools. Occasionally collected data include information on project implementation, such as its timeline, descriptions of community participation in the project, and assessments of UCs' and village cadres' fulfilment of their responsibilities. Process monitoring data collection extended from November 2016 until June 2019.

Quantitative survey data. For impact evaluation purposes, an independent survey team collected quantitative survey data during three different periods (Gaduh et al., 2021). A baseline survey was conducted between November 2016 and February 2017. After roughly one year of implementation (and soon after SAM facilitation was handed over from project facilitators to village cadres), an endline survey was conducted between February and mid-April 2018. A follow-up survey was undertaken from March to May 2019 to assess the impact of the project intervention one year after it concluded. All 270 treatment and control schools were included in the baseline and the endline surveys. However, due to budget restrictions, the relatively disappointing endline findings on SAM+Score, and the government's stated preferences, only SAM+Cam, SAM, and control schools were included in the follow-up survey.

The quantitative survey encompassed evaluation tests of student learning, teacher absence surveys, and interviews with stakeholders participating in the project. The student learning tests assessed basic literacy and numeracy competencies based on the 2006 national curriculum. Students in grade 1–5 in participating schools took part in the assessment during the baseline survey, and subsequently in the endline survey when most of them were in grades 2–6. The teacher absence surveys aimed to provide a representative description of teacher attendance in participating schools through an unannounced visit during normal school hours on the first day of each quantitative survey round.

The quantitative interviews with local stakeholders captured background information on the perceptions, characteristics, and dynamics of stakeholders at the village level. School leaders, teachers, parents, school committees, and village heads were interviewed in all three rounds of the survey. UC members were only interviewed in the endline and follow-up surveys as the committees had not been established at baseline.[8] In this paper, the primary areas of interest from the surveys are perceptions of teacher performance, school leadership, the extent of local government support, parental participation, UC effectiveness, and indicators that help to triangulate qualitative case study data on the relationships between stakeholders.

Service indicators. At the start of KIAT Guru, teachers, village leaders, and parents in all treatment schools formulated joint service agreements to improve education quality, as discussed above in Section 2.2. Teacher service agreements were operationalized into teacher scorecards containing five to seven locally defined service indicators, which always included an indicator of teacher presence. These scorecards were used by UCs to monitor and evaluate each teacher every month. The scores were then reported to the district government. After each six-month semester, local stakeholders met to evaluate these scorecards for relevance, during which service indicators were sometimes amended to reflect identified needs.

There was a total of 2,542 service indicators in the beginning of the project, which grew into 10,508 service indicators following the first amendment of the scorecard. This large increase was attributed to growing need for parents and teachers to tailor each scorecard to individual teachers' subject- or grade-specific responsibilities.

Qualitative case studies. To complement the process monitoring and impact evaluation, a series of qualitative case studies were conducted in nine of the 203 treated schools (World Bank, 2020). The nine schools were located in three of the five project districts: Ketapang and Landak (in West Kalimantan province), and Manggarai Barat (East Nusa Tenggara). In each district, three villages with similar characteristics—such as student learning outcomes, geography, and community characteristics—were chosen as case study sites, each representing one of the three treatment groups. The qualitative data collection was designed to identify changes in stakeholders' views about education quality and the social accountability process. The three SAM-only schools were Sangka, Engkangin, and Sungai Laur; the SAM+Score schools were

[8] School committees are pre-existing entities appointed by the school leader and primarily involved as counterparts in school management and the utilization of school operational funds. UCs are elected bodies, newly established under KIAT Guru to represent the wider village community in administering the social accountability mechanism. UCs are formally appointed by the head of the village in which the school is located.

Konang, Sungai Keli, and Simpang Dua; and the SAM+Cam schools were Kondok, Sampuraneh, and Usaba Sepotong.[9]

Researchers visited each school three times: in October–November 2016 before KIAT Guru started, in August–September 2017 after monthly teacher evaluations had been conducted several times in the schools, and in February-March 2018, shortly after the KIAT Guru project facilitators had handed over implementation responsibilities to village stakeholders. The qualitative research questionnaires were developed by three principal investigators,[10] who recruited and trained field researchers who had master's degree in social sciences. The questionnaires were piloted in collaboration with the field researchers and then revised prior to each field visit. Each of the nine schools wase studied by two researchers, who visited the field together and shared data collection responsibilities. They interviewed students, teachers, parents, school leaders, UC members, government officials, and members of the community; attended UC meetings; facilitated focus group discussions with teachers, students, parents, and UC members; examined school records; observed lessons; and collected evidence of student learning outcomes from school report cards. Interviews and focus group discussions were transcribed by the field researchers who conducted them. These transcripts are the main sources of qualitative data in this paper.

4.2 *Analytical approach*

As noted above, we analyze these four data sources to explore the extent to which different aspects of coherence in accountability relationships can explain the overall improvements in student learning across treatment groups, as well as the relative effectiveness of the SAM+Cam treatment. We draw on a mix of sources throughout the analysis, depending on data availability for the constructs and issues in question. For example, our analysis of teacher attendance and incentives mainly uses quantitative data from process monitoring and surveys (see Section 5.3.4), while our discussion of power dynamics in the voice & choice relationship mainly uses qualitative case study data (see Section 5.3.1).

For the quantitative data sources—process monitoring data, quantitative surveys, and service agreement indicators—we use straightforward descriptive statistics, along with some significance tests. Where relevant, we also refer to models estimated using these data in the impact evaluation papers (Gaduh et al., 2020, 2021). We analyze a wide range of *process monitoring* and *quantitative survey* variables that we identified, ex ante, as being related to student learning and coherence in accountability relationships between village-level stakeholders. Based on these findings, we iteratively refined our arguments, weighing evidence both for our hypotheses and for alternative hypotheses. The analysis of process monitoring and quantitative survey data is summarized in Appendix A, and the most salient indicators are discussed in Section 5.1 on coherence for learning and Section 5.3 coherence between design elements in the voice & choice relationship.

The raw *service agreement indicators* from village-specific scorecards were first grouped into larger categories, with 20 categories for teacher service agreement indicators (e.g., "Teacher applies fun and motivating learning techniques in classroom") and 14 categories for school leader service agreement indicators (e.g., "Principal sets good example through their attendance and behavior whilst in school"). These categories were then classified according to whether or not they were directly related to student learning. Changes in distribution of indicators across categories before and after the first amendment process are outlined in Section 5.1 on coherence for learning, with further descriptive statistics in Appendix B. We focus on the first round of semesterly amendments to the service agreements because this gives a snapshot of the extent to which village stakeholders' pre-treatment educational priorities, as articulated in the initial service agreements, were altered through participation in KIAT Guru.

---

[9] Although individual participants in the field data collection are identified by their titles or designations (e.g., village head, UC member) rather than by name, we identify schools by name for consistency with the previous qualitative study (World Bank, 2020).

[10] Christopher Bjork, Raihani, and Dewi Susanti. Bjork is the Dexter M. Ferry, Jr. Professor of Education and Chair & Coordinator of Teacher Education at Vassar College in the United States. Raihani is a Professor of Islamic Education Studies at the State Islamic University of Sultan Syarif Kasim in Indonesia.

For the *qualitative case study data*, we read the transcripts of interviews and focus group discussions with village-level stakeholders in full, coding them according to a scheme based on three thematic areas of interest: (a) coherence, because of the conceptual framework of the paper; (b) the quality of information, because this was identified as a key variable in the initial analysis of the qualitative data (World Bank, 2020); and (c) power dynamics, because both the wider literature and prior experience and analysis of KIAT Guru suggested that this would be an important explanatory factor. However, these codes were neither applied comprehensively nor analyzed in a structured manner. Rather, we use them to gain traction and highlight points of interest in a large corpus of 180 interview transcripts and 122 focus group discussion transcripts of village-level stakeholders. On occasion, we also draw on interview transcripts with stakeholders from other administrative levels (e.g., project facilitators who coordinated KIAT Guru implementation across several schools) as well as from field notes and summary reports from the field researchers. Quotes included in the text have been translated from Indonesian to English and lightly edited for readability.

In addition to this broad analysis of key themes in the interview data, we conducted a systematic analysis of stakeholders' views of each other, which is presented in Section 5.2 on coherence between stakeholders in accountability relationships, with further summary tables in Appendix C. For this analysis, we extracted quotes from interviews and focus groups in which a stakeholder (or set of stakeholders) expressed opinions about another stakeholder (or set of stakeholders). We then summarized each set of views with one classification for each stakeholder pair in each direction, for each data collection round in each school (e.g., there is one classification for Sangka parents' views of the school leader at midline, and one classification for the Sangka school leader's view of parents at midline). Each classification was assigned by one of us and then checked by a co-author. Any divergences in classification were discussed and resolved by consensus. While some of these classifications represent an individual's views (e.g., the school leader), others span multiple individuals (e.g., the views of multiple parents summarized into a single classification) as well as multiple interviews and focus groups (e.g., the views of multiple teachers across a focus group and a few individual interviews from the same field visit summarized into a single classification). The classifications are: *negative*; *neutral or "don't know"*; *mixed* (i.e., having positive views of some actions but negative views of others; or some members of the stakeholder group having positive views and others having negative ones); *positive* (which can include some mildly negative views if the overall opinion is obviously positive); and *positive and oriented toward student learning.*

To illustrate the difference between the last two categories, this quote from the midline interview with the school leader at Engkangin represents a generically positive view of teachers:

> There's been a small change. Previously they were a bit tardy. Now they're more punctual. Maybe they're getting more and more aware of the roles and responsibilities in which they need to show determination, for the sake of their performance [on the scorecard rating].

In contrast, during the endline round of data collection, the same school leader's view of the teachers in his school was not only positive, but also distinctly oriented toward student learning:

> Over the last six months, there have been quite a lot of changes. First, there's been change in the fulfilment of their duties. Before, they were maybe less effective in delivering teaching materials, but now that there are indicators for foundational skills they're right on target.

As discussed above, this distinction between generic positive views and positive views oriented toward student learning is crucial because actors within education systems can have a range of priorities for schools and teachers on the frontline, some of which support student learning, and some of which compete with it for attention and other resources.

To strengthen the validity of our argument, in examining the third aspect of coherence (i.e., coherence between design elements within the voice & choice relationship) we weigh the evidence for our explanation of the differential effectiveness of the KIAT Guru treatment groups against the evidence for two alternative hypotheses. This aligns with both with statistical approaches to hypothesis testing and with realist arguments about validity in policy research (Pawson & Tilley, 1997) as well as educational research (Porpora, 2015).

5 Results

5.1 Coherence for learning

To examine the extent to which the KIAT Guru treatments increased coherence for learning in the village-level voice & choice relationship, we examine, firstly, service agreement indicator data for changes over time in the distribution of scorecard weightings (out of 100 percentage points) for indicators that are directly related to student learning compared to indicators that are less related to learning; and, secondly, quantitative survey data for changes in the extent to which village-level stakeholders took actions that directly support the learning process.

5.1.1 Prioritizing student learning in service agreements

Beginning with the service agreement indicators, we focus on the first round of amendments to the service agreements, which took place after the first semester of KIAT Guru implementation as well as an additional round of rapid diagnostic tests to inform the amendment process. We examine whether these amendments resulted in heavier weighting of service agreement indicator categories that are directly related to student learning. We interpret this as a proxy for the degree to which village stakeholders prioritized learning (versus other goals) in monitoring the performance of teachers and school leaders. For teachers, eight out of the 20 categories of service indicators are directly related to student learning (e.g., "Teacher strives to ensure students' learning comprehension, including in providing feedback"), while the other twelve are indirectly related at best (e.g., "Teacher inculcates patriotism and values of obedience and orderliness in students"). For school leaders, three out of the 14 service indicator categories are directly related to learning.[11] An example of the service agreements for teachers in one of the qualitative case study schools, SDK Kondok, is shown in Table B1 in Appendix B .

Across all three treatment groups, there was a clear post-amendment shift toward indicators directly related to student learning.[12] Prior to the amendments, the eight indicator categories in the teacher service agreements that were directly related to student learning had an average total weighting of 36.6 percentage points, which increased to an average of 52.1 percentage points post-amendment. This increased emphasis on learning-related indicators was consistent across treatment arms, with pre-amendment weightings ranging from 35 to 40 percentage points, and post-amendment weightings ranging from 51 to 54 percentage points (see Table B2 in Appendix B for more details). For school leaders, there was a similar but smaller shift, with the three learning-related school leader service indicator categories having an average total weighting of 22.6 percentage points pre-amendment and 29.3 percentage points post-amendment (see Table B3 for more details). Again, changes were of comparable magnitude across treatment groups.

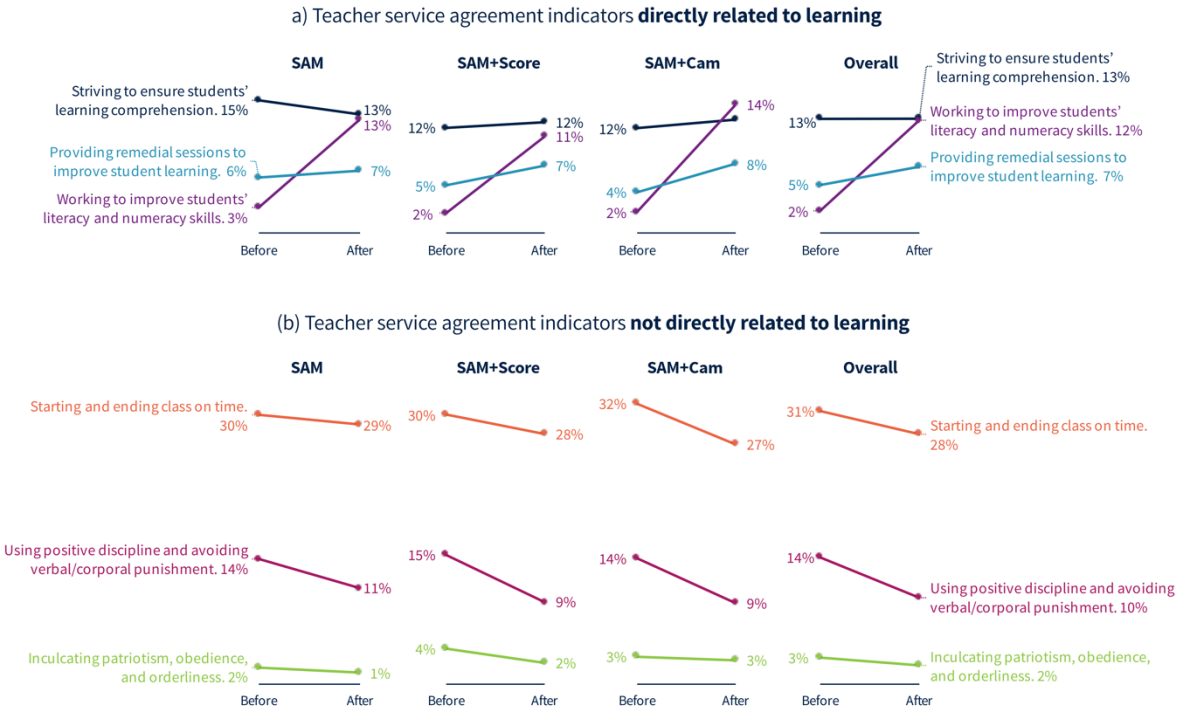
[11] We are not implying that the service indicator categories that are indirectly related to learning are unimportant or illegitimate. Some of these categories represent other goals that matter tremendously for holistic development but may inadvertently compete with learning goals, e.g., "Teacher instils religious, cultural and social norms in students." (See also Kurniasih et al., 2018, on the character education policy in Indonesia.) Other categories represent inputs that are necessary but far from sufficient for student learning, e.g., "Teacher starts and ends class on time." While such inputs are obviously important, many low-performing education systems fail to cultivate adequate levels student learning because they focus excessively on inputs rather than on how these inputs contribute to meaningful classroom interactions (Pritchett, 2013). In this analysis, we focus on increases in coherence for learning because the outcome of interest that we are trying to explain is gains in student learning.

[12] It is unclear the degree to which this increased weighting of learning-related indicators was due to changes in village stakeholders' priorities throughout the intervention, or to the amendment process itself. This probably varied across schools. For example, field notes from the qualitative case study of SDN Simpang Dua indicate that the project facilitator who led the amendment meeting at this school consistently asked meeting attendees whether each indicator had a direct impact on student learning outcomes; whereas field notes from the amendment meeting at SDI Konang do not indicate that the project facilitator played such an explicitly learning-oriented role.

This increased emphasis on learning-related indicator categories is also evident when looking at pre- and post-amendment changes in specific indicator categories. For example, out of the top three indicator categories for school leaders, pre-amendment, only one was directly related to student learning, i.e., "School leader develops a teaching schedule for teachers, conducts supervision over teachers' discipline, and ensures learning activities are implemented;" while the other two were related to the school leaders' management of school facilities and their attentiveness to community needs and communication. In contrast, after the amendment process, the indicator categories about facilities management and community needs had been displaced by an indicator category about the school leader setting a good example while in school along with a learning-related indicator category, i.e., "Principal develops curriculum and ensures that teachers develop lesson plans and deliver fun learning activities" (see Table B3 in Appendix B for more details). This shift across the top three indicators was consistent across treatment groups.

Similarly, for the teacher service agreement indicators, as summarized in panel (a) of Figure 4, all three treatment groups saw a large increase in the weighting of the indicator category "Teacher works to improve students' literacy and numeracy skills," from an average weight of 2.4 percentage points pre-amendment to 12.4 percentage points post-amendment. In contrast, there were consistent declines in the weighting of the indicator category "Teacher uses positive discipline with students, and avoids any form of verbal or corporal punishments," which is not directly related to student learning. That said, it is worth noting, as suggested by the qualitative case study data, that this reduced emphasis on positive discipline approaches may be due to teachers' beliefs that eliminating corporal punishment led to unmanageable classroom disruptions given that children had been socialized into expecting corporal punishment—rather than to a specific desire to prioritize learning-related indicators. It is also worth noting that "Teacher starts and ends class on time" had, by far, the heaviest weighting both pre- and post-amendment.

Figure 4. Average weighting of selected indicator categories in the teacher service agreements, before and after amendment



Notes. Each panel shows the three indicator categories with the highest overall average post-amendment weighting, out of a total of 8 indicator categories directly related to student learning for panel (a), and 12 indicator categories not directly related to student learning for panel (b). The weightings of all indicators in each teacher service agreement sum to 100 percent. For a more detailed view of pre- and post-amendment service agreement indicator weightings, see Appendix B.

5.1.2 Stakeholders' contributions to the learning process

While the service agreements identify the goals that the village-level principals (service recipients, e.g., families and village government stakeholders) most prioritize in service delivery by village-level agents (frontline providers, i.e., teachers and school leaders), the actions of the principals themselves also provide valuable information about the degree to which these bidirectional accountability relationships were oriented toward learning. Here the quantitative survey data are encouraging, but perhaps less conclusive than the service agreement indicators.

Looking first at parents' willingness to contribute directly to the learning process, there were small increases between the baseline and endline self-report surveys in the average amount of time that parents allocated toward supporting their children's learning at home each day. These increases were comparable across treatment groups, with increases of 10.3 percent in the SAM-only treatment group (from 32.1 minutes daily at baseline to 35.4 minutes daily at endline) and 8.5 percent in both SAM+Score and SAM+Cam. Average increases notwithstanding, there was substantial variation across families (e.g., for SAM-only, standard deviations were 21.7 minutes at baseline and 26.9 minutes at endline). Moreover, there were no significant differences between treatment groups at either baseline or endline, based on t-tests at 5-percent significance.

Similar results were seen in families' average financial expenditure on education, as reported by parents. Comparing baseline and endline averages, there was an increase in parent-reported educational expenditure of 10.8 percent for SAM, 4.7 percent for SAM+Score, and 18.1 percent for SAM+Cam. Comparing the endline averages for each treatment group with that of the control group, the typical family in the SAM-only group spent 6.3 percent more on education than the typical family in the control group; with increases of 5.3 percent and 8.7 percent for SAM+Score and SAM+Cam, respectively. Again, there was considerable variation between families, and no significant differences between groups at endline.[13] Data on other aspects of parental participation in education are available in Appendix A.

Besides families, the village government was willing to contribute to KIAT Guru improvements, but to a limited extent. A high proportion of village heads attended the monthly KIAT Guru meetings: 82.8 percent for SAM, 89.8 percent for SAM+Score, and 93.0 percent for SAM+Cam. Additionally, village governments for 83 percent of treated schools allocated some of the village budget for KIAT Guru implementation. However, when it came to more direct contributions, less than half of the UCs reported that their village government had undertaken any novel educational initiatives to honor the service agreements (44.8 percent for SAM, 47.8 percent for SAM+Score, and 42.7 percent for SAM+Cam). Data on other aspects of local government support for KIAT Guru are available in Appendix A.

5.1.3 Summary

To summarize, data from service agreements and from quantitative stakeholder surveys indicate that there was a shift toward teacher and school leader service indicator categories that were directly related to student learning, suggesting a greater emphasis on learning in *delegation* from principals (i.e., service recipients) to agents (i.e., frontline providers) in the voice & choice relationship at the village level. Additionally, there are slight increases in parents' and village leaders' willingness to contribute to the learning process and/or to the KIAT Guru social accountability process, suggesting an increase in *finance* (i.e., resources) in the voice & choice relationship oriented toward student learning.

All of this supports hypothesis (a), which attributes the gains in student learning across all three KIAT Guru treatment groups to corresponding gains in coherence for learning in the accountability relationships across all three treatment groups. However, the data sources discussed in this section showed neither large nor consistent differences between treatment groups. Thus, they do not lend any

[13] However, there was a significant difference at baseline between SAM+Score (mean=IDR326075.9) and SAM+Cam (mean=IDR298329.5).

support to hypothesis (b), which posits that the larger student learning gains in SAM+Cam are due to greater increases in coherence across accountability relationships in this treatment group (relative to the other treatment groups).

5.2 Coherence between stakeholders in accountability relationships

As noted above, there are many aspects of coherence in accountability relationships for education systems. Having used data from the service agreement indicators and quantitative stakeholder surveys to examine changes in coherence for the purpose of learning, we now weigh the evidence for changes in coherence between stakeholders in accountability relationships, drawing on interview and focus group data from the qualitative case studies in nine KIAT Guru schools.

In this part of the analysis, we interpret a stakeholder's positive view of another stakeholder as an indication of (some degree of) coherence in their relationship, especially when both sets of stakeholders in the relationship express favorable views of each other. This requires the strong assumption that positive views of another stakeholder correspond to the belief that this stakeholder is fulfilling their obligations as principal or agent in an accountability relationship. This assumption may fail for multiple reasons, ranging from dispiriting reasons (e.g., if parents' expectations of teachers are so low that they view teachers positively as long as the latter show up for part of the school day) to more encouraging ones (e.g., if parents have commendably aspirational but unrealistically high expectations of teachers such that they view teachers negatively despite the latter's best efforts). Although we cannot conclusively test the extent to which this strong assumption is met, this part of the analysis nonetheless contributes to the overall understanding of complexity and coherence in community accountability in KIAT Guru schools.

5.2.1 Changes in how favorably stakeholders regard each other

Many opinions expressed in the qualitative interviews and focus groups suggest that KIAT Guru improved the coherence between village-level stakeholders, particularly in their expectations of each other (i.e., *delegation*) and in communication with each other (i.e., *information*), across all three treatment groups. For example, when asked in a midline interview about changes in the school over the last six months, the school leader at the SAM-only school Engkangin said that there had been a change in "motivation for educational cooperation at school":

> Before this, there were still a lot of parents who were ignorant. … There was often the assumption that the school would bear all 100 percent of [the responsibility for] their children, but after we had those socialization meetings then they understood, so they also contribute.

During an endline focus group discussion at the SAM+Cam school Kondok, a community elder said that there was "closeness" in the relationship between the school leader and the UC:

> … if there are any issues that the UC doesn't fully understand, they ask the teachers and the school leader. There's good communication between the school and the UC. They are like partners. What needs to be improved is that from all parties, starting from the school, the committee, and the UC, there needs to be even more socialization and alignment of perceptions so that all can understand.

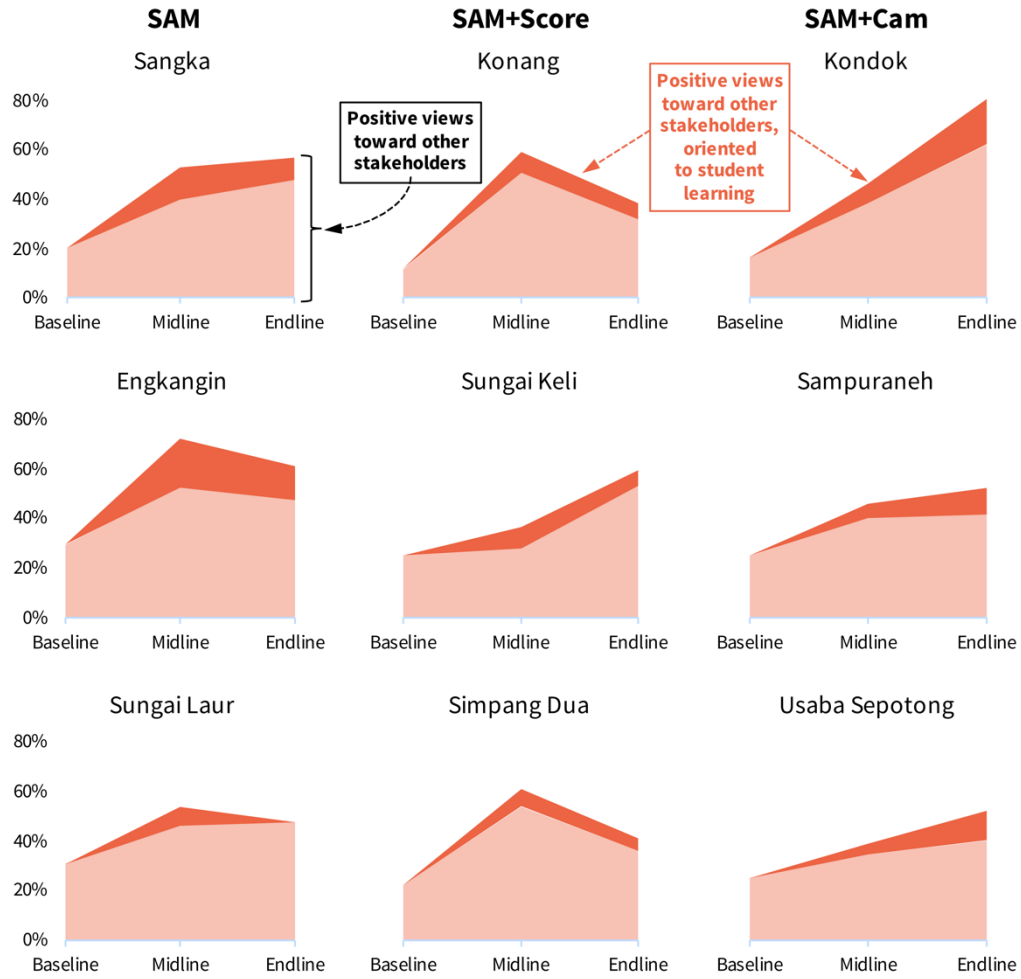
Similarly, during an endline interview, a teacher at the SAM+Score school Konang said: "The UC has become a bridge connecting parents' aspirations with the school."

To develop a more systematic view of the extent to which these positive views represent overall changes in village-level educational accountability relationships, we extracted and classified quotes pertaining to stakeholders' views of each other, as described above in Section 4.2. To begin with a high-level, coarse-grained view, Figure 5 summarizes overall changes over time in the proportion of positive views and positive learning-oriented views among village-level stakeholders expressed in interviews and focus group discussions in the nine case study schools. The graphs in Figure 5 are based on straightforward proportions of positive and positive learning-oriented views among the total count of views expressed,

where each unit included in the count represents one stakeholder group's views of another stakeholder group.[14]

As shown in Figure 5, all nine case study schools saw increases in the proportion of total positive views (i.e., generic positive + learning-oriented positive) between the baseline and endline field visits. To the extent that positive views of other stakeholders may indicate that accountability relationships are functioning effectively, this suggests that there may have been increases in the coherence of accountability relationships as the project progressed. Also, across all nine case study schools, there was an increase in the proportion of positive views that were oriented toward student learning, from 0 percent at baseline in all cases.

Figure 5. The proportion of positive views of other stakeholders expressed in interviews and focus groups in the nine case study schools



Notes. Stakeholders included in the analysis were: students, teachers, school leaders, parents, school committee members, user committee members, village cadres, and village heads. Views of other stakeholders were extracted from interview/focus group transcripts and were classified, for each pair of stakeholders in each round of data

[14] In this broad-brush summary, we use proportions (e.g., the proportion of views that were positive) rather than counts because the total number of units (i.e., the denominator) varies across each round of data collection for each school, ranging from 25 (sets of) views at baseline in Sangka to 51 (sets of) views at endline in Engkangin. This wide variation is due to a several practical reasons—such as participants choosing not to answer certain interview and focus group questions about other stakeholders, participants offering opinions about specific stakeholder groups in response to general questions, or, in a few cases, interviews or focus groups that could not be conducted because the stakeholders in question were unavailable during the field research teams' visits. Across all case study schools, the total number of views included in the classification increased between the baseline and the endline, largely because endline data collection included UC members and village cadres, who had yet to be appointed at baseline.

collection (e.g., teachers' views of students at baseline) as either negative, neutral, mixed, positive, or positive and oriented toward student learning.

However, these increases were not uniform. In four out of the nine schools (Engkangin, Sungai Laur, Konang, and Simpang Dua), the proportion of total positive views declined between midline and endline (while remaining higher than the baseline level). Similarly, in six schools, there was a higher proportion of learning-oriented positive views at midline than at endline. (In Sangka, this decline between midline and endline was such that none of the participating stakeholders expressed learning-oriented positive views of others at endline.) It is worth noting that all six of these schools were in the SAM-only and SAM+Score treatment groups. Put differently, all three SAM+Cam case study schools saw monotonic increases in the proportions of total positive views *and* of learning-oriented positive views in successive rounds of data collection. These patterns must be interpreted with a great deal of caution given the very small sample of schools involved, as well as the fact that the absolute proportions of positive views were lower in some SAM+Cam schools than in some schools in the other treatment group (e.g., among the three schools in Landak district, the SAM+Cam school Sampuraneh had lower proportions of total positive views at endline than the SAM-only school Engkangin and the SAM+Score school Sungai Keli). Nevertheless, these patterns at least fail to contradict hypothesis (b) that SAM+Cam was the most effective treatment at raising student learning outcomes because it was most effective at improving the coherence of accountability relationships.

For a more granular view, Figure 6 shows the classifications of principals' and agents' views of each other in the voice & choice relationship and the management relationship. (For brevity, this figure does not include principals' views of other principals in the voice & choice relationship; e.g., it does not include students' views of the school committee, nor parents' views of students, etc. The full classifications of all stakeholder pairs' views of are available in Appendix C.) The figure shows a general trend toward stakeholders' views becoming more positive in successive rounds of data collection. For example, representative views held by parents at the SAM+Score school Konang were decidedly mixed at baseline:

> In terms of teacher quality, we see there are some teachers who have never gone into the school. … Teachers aren't very active in teaching, to the point where children are always playing. We parents feel disappointed. In my view, [only] 80 percent are active, we don't know why.

At midline, parents at the same school had generally positive views of teachers, such as:

> The change in this school is that teachers come on time. Secondly, the way they teach has changed. Like in terms of behavior, previously my child wasn't neat [i.e., didn't dress neatly] because they were only in second grade, but now their teacher has given encouragement and shown examples, so they now look neater when they go to school.

By the endline data collection, parents' views of teachers in Konang were not only positive, but also clearly oriented toward learning. For example, one parent said, "With KIAT Guru, teachers are also able to open their eyes well to educate children well," later commenting that "The impact [of KIAT Guru] is that children's learning outcomes have gone up."

Figure 6. Principals' and agents' views of each other in the voice & choice and management relationships at baseline, midline, and endline

5.2.2 *Disparities between principals' and agents' views of each other*

Despite this overall increase in positive views and decrease in negative views, the snapshot in Figure 6 also indicates one distinct shortfall in the coherence (broadly construed) of village-level accountability relationships: the fact that principals' views of agents tended to be much more favorable than agents' views of principals. For example, while parents' (i.e., one set of principals) views of teachers (i.e., agents) in Konang markedly improved, as illustrated in the quotes above, teachers at the same school were more circumspect in their views of parents. For example, one teacher commented at a baseline focus group discussion that:

> Almost all of the children don't have time to study at home, except for the children of [village/government] officials. There's almost no time, with parents just trying to earn money. The problem of learning is just handed over to the school. At home, teaching is only limited to things like manners. How could they teach their children, when they can't read?

This same teacher was somewhat more positive during a midline interview, noting that, "there's been a change in parents' motivation to make sure that their children come to school." However, he also commented that roughly half of all parents failed to put their signatures on their children's homework, a required action under the KIAT Guru service agreements that parents support their children's learning at home. He attributed this negligence to "the reason being that some parents still don't have awareness." In his endline interview, this teacher's mixed views about parents persisted:

> It's true that a portion of parents, 60 percent, have fulfilled the service agreement. … Some parts of the community don't yet understand their roles and responsibilities in supporting children at home.

These mixed views of parents in the qualitative case study schools were echoed in the quantitative surveys across all KIAT Guru schools, where roughly half of all teachers answered affirmatively when asked whether they think that parents have a good level of involvement in their children's education. There was little variation in this proportion between the baseline and endline surveys, and between treatment groups (ranging from 46.9 percent across in SAM+Cam schools at baseline to 51.7 percent in both SAM+Score schools at baseline and SAM-only schools at endline). However, neither the qualitative case study interviews and focus group discussions nor the quantitative surveys allow us to disentangle the degree to which these less positive views that teachers held of school leaders (in the management relationship) and of other village stakeholders (in the voice & choice relationship) were due to these other stakeholders being less effortful or effective in fulfilling their parts of the service agreements, versus teachers simply being more critical or exacting in their judgements of other stakeholders. Power dynamics related to the latter point are explored further in Section 5.3.

5.2.3 Summary

To summarize, this analysis suggests firstly that stakeholders viewed each other more positively as the project progressed and secondly that these positive views became more oriented toward student learning. This aligns loosely with hypothesis (a) that attributes the student learning gains across all treatment groups to cross-treatment increases in the coherence of accountability relationships for student learning. However, as noted above, there was limited concurrence between principals' more positive views of agents' and agents' less positive views of principals; such that the SAM+Cam school Kondok was the only school where principals and agents in both the management and voice & choice relationships had generally positive views of each other at endline, as shown in Figure 6.

As for hypothesis (b), which proposes that there were greater increases in the coherence of accountability relationships in SAM+Cam, there is some limited support in the coarse-grained summary shown in Figure 5, in that the three SAM+Cam schools were the only case study schools that saw uniform increases in the proportions of both generic positive and learning-oriented positive views over time. Yet these advantages of SAM+Cam schools are neither large nor absolute, nor are they clearly apparent in the more granular representation in Figure 6. In short, at this level of detail, there is no distinct relationship between each treatment group's impact on student learning and changes in coherence between stakeholders in accountability relationships. This absence of a clear relationship applies not only when comparing student learning gains across treatment groups, but also when looking at changes in student learning outcomes and coherence between stakeholders across the nine individual case study schools. In the next section, we move to a greater level of specificity: looking not only at overall coherence in stakeholders' views of each other, but at design elements within stakeholder relationships.

5.3 Coherence between design elements within the voice & choice relationship

In this part of the analysis, we examine how the differences between the treatments manifested in the design elements of the voice & choice relationship, and how this affected the degree to which this relationship was coherent for learning. Accordingly, we focus on hypothesis (b), i.e., that SAM+Cam had the largest student learning gains because it also had the greatest improvements in coherence for learning.

As noted in Section 3, KIAT Guru affected both the management relationship and the voice & choice relationship. That said, in this section we focus on the village-level voice & choice relationship because the qualitative case studies offer particularly rich data from interviews and focus groups with stakeholders within this relationship. We posit that SAM+Cam was the most effective treatment arm because it was the most effective in coherently aligning the different elements of the voice & choice relationship. A key design element in this coherence analysis is *support* to strengthen the accountability relationship. In the remote schools where KIAT Guru was implemented, the hierarchical differences between higher-status teachers and school leaders (agents) and lower-status parents and community members (principals) meant that UCs in SAM-only and SAM+Score were insufficiently empowered for a strong accountability relationship between these committees and teachers/school leaders. In SAM+Cam, the cameras provided UCs with teacher attendance data that was neutral rather than observer-dependent, which not only improved the quality of *information* but also gave *support* by boosting the perceived legitimacy of UCs' evaluations of teachers.

In this section, we first draw on the qualitative case study data to show the extent of the power imbalance between teachers and community members, and to compare the coherence between design elements within the voice & choice relationship across the three treatment groups. We then return to the process monitoring and quantitative survey data to weigh the evidence for two alternative hypotheses that could weaken the validity of our explanation.

5.3.1 Power dynamics in the voice & choice relationship

As in similar settings, teachers and school leaders in the remote villages where KIAT Guru was implemented tend to have significantly higher social status than the children and families whom they serve (see Section 3 for some examples). In other words, the agents in the voice & choice relationship hold more power than their principals. This can affect the local community's ability to hold teachers accountable. For example, a Sangka UC member said, in an endline focus group discussion, "We have never reprimanded them about their work [because we're] afraid they would be offended." This fear may be well-founded: in Sungai Laur, a teacher who disagreed with their scorecard ratings angrily confronted the UC "to the point where the whole village knew" about the conflict (Sungai Laur midline focus group with UC members).

The main reason why teachers are more empowered than most village-level stakeholders is due to their education. This is partly a question of social status. According to a parent in an endline focus group in Konang:

> KIAT Guru also had some negative impact because, actually, UCs don't have the right to evaluate teachers. As far as we know, only inspectors can evaluate. … In my view, teachers might just say to themselves: you farmers, on what basis could you evaluate me? They might just feel, in their hearts, that the community isn't eligible to do evaluations.[15]

Besides social status, education levels also affect local perceptions about capacity for accurate evaluation. For example, the village secretary in Sangka said in an endline interview that some UC members "don't suit the standard" because, in his opinion, all UC members should at least have completed upper secondary school; a sentiment that was echoed by a Sungai Laur parent in a midline focus group. To some extent, this emphasis on formal educational qualifications reflects real capacity issues in UC scorecard appraisals, which were mentioned by village stakeholders in all three treatment groups. For example, a school committee member in the SAM-only school Sangka said that "the service agreements

[15] See Broekman (2015) for similar views expressed from the teacher's perspective in a separate study in Indonesia.

don't make sense, because it's the teachers who understand how to teach the children, while the UC isn't able to," a sentiment echoed in an endline interview with a teacher in the SAM+Cam school Usaba Sepotong. Also, teachers in an endline focus group in the SAM+Score school Sungai Keli remarked that UC members didn't fully understand their duties, partly because their relatively low education levels hampered their ability to interpret the language of the service agreements. Specific errors in appraisal mentioned by teachers and school leaders include a teacher being penalized for an action that had in fact occurred the year before KIAT Guru was instituted (Simpang Dua midline focus group with teachers), and a teacher scorecard rating of 113 percent (Simpang Dua midline interview with school leader).[16]

Another factor that tilted the power balance toward teachers was their direct influence over individual children's school experiences. For example, UC members expressed the fear that agreeing to serve on the UC would result in teachers threatening not to promote their children to the next grade (Sungai Keli midline focus group with UC members) or in teachers not paying any attention to their children during classroom lessons (Konang midline focus group with UC members). In Sangka, one UC member said in an endline focus group that children were afraid of speaking to the UC because teachers had allegedly threatened to beat any children who gave information to the UC about teachers' work. (That said, it is also important to note that children in KIAT Guru schools also had agency. They were not necessarily passive service recipients. In Sungai Laur, some children had intentionally lied or exaggerated unfavorably about their teachers to the UC; as mentioned separately by teachers in a midline focus group, the school leader in both midline and endline interviews, and the village cadre in an endline interview. Some children also took advantage of the scorecard prohibitions on corporal punishment to actively provoke teachers and behave rudely, because "they know that if the teacher gets mad, if they pinch me, I can tell the UC and ask them to fire that teacher," as an Engkangin UC member related in an endline focus group.)

Besides teachers' educational status and their sway over individual children's experiences, another factor in the power dynamic between principals and agents was legitimation from higher authorities, or the lack thereof. Such legitimation is highly valued. This was apparent during the midline round of data collection in Sungai Laur, which took place shortly after the chair of the UC had failed to travel to the national capital, Jakarta, for some KIAT Guru capacity building sessions at the invitation of the central government. This omission was mentioned by multiple stakeholders, from other UC members who said that they were "deeply disappointed"; to teachers who said that they had "all lost out" on the information that would have been shared in Jakarta; to the village head who said that the errant UC chair "actually isn't suitable to be the leader"; to a non-UC parent who said they had been "proud" when the UC chair was invited to Jakarta but "disappointed" by his failure to attend. In the field research team's daily report, they note that the UC chair was regarded as having "besmirched the names of the school and the village" in failing to pursue this connection to higher administrative levels.

The interplay between village-level status hierarchies and legitimation from higher administrative levels was evident in this exchange between the Usaba Sepotong (SAM+Cam) school leader and a field researcher in an endline interview:

> *School leader*: Because of their lack of knowledge, we are evaluated by people who have a lower education level than us. That's what the teachers objected to, having former students evaluate us. So there needs to be tolerance from them [i.e., lenience from the UC in appraisals].
> *Interviewer*: Have you said this to the UC before?
> *School leader*: Yes, I have.
> *Interviewer*: The UC wasn't angry?
> *School leader*: No, but the teachers have been angry with the UC. I said to the teachers, whatever scores they give, we will accept it, the important thing being that it's not below 90.
> *Interviewer*: What was the motivation for teachers [to accept it]?

[16] Another cause of UCs' capacity issues was the fact that UC members were operating on a largely voluntary basis. Many UC members were farmers who could not leave their fields to monitor teachers' arrivals and departures at school firsthand. Instead, they had to rely on school documents such as the teacher attendance register, as well as word of mouth from other children and other villagers—all of which could be unreliable (as noted by Konang UC members in a midline focus group).

> *School leader*: Because it's a government program, it must be done, and whatever happens it's important that we submit.

This suggests firstly that the teachers would not have accepted the UC's scorecard appraisals as legitimate on their own merits, but also that the authority conveyed by KIAT Guru being "a government program" was a crucial support in the UC–school accountability relationship. In turn, the UC members appear to have accepted this mantle of this government legitimation. During a midline focus group at the same school, UC members spoke of a teacher who had refused to have her picture taken using the KIAT Kamera application and who had "once said, while pointing with her hand, that 'you're my former students.'" One UC member deemed these actions to have "lower[ed] our status as UC members"—implying both that relative status matters and that UC members regard themselves as having special status within the village hierarchy. As we show below, these power dynamics substantially influenced the effectiveness of the three different KIAT Guru treatments.

5.3.2 Design elements and coherence in the voice & choice relationship

As discussed in Section 3 on the conceptual framework, the design elements within any given accountability relationship—as well as the coherence between those design elements—are pivotal to effective functioning of the relationship. We now examine design elements within the voice & choice relationship across the three KIAT Guru treatment groups. These design elements are summarized in Table 2, as contrasted with the status quo in the control group.

Table 2. Design elements of the voice & choice relationship between families/community members (principals) and teachers/school leaders (agents), for each KIAT Guru treatment arm

Legend: **+** improvement relative to the control group, given the specific context (where the agents are more empowered than the principals, etc.)

	Control	SAM-only	SAM+Score	SAM+Cam
Delegation What do principals want agents to do?	There is no clear delegation from the community to teachers and school leaders (besides the government-specified minimum service standard).	**+** The community specifies service standards, as contextualized in service agreement indicators, for teachers and school leaders.	**+**	**+**
Finance What resources do principals provide to agents?	Parents sporadically contribute time, effort, and money to the learning process.	**+** Parents, village governments, and the community contribute various amounts of time, effort, and money to the learning process, and are regularly reminded to do so via SAM processes.	**+**	**+**
Information How do principals know that agents are performing?	Community monitoring is informal and ad hoc.	**+** User committees monitor teachers and school leaders using community perceptions and learning assessments, based on the service agreements.	**+**	**+** In SAM+Cam, user committees also use cameras.
Motivation How does agents' wellbeing depend on performance?	Teachers and school leaders face some social rewards and penalties, but these are not systematic.	**+** All teachers and school leaders face social rewards and penalties aligned around the social accountability mechanism.	**+** In SAM+Score and SAM+Cam, allowance-eligible teachers and school leaders also face financial penalties.	**+**
Support How is the accountability relationship strengthened?	School committees receive little (if any) funding and technical support.	**+** Communities have awareness-raising sessions about rights, responsibilities, and quality in education. User committees have commissioning decrees and some funding from the village government, and visits from the project facilitator.	**+**	**+** In SAM+Cam, the camera also provides more authority to user committees.

Beginning with *delegation*, or the specification of what principals want from agents, the social accountability mechanism within KIAT Guru improved the clarity and the degree of local consensus about performance expectations for teachers. In the control group, dominant expectations of teachers

came from the central government's minimum service standards for all teachers, which are general, rather than specifically calibrated by principals and agents at the local level.[17] In contrast, the process of discussing and agreeing on service agreements by village-level educational stakeholders in KIAT Guru treatment groups yielded much more specificity and immediacy in performance expectations for teachers. For example, a teacher in the SAM+Cam school Kondok said in an endline interview that

> For me personally, I don't run away from the duties of serving as a teacher. These come from service agreements from teachers, from the community, from parents, so that's good for me. The benefit for me is that I can do my job well, as a guide for me.

This clear, relational, personally endorsed delegation was due to the SAM process rather than the use of cameras for performance pay—as was evident in this teacher's remark, immediately before the quoted statement, that "if it's the camera on its own, [some] people intentionally won't operate the camera." Teachers across the three treatment groups similarly affirmed the value of service agreements and scorecards as guides that orient them toward their duties.

Another shift that was largely consistent across treatment groups was improvements in *finance*, i.e., greater resource provision from families, the community, and the village government toward children's education. In the words of a Simpang Dua (SAM+Score) teacher in a midline focus group:

> In the past, before there was a UC, it was difficult to ask parents for help. For example, if we wanted to do something they'd complain, but now they are willing to contribute to providing uniforms. This means they have become open and supportive of school activities.

These anecdotal reports of improvements in resource provision from the community were borne out in the quantitative survey data, 5.1.2with increases across treatment groups in the amount of time and money that parents contributed to the learning process (see Section 5.1.2). Still, there was considerable variation in this increased provision, which affected teachers' mixed views of parental contributions (see Section 5.2.2). Besides parents, village governments also increased their educational contributions under KIAT Guru. Again, there was wide variation in the resources that village heads allocated to support KIAT Guru (World Bank, 2020), with roughly half of all village governments starting new educational initiatives to honor the service agreements (see Section 5.1.2).

There were, however, more distinct differences between treatment groups in how KIAT Guru affected the design elements of motivation, information, and support. For *motivation*, two of the treatments involved performance-based deductions from a special allowance for eligible teachers working in remote and otherwise deprived areas. Given that the special allowance was roughly equal to a teacher's base salary, these deductions could affect agents' (i.e., teachers' and school leaders') wellbeing considerably. Both principals and agents in SAM+Cam and SAM+Score schools agreed that the performance-based deductions were a significant motivator. For example, when asked about what had changed at the SAM+Score school Konang, UC members said during an endline focus group discussion that:

> *UC member 1:* Before KIAT, teachers were not very active.
> *UC member 2:* Looking at it as a community representative, we were afraid when a teacher was late. We couldn't say too much. Now there is definitely a change in teacher attendance. Previously, they were not very aware of their duties, now they're more hardworking.
> *UC member 3:* Maybe they're afraid of the program. This links to their salary.
> *UC member 4:* Because of this UC duty, watching teachers from a distance early in the morning, so they are afraid. They arrive on time because there's someone monitoring.
> *UC member 5:* … They are afraid that their salaries and benefits will go down.

In some cases, teachers themselves appreciated the role of performance pay in teacher motivation. When asked whether the allowance deductions were effective in improving teacher performance and student learning, a teacher at the SAM+Cam school Usaba Sepotong said in an endline interview that: "If it's

[17] In remote areas, this conceptual distance in *delegation* often goes hand-in-hand with physical distance that affects *information*, because district supervisors and other education bureaucrats who can facilitate accountability and support are often headquartered in district capital cities and may lack adequate transport resources to visit the schools under their purview, as documented in a study of Papua and West Papua provinces in Indonesia (UNICEF, 2012).

reasonable, this deduction in value is good for teachers. Without it, then there's nothing to control the teachers."

Social rewards and penalties also helped to align teacher motivation with agree-upon delegation. Crucially, this reputational pressure was deeply felt even in the treatment arm that lacked financial incentives. In an endline interview, the school leader in the SAM-only school Sangka observed that KIAT Guru had led to "a new spirit" among both teachers and students, attributing this change to the fact that:

> There's a shared commitment to strengthening teachers' attention and concentration. If this is ignored, then the community will draw unfavorable conclusions. Now teachers are afraid of community assessment through the UC.

Teachers in the SAM-only school Sungai Laur described the extent of this social pressure in a midline focus group discussion, directly linking the reputational pressure to central government legitimation of KIAT Guru:

> *Teacher 1:* Because [KIAT Guru scores] are reported to the center, its influence is on morale.
> *Teacher 2:* They said it has no effect on our allowance, but it has to do with morale, loyalty to duty, good conduct.
> *Teacher 3:* Like when I got scolded at the UC meeting, I felt crushed as a human being. … Those scores go directly to the District Education Management Unit, where all of them are my friends, so now they know that [Teacher 3] from that school is bad; so my friends from my schooldays have seen it.

This social pressure was also pivotal in improvements in practice from uncertified teachers in SAM+Cam and SAM+Score schools who were ineligible for the teacher special allowance and, accordingly, were unaffected by the financial disincentive.[18] One such teacher from the SAM+Score school Konang spoke about the impact of KIAT Guru on his sense of responsibility:

> Before the program, sometimes I just did as I pleased. With this program, I am increasingly aware of what our duties as teachers should be. I think this is a very extraordinary impact … In the beginning, I was the one who had the lowest score, so I reflected and thought that my friends were chasing points, chasing 100, because of the allowance. But then I was still motivated because even though I don't get the allowance, this had become my main duty whether I liked it or not, so I had to improve myself.

Given the extent of this social and reputational pressure on teacher motivation, we cannot conclude that the different financial incentives of the treatment arms were sufficient to explain the greater effectiveness of SAM+Cam.[19]

Finally, we turn to the two remaining design elements, information and support. The cameras in SAM+Cam gave this treatment group a distinct advantage in both of these elements (see Table 2). In terms of *information,* teachers and UC members across all three SAM+Cam schools said that they appreciated the cameras because they were harder to manipulate than other information sources, whether

[18] The KIAT Guru impact evaluation (Gaduh et al., 2020) found that KIAT Guru negatively affected the attendance and effort of teachers who were ineligible for the teacher special allowance—and that this negative effect was significant among ineligible teachers in the SAM+Score treatment group. However, among the small subsample of schools in the qualitative case studies, there were no obvious between-treatment differences in the sentiments of teachers who were not eligible for the allowance. Across all three treatment groups, stakeholders noted that these ineligible teachers, who were subject to KIAT Guru appraisals despite receiving none of its financial benefits, felt disappointed and jealous of their colleagues. This also manifested in the non-performance-pay SAM treatment because many stakeholders mistakenly associated the provision of teacher special allowance with KIAT Guru, although they were distinct policy programs. The point here is not to support or contradict the impact evaluation observation of perverse effects among allowance-ineligible SAM+Score teachers, not least because the sample size of three schools per treatment arm would hardly permit such conclusions. Rather, the point is to show that the KIAT Guru social accountability mechanism generated non-financial incentives that substantially influenced teachers' behavior, whether or not they were subject to performance pay (and whether the presence/absence of this financial disincentive was due to individual-level circumstances or to school-level treatment assignment).

[19] We further explore this alternative explanation related to incentive structures in Section 5.3.4.

by lying or by lobbying. Notably, a few frontline providers (e.g., the Sampuraneh school leader in his midline interview) mentioned that the cameras also reduced manipulation by UC members who wanted to preferentially inflate certain teachers' scores. These favorable perceptions of the cameras for improving information on teacher attendance were also reflected in the quantitative survey. On average, SAM+Cam UC members rated the cameras 9.42 out of 10 for helpfulness in evaluating teacher attendance, with two-thirds of UCs choosing the highest possible rating.

In addition to their immediate role in improving the quality of information, the cameras also played an indirect but vital role in *support*. Specifically, the cameras strengthened the voice & choice relationship by lending authority to the relatively disempowered principals in their interactions with higher-status teachers. According to a project facilitator who supported KIAT Guru in Kondok and five other schools:

> The most accurate [scores] are those using cameras, because teacher attendance is really evidenced with those cameras. They are more scared of the cameras than of the UCs. Especially when it is connected to their allowance.

That is, the accuracy of the camera-mediated information, together with the threat of allowance deductions, added legitimacy to the social accountability process. This was echoed by a Sampuraneh school committee member in an endline interview, who observed that teachers were afraid of the camera-wielding UC but had no such fear of the preexisting school committee. The cameras in SAM+Cam rebalanced power dynamics toward the community as represented by the UC. The fact that this was a question of power, legitimacy, and status rather than simply a question of financial penalties is apparent in the way teachers spoke about the cameras. For example, a contract teacher said in the endline teacher focus group in Usaba Sepotong that, "Once, I didn't want to be photographed, because I considered it an oppression of teachers." This teacher was ineligible for the teacher special allowance and thus unaffected by the financial incentives. Yet they viewed the camera not only as an instrument of information, but also an instrument of authority.

Although UCs under SAM+Cam enjoyed legitimation from the visibly official, purposefully designed, tamper-proof smartphone cameras, there were also other sources of support that strengthened the voice & choice relationship in all three treatment arms. First, as described in Section **Error! Reference source n ot found.**, the village government formally appointed the UCs and the village cadre and allocated some funding for the implementation. Second, project facilitators visited the schools at regular intervals to coordinate meetings, provide on-the-job training for UCs and the village cadre, and facilitate the communication of challenging issues between stakeholders. These supports positioned the UCs, and the village-level voice & choice relationship more generally, to be far more effective and empowered than the long-established but circumscribed school committees.

Notwithstanding these sources of support that were common across treatments, our conclusion after examining the data is that SAM+Cam was the only treatment group in which the design elements in the voice & choice relationship were coherent for improvements in student learning in the context of the principals and agents in these remote village schools.[20] Across all treatment groups, the service agreements improved *delegation* between principals and agents, particularly in the clarity, contextualization, and emphasis on student learning. The social accountability mechanism also mobilized greater contributions of *finance* and resources from the community toward educational provision, which made teachers more amenable to the more demanding expectations. The monthly appraisal meetings, along with various instruments for collecting information on teachers' and school leaders' performance, also improved the quality of *information* in the voice & choice relationship. This was particularly true of SAM+Cam with its manipulation-resistant teacher attendance records, which teachers were less likely to question—unlike the scorecard appraisals, which some teachers did question extensively due to their doubts about UCs' appraisal capacity, as noted above.

However, in SAM-only and SAM+Score, the voice & choice relationship still fell short of coherence in the intervention contexts. Specifically, the improved delegation via the social accountability mechanism

[20] This also aligns with Dewi's, Sharon's, and Usha's firsthand observations of KIAT Guru as they were involved in the project as researchers throughout its implementation.

placed greater demands on the relationship than it was able to deliver. Lacking technical and symbolic *support* from the camera, UC members in SAM-only and SAM+Score could not marshal enough authority, as lower-status villagers, to credibly hold teachers accountable to the new and more demanding performance expectations. In SAM-only schools, the imbalance was worsened because UCs could not deploy financial disincentives, such that the performance expectations that they delegated to teachers had too little influence over teachers' *motivation*. In SAM+Score schools, the imbalance was worsened because UCs were seen as having too much influence (in the context of established local power dynamics) on the financial sources teachers' motivation, such that teachers challenged UCs' perceived overreach. Given these imbalances between the design elements, the village-level voice & choice relationships in SAM-only and SAM+Score schools were not coherent enough to sustain the stakeholder cooperation needed to improve student learning. In contrast, the new demands on UCs in SAM+Cam were balanced with support, motivation, and information, such that the voice & choice relationship was adequately coherent for improvements in student learning.

From the previous two paragraphs, it is apparent that our argument for the greater effectiveness of SAM+Cam is fairly complex. The argument requires consideration not only of several design elements that interact with different degrees of coherence, but also of context-specific social hierarchies and power dynamics. To stress-test this argument, we weigh the evidence for two alternative, more parsimonious hypotheses. As shown in Table 3, the first hypothesis focuses only on the design element of information, which prioritizes the technical aspect of the accountability relationship, rather than the social aspects emphasized in our hypothesis. The underlying assumption here is that for agents to perform well, it is sufficient for principals to monitor them accurately. In turn, the second alternative explanation focuses on the design elements of delegation, motivation, and support. This alternative explanation prioritizes agents and the incentive structures that they face. The assumption here is that for agents to perform well, it is sufficient to ensure that they face clear incentives. As noted in Section 4.2 on the analytical approach, we weigh each alternative hypothesis by looking at both the strength of evidence for the alternative hypothesis itself, and for its explanatory power in accounting for other findings from the field data.

Table 3. Design elements in the village-level voice & choice relationship that are incorporated into hypotheses for the greater effectiveness of SAM+Cam

	Our hypothesis	Alternative hypothesis 1	Alternative hypothesis 2
Delegation	✓		✓
Finance	✓		
Information	✓	✓	✓
Motivation	✓		✓
Support	✓		

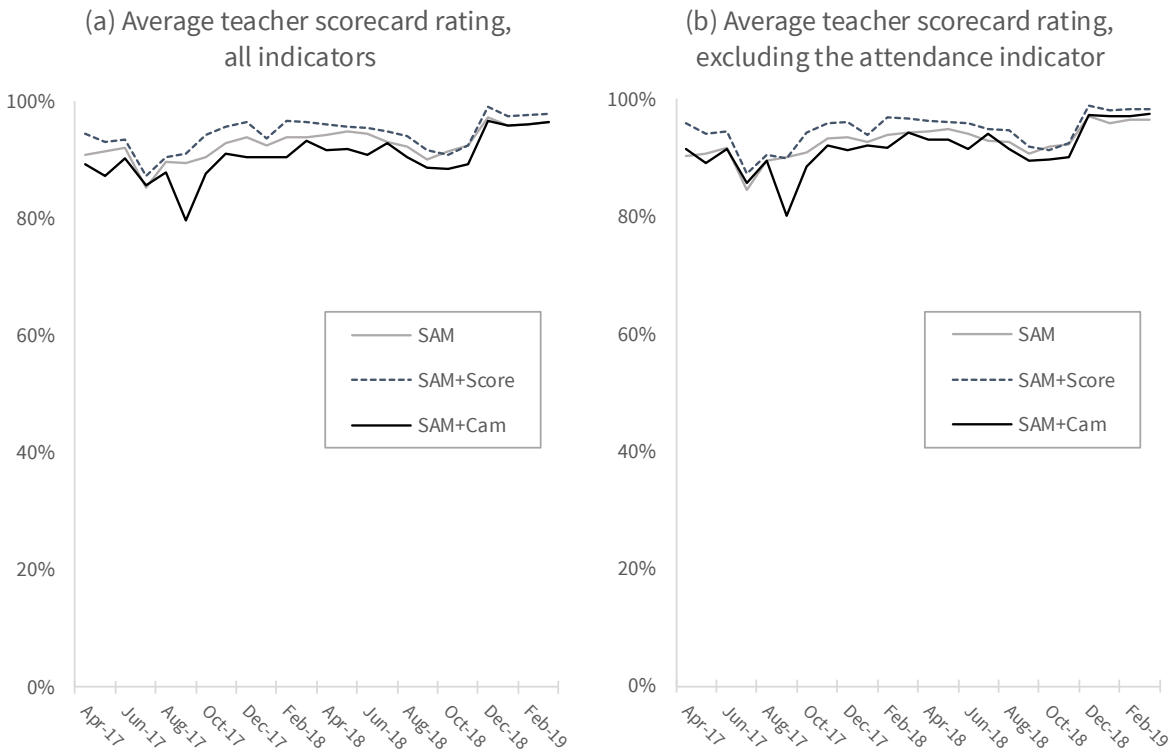
5.3.3 Alternative hypothesis 1: Not coherence, but better information

The first alternative hypothesis for the comparative success of the SAM+Cam treatment focuses on information. Specifically, it is possible that the improvements under SAM+Cam were driven primarily by the higher-quality teacher attendance information from the cameras. This would imply that the cameras did not confer symbolic *support* for the UCs' authority; and also that the driving factor was not coherence between all the design elements, but rather the superiority of one design element on its own.

As noted above, the theory of change here is that technical improvement in monitoring was sufficient for improving teacher performance and student outcomes. This implies that the binding constraint in the other treatments—or, at least, in SAM+Score, which also had a performance pay element—was the quality of information. A possible mechanism here that teachers' financial penalties in SAM+Cam were based on this higher quality, more neutral information source that made the treatment more acceptable to teachers. In contrast, the less reliable scorecard ratings, as in SAM+Score, resulted in teachers becoming dissatisfied with KIAT Guru and, consequently, demotivated (Gaduh et al., 2021; World Bank, 2020).

The empirical data offers some suggestive evidence that UCs under SAM+Cam had higher-quality information, and that the quality of information mattered. For example, one piece of suggestive evidence that SAM+Cam had better information is that average teacher scorecard ratings were slightly but consistently lower—that is, less inflated—in SAM+Cam than in SAM+Score and SAM-only, as shown in Figure 7. This is true whether or not the attendance indicator is included in the weighted average ratings. It is also true whether aggregating across all monthly appraisal rounds or looking at monthly within-treatment averages across the full duration of the study (with minor deviations from this pattern in the month-by-month case).[21]

Figure 7. Average teacher scorecard ratings by treatment arm, with and without the attendance indicator



However, for this alternative hypothesis to be persuasive, the field data also need to indicate that information quality in SAM+Cam improved sufficiently to alleviate what would otherwise have been a binding constraint, as experienced in the other two treatment groups. The data do not show this conclusively. For one thing, it is not obvious that the information available to UCs under SAM+Cam was substantially better than that available to UCs under the other two treatments. Despite the cameras to verify teacher attendance registers, the teacher attendance indicator in SAM+Cam schools was not immune to subjectivity—a fact noted by stakeholders in all three SAM+Cam qualitative case study schools. In a midline focus group discussion, an Usaba Sepotong teacher reported that:

> The other day, I didn't want to use the camera because I actually did go home early. A user committee member said I could just go home first and take a photo later, but I didn't want to tell a lie.

Similarly, a Kondok teacher initially noted in an endline interview that "if it's only manual [i.e., written recording of attendance], sometimes teachers aren't honest, but if it's with cameras then it can't be

[21] For aggregated averages across all intervention months, the average scorecard rating including the attendance indicator was 92.6% for teachers under SAM, 94.3% for SAM+Score, and 90.4% for SAM+Cam. For weighted average ratings after removing the attendance indicator, the average ratings were 92.7% for SAM, 94.7% for SAM+Score, and 91.6% for SAM+Cam. In Figure 7, the dip in average scores in SAM+Cam schools coincides with the first round of assessment following the first round of amendments to service agreement indicators; however, it is unclear why there was a sharp drop in SAM+Cam average scores but not the other treatment groups.

manipulated"—before adding, on reflection, that, "if it's just the camera alone, then people can purposely not operate the camera." A Sampuraneh UC member said in an endline focus group that they would prefer it if KIAT Guru did not use cameras, because they preferred a more relational, heartfelt accountability, and that there was always the possibility that the camera would fall out of use—also noting that, "It may be a camera, but those who operate it will always be human." It is worth noting that human manipulation of ostensibly tamper-proof monitoring equipment—both through arbitrary exemptions and purported "failures" to use the equipment—was also a factor in the eventual failure of a performance-pay-for-attendance program for nurses in Rajasthan, India (Banerjee, Duflo, & Glennerster, 2008)

Additionally, process monitoring data from across all KIAT Guru schools also cast doubt on the proposition that SAM+Cam scorecard ratings were more rigorous than in the other treatment arms. Table 4 shows correlations between school-level average teacher scorecard ratings at endline (both with and without the camera-supported attendance indicator) and average student learning outcomes in both language and mathematics at endline as well as baseline-to-endline changes. If SAM+Cam scorecard ratings were of higher quality than the scorecard ratings of the other treatment groups, then we would expect a consistently stronger correlation in SAM+Cam between these ratings and the desired outcome, i.e., student learning levels. Instead, there are no consistent patterns. For the scorecard ratings that include the attendance indicator, SAM+Cam schools do have a stronger association in the expected direction with averages student learning outcomes—but these correlations are not large (0.260 for language and 0.249 for mathematics). Moreover, this association does not hold for changes in scorecard ratings and in learning outcomes.[22] Neither does it hold when the teacher attendance indicator is excluded, suggesting that any gains in information quality were limited just to the camera-verified indicator, without any spillovers to other indicators. In fact, school leaders in SAM+Cam schools gave UCs slightly lower ratings for their ability to choose scorecard indicators that are important to children's learning (7.21, on a 10-point scale) compared to their counterparts in SAM-only (7.53) and SAM+Score (7.39). If the average UC in the other two treatment groups was at least as competent as the average SAM+Cam UC in appraising teacher performance on more complicated behaviors beyond merely showing up, it seems unlikely that the camera-mediated improvements in teacher attendance information were sufficient to alleviate a posited binding constraint in the quality of information that the other treatment groups hypothetically faced. All of this weighs against the argument that better-quality information, in and of itself, was the main driver behind the larger student learning gains in SAM+Cam.

Table 4. Correlations between school-level student learning assessment scores (in language and mathematics) and endline teacher scorecard ratings (with and without the attendance indicator)

		Endline teacher scorecard ratings, *all indicators*			Endline teacher scorecard ratings, *excluding the attendance indicator*		
		SAM-only	SAM+Score	SAM+Cam	SAM-only	SAM+Score	SAM+Cam
Student learning, *language*	*endline*	-0.208*	-0.239*	0.260**	-0.234*	-0.130	0.097
	change	0.184	0.224*	0.158	-0.368***	-0.101	0.117
Student learning, *math*	*endline*	-0.339***	-0.127	0.249**	0.122	0.252**	0.113
	change	0.183	0.178	0.039	0.135	0.177	0.019

Notes. 'Average' denotes the average within-school student learning assessment/teacher service agreement score at endline. 'Change' denotes the within-school percentage change in student learning assessments (comparing endline to baseline) and teacher service agreement scores (comparing endline to community-assessed retrospective ratings of how teachers performed prior to the start of the project). */**/*** denotes 10/5/1 percent significance levels.

[22] Note that the changes in teacher scorecard ratings were based on comparing the average endline teacher scorecard rating with community members' retrospective evaluations of how teachers performed prior to the inception of KIAT Guru. Although these retrospective evaluations were recorded only a few months after the intervention started, they are nonetheless subject to recall bias.

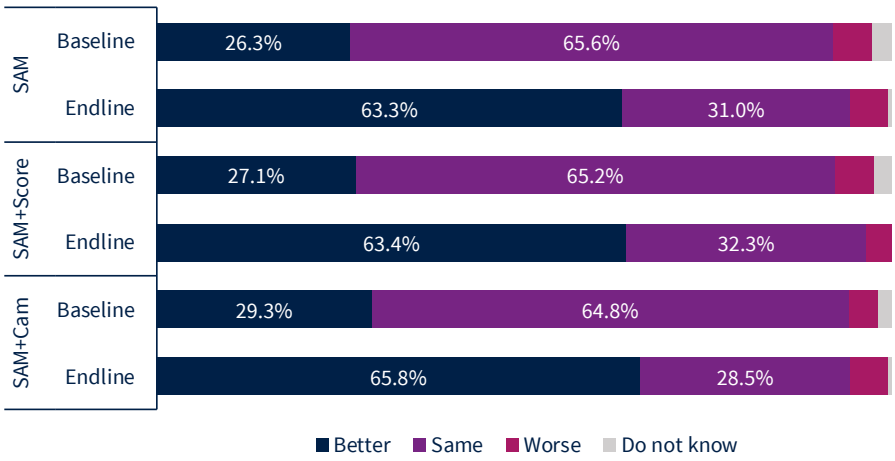
Besides the limited evidence for the alternative hypothesis that information was the most important design element in the greater effectiveness of SAM+Cam, another challenge to this hypothesis is its limited explanatory power. Specifically, this alternative hypothesis cannot, on its own, account for the fact that SAM+Cam led to multifaceted improvements in teacher performance above and beyond attendance rates, which were the only area for which information quality improved. For example, a UC member at the SAM+Cam school Kondok said in an endline focus group that:

> There has been a change in the attendance of teachers such that they are more careful with their service agreement. More active and disciplined. … [In terms of] relationships between teachers and parents, there are many examples of good communication with the community … Teachers have fulfilled their promise to use a variety of teaching and learning methods, and the children feel happier.

These encouraging improvements are evident not only across the case study schools, but also in the quantitative survey data.

As shown in Figure 8, the endline surveys saw a much higher proportion than the baseline level of parents reporting that the quality of their children's education (29.3 percent at baseline compared to 65.8 percent at endline)—which implies far more than teacher attendance—had improved over the previous year (with similar distributions among SAM and SAM+Score parents). Similarly, among SAM+Cam UCs, 64.7 percent agreed that teacher performance had greatly improved since the UC had been established (with 62.7 percent of SAM-only UCs and 64.2 percent of SAM+Score UCs agreeing). If we take these data points at face value, they suggest that there were general improvements in teacher performance under SAM+Cam (as in the other treatment arms) that went beyond the performance area for which information quality improved. This weighs against the possibility that teachers were responding to improvements in information alone. If instead we interpret the similarity in survey response patterns across treatment groups as an indication that UCs and parents had limited capability for discerning meaningful improvements in performance, this would weigh against the possibility that information was a binding constraint in the other treatment groups but not in SAM+Cam.[23] Either way, these survey indicators on improvements in educational quality, together with qualitative field data to the same effect, weaken the case for the information-focused hypothesis.

Figure 8. Parents' perceptions of the quality of their children's education compared to the previous year



[23] It is worth noting that these apparent similarities across treatment arms could be due either to (a) insufficiently granular response categories in the quantitative surveys, or to (b) UCs and parents choosing to favorably inflate scorecard ratings to make their villages and schools look good. These possibilities may weaken the validity of the quantitative survey data on perceived performance improvements for the purposes of weighing alternative explanations. However, the arguments above from the process monitoring data (i.e., scorecard ratings and learning assessments) and the qualitative case study data still stand.

In short, although information certainly played a role in the greater effectiveness of SAM+Cam, and although the quality of teacher attendance information in SAM+Cam was probably better than in the other treatment groups, the data do not provide convincing support for the argument that information, on its own, could sufficiently explain the differential effectiveness of the KIAT Guru treatments.

5.3.4 Alternative hypothesis 2: Not coherence, but clear incentives (information + delegation + motivation)

A second alternative hypothesis is that SAM+Cam was more effective because it created an incentive structure focused on a single, easily fulfilled metric, rather than the more complicated demands of SAM+Score. Teachers responded to this straightforward, formal incentive structure by improving their performance, leading to gains in student learning.[24] In other words, the pivotal design elements were delegation, information, and motivation, as shown in Table 3 above, rather than overall coherence across all five design elements.

Some of the quantitative survey data on the relationship between UCs and teachers may appear, at first glance, to support the incentives-focused hypothesis. When asked whether they feel pressured to give better scorecard ratings to teachers, 20.6 percent of SAM+Score UCs responded affirmatively, compared to just 8.1 percent of SAM+Cam UCs and 8.5 percent of SAM-only UCs. Similarly, 16.4 percent of SAM+Score UCs reported receiving threats not to give bad scores to teachers, in comparison to just 7.4 percent of SAM+Cam UCs. According to the impact evaluation paper, these differences between SAM+Score and the other treatment arms were significant (Gaduh et al, 2020).

The greater pressure faced by UCs in SAM+Score could be interpreted in (at least) two ways. One interpretation is consistent with the incentives-focused alternative hypothesis: teachers pressured UCs about their scores when those scores would have direct effects on their compensation, which was true in SAM+Score but neither in SAM-only, which did not have performance pay, nor in SAM+Cam, where financial incentives depended only on the non-negotiable attendance metric. However, contrary to this interpretation, the camera-based teacher attendance indicator was, in fact, negotiable. The camera-recorded attendance data did not feed mechanistically to scorecard ratings and financial penalties. Rather, SAM+Cam UCs had to verify what was recorded by the cameras and to triangulate the photographic records with formal leave allowances that had been approved by the school leader. This human-led verification process created room for "toleransi" (tolerance or lenience), as the village stakeholders called it. Tolerance could occur either on an individual basis, as reported in the quote above from an Usaba Sepotong teacher who said a UC member said she could leave school early and simply return later to take an end-of-day photo, or on a schoolwide basis, as reported in the Sampuraneh midline report from the field research team who observed that "there was tolerance for tardiness of up to 30 minutes that had been agreed by parties on both sides" and was "not in line with what was written in the service agreement indicators."

Another interpretation of the greater pressure reported by SAM+Score UCs comes from the coherence-focused hypothesis: teachers in both SAM+Cam and SAM+Score schools had incentives to pressure the UC to give them higher scores (whether in attendance or overall), but the cameras lent the SAM+Cam UCs an authority that reduced the likelihood that teachers would question their scorecard ratings. (As for SAM-only UCs, they faced less pressure than their SAM+Score counterparts because their scorecard ratings had social consequences rather than having both social and financial ones.) Compared to the first interpretation, this interpretation is more consistent with the fact that UCs in both performance pay treatments could directly influence teacher salaries. Hence the coherence-focused hypothesis can account for these patterns in UC–teacher relationships more persuasively than the incentives-focused hypothesis.

---

[24] In one sense, this could be taken as a benevolent form of Holmstrom and Milgrom's (1991) multitask problem: in low-resource, low-performance settings, perhaps focusing on a single, straightforwardly quantifiable outcome may be more effective than expecting agents and principals alike to divide their attention to multiple tasks. However, Holmstrom and Milgrom assess the allocation of time and attention across multiple tasks; whereas in this case the focal 'task' is showing up at work to begin with, such that there is enough time to be allocated to job tasks in the first place.

Besides the survey data on pressure and threats to UCs, perhaps an even more challenging set of data points for the incentives-focused hypothesis is that the improvements in teacher attendance were neither large enough nor sufficiently different between treatment groups to convincingly demonstrate that attendance was the main driver of student learning gains. As shown in column (4) of Table 5, gains in teacher attendance under SAM+Cam were non-significant (which was also true across all intervention groups). Moreover, as shown in columns (5), (6), and (7), respectively, two-sample t-tests also show that there was no significant difference between changes in teacher attendance under SAM+Cam and changes in the control group nor under SAM-only and SAM+Score. The only significant between-group difference in baseline–endline teacher attendance was between SAM-only and SAM+Score (p=0.10, not shown in the table).[25] This aligns with the regression analysis in the impact evaluation, which indicates that the main differential effect of KIAT Guru treatments was that SAM+Score had a significant *negative* effect on teacher attendance, which was driven by attendance declines among teachers who were ineligible for the teacher special allowance (Gaduh et al., 2021). Among allowance-eligible teachers, SAM+Cam had a small positive treatment effect on teacher attendance (a gain of 5 percentage points over the control group mean of 84 percent attendance), but this effect was only significant at the 15 percent level—and, unlike the student learning gains under SAM+Cam, the effect on eligible teachers' attendance did not persist in the follow-up study one year after the project facilitators had left (ibid).[26] In short, the salient point here is that SAM+Cam did not significantly improve teacher performance on the incentivized service agreement indicator, i.e., attendance, which weighs against the hypothesis that the effectiveness of SAM+Cam was driven primarily by clear incentives.

Table 5. Average teacher attendance within KIAT Guru treatment and control groups at baseline and endline, with two-sample t-tests for between-group differences (%)

		Mean (standard errors)				Difference between SAM+Cam & … (p-value)		
		Control (1)	SAM-only (2)	SAM+Score (3)	SAM+Cam (4)	Control (5)	SAM-only (6)	SAM +Score (7)
Average teacher attendance at	*baseline*	77.25 (20.08)	76.72 (16.43)	79.22 (21.11)	79.45 (18.61)	2.20 (0.51)	2.73 (0.37)	-0.23 (0.95)
	endline	80.31 (18.51)	83.93 (17.70)	79.66 (16.87)	85.03 (16.09)	4.71 (0.12)	1.09 (0.71)	-5.37* (0.06)
Difference between baseline & endline		3.06 (20.85)	7.21 (21.38)	0.44 (25.50)	5.58 (18.99)	2.51 (0.47)	-1.63 (0.64)	-5.14 (0.19)
N		67	68	67	68	135	136	135

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

[25] The corresponding t-tests for student attendance also failed to find significant gains in student attendance in the treatment groups (from baseline average student attendance ranging from 86.5 percent to 90.4 percent across the treatment groups). Also, changes in student attendance rates between baseline and endline in each treatment group were also not significantly different from changes in the control group. This aligns with the hypothesis that student learning gains in KIAT Guru were not due to increases in time spent in the classroom (by teachers and/or students), but to increases in the overall coherence of accountability relationships that led to changes in practice during time spent in the classroom and at home.

[26] Beyond teacher attendance, the impact evaluation found that none of the treatments resulted in significant gains in the number of hours that teachers allocated weekly for school-related activities. However, teachers in SAM and SAM+Cam changed how they allocated these hours, such that there was a significant increase in time allocated to activities that were positively correlated with student learning (Gaduh et al., 2021, Table 8). This reinforces the argument that SAM+Cam (and KIAT Guru more generally) improved teacher performance not through the amount of time that teachers supplied, but in how they used that time. This weighs against both the information-focused alternative hypothesis (because the informational difference in SAM+Cam focused on teacher attendance) and the incentives-focused alternative hypothesis (because the incentives in SAM+Cam similarly focused on attendance).

5.3.5 Summary

On balance, the coherence-focused hypothesis can explain more of the field observations than either the information-focused hypothesis or the incentives-focused hypothesis. This is not to imply that *information* and *motivation* were unimportant, nor that the alignment between *delegation*, *motivation*, and *information* in the formal incentive structures was unimportant. However, these elements in isolation cannot sufficiently explain the observed patterns without broadening the explanation to include the social nature of the accountability relationship—which calls attention to village-level status hierarchies that require *support* in order to sufficiently empower UCs to enforce the more demanding delegation of KIAT Guru, as well as to the value of reciprocity in the accountability relationship, as demonstrated in greater *finance* and other resources that village stakeholders contributed to complement the greater efforts demanded of teachers. Emphasizing the coherence between elements in this fundamentally social relationship of accountability may not be the most straightforward explanation, but it has the most explanatory power.

6 Discussion

Our analysis of the KIAT Guru project for accountability in education in remote Indonesian villages aligns with arguments for the importance of coherence between stakeholders in service delivery for the poor (World Bank, 2003; Pritchett, 2015). We find that one key element of coherence in the intervention context, where most local community members were relatively disempowered compared to their frontline service providers, is support that comes from tools, processes, and structures that strengthen the accountability relationship, especially by empowering the weaker set of actors. Where communities have few resources, local-level accountability is most effective when communities are empowered with both the technical tools and legitimacy to hold teachers and school leaders accountable for education service delivery. In the SAM+Cam treatment, using dedicated smartphone cameras to monitor teacher attendance improved both the technical quality and the symbolic authority of user committees' monitoring of teachers.

Our analysis also aligns with arguments for combining community-led voice with state-led teeth in social accountability (Fox, 2015; see also Kosec & Wantchekon, 2020). KIAT Guru improved the coherence between the voice & choice relationship (between communities and frontline providers) and the management relationship (between education authorities and frontline providers) by involving school leaders and a government-funded teacher allowance in a community-monitored performance pay scheme. By combining community-led voice with state-led teeth in a way that is coherent with the configuration of principals and agents and other aspects of the implementation context, KIAT Guru bucked the trend of school-based management committees in many parts of the Global South that present a formal and politically appealing image of educational decentralization but fail to respond to local needs and preferences (e.g., Bano, forthcoming).

If we had not looked at village-level power dynamics and the importance of support, one (erroneous) conclusion might have been that SAM+Cam was the most effective treatment for raising student learning because teaching in these low-resourced remote schools is a straightforward task where the most important thing is the teacher showing up; such that potential side effects from skewed allocations of attention toward the incentivized behaviors (Holmstrom & Milgrom, 1991; Murnane & Cohen, 1986) were irrelevant. However, taking power dynamics and social status hierarchies into account makes it clear that SAM+Cam was the most effective treatment precisely because teaching was regarded as a complex task. Consequently, the UCs were regarded as lacking the legitimacy to evaluate teachers, especially when this evaluation affects teacher pay. To use the terminology of Honig and Pritchett (2019), SAM+Cam struck the right balance between accounting-based accountability that linked official financial incentives to standardized, verifiable camera-monitored attendance data, and account-based accountability that linked social rewards and penalties to multifaceted community scorecards that were part of a deliberative, interactive social accountability process. Crucially, these two aspects of the voice & choice accountability

relationship were mutually reinforcing, i.e., coherent.[27] That said, it is worth noting that in a different context where parents and community members were at least as well-educated as their children's teachers, SAM+Score could hypothetically be the more coherent intervention. In such a context, the community may have the technical capacity and social legitimacy to evaluate scorecards tailored to improving the complex task of teaching, and hold teachers accountable through financial incentives based on the scorecards.

Methodologically, our analysis highlights the value of combining different types of data and levels of analysis in explaining the mechanisms underlying an intervention. Our argument about the importance of support and of local power dynamics would have been far weaker without the qualitative data from interviews and focus groups in the nine case study schools. Yet the qualitative data from this small sample of schools, however richly detailed, would not have been as persuasive if it had not been complemented by the quantitative process monitoring data, survey data, and service agreement indicator data from across all 203 treatment schools. It is also worth noting that different types and levels of data were suitable for different aspects of the analysis, e.g., granular interview and focus group data to understand how teachers perceived the intervention, and aggregated data from unannounced visits to schools to gauge whether the intervention affected teacher attendance rates.

7 Conclusion

Frontline service delivery failures, particularly by those serving poor communities and/or in remote locations, remain frustratingly common throughout low-income countries and some middle-income countries. The case of KIAT Guru shows that strengthening the capacity and authority of parents and the broader community could be one pathway for improving service delivery outcomes and holding service providers more accountable. However, this requires strong policy support, particularly in a context where power relations between the two parties are not balanced.

It is worth restating that all three KIAT Guru treatments improved learning outcomes. In other words, independent of whether teachers faced financial disincentives, the social accountability mechanism seemed to work. However, the empowerment of poor remote communities needs multiple reinforcements: regulations and resources (from government), on-site guidance and endorsement (from project facilitators), knowledge of which information and data are relevant (from capacity building), and willingness to contribute time (from the community themselves).

While successfully improving learning outcomes, KIAT Guru project implementation was too complex an intervention to be scaled up. Given that SAM is a package of interventions (even apart from the different performance pay mechanisms), it would be useful, from a cost-effectiveness perspective, for future studies to identify which elements matter most; whether information on benchmarked learning outcomes, joint service agreements, community monitoring, or monthly meetings.

To that end, between 2019 and 2020, the same project partners expanded a simplified version of the project to expand the scope to 410 schools. This Phase 2 project streamlined the SAM and digitized some of the administration process (including learning outcomes and service agreements). It also varied the configuration of involved stakeholders to test the influence of different accountability relationships: one more external to the school, as in the Phase 1 treatments discussed in this paper, and one more internal to the school, where the pre-existing school committee took on the role of the user committee. Although some of these Phase 2 iterations seemed to improve program effectiveness, the impact evaluation could not be completed due to the pandemic (World Bank, 2021).

Overall, the three treatments discussed in this paper differ in their potential for sustained collaboration and impact. This study validates findings from the previous KIAT Guru studies, suggesting that SAM-

[27] For an example of complementary interventions that improved the overall coherence of the management relationship in a given context and resulted in student learning gains, Mbiti et al. (2019) find that providing both financial grants and test-based teacher financial incentives to schools in Tanzania improved student learning far more than providing either the grants or the incentives in isolation.

only and SAM+Score were not as effective as SAM+Cam because they were less coherent. SAM+Score generated the most pushback from teachers who deemed community members to lack the capacity and legitimacy to evaluate them. As for SAM-only, the follow-up KIAT Guru impact evaluation found that student learning gains were not sustained after the project facilitators stopped supporting the project in each village (Gaduh et al, 2021). However, SAM+Cam appears to be coherent for improvements in student learning in these implementation contexts because teachers deemed the evaluation to be fair, while community members found that the cameras boosted their limited capacity.

8 References

ACDP (Analytical and Capacity Development Partnership). (2014). Study on Teacher Absenteeism in Indonesia 2014. Jakarta, Indonesia: Education Sector Analytical and Capacity Development Partnership.

ASER. (2014). Annual status of education report (rural) 2013. New Delhi, India: ASER Centre

Andrabi, T., and Brown, C. (forthcoming). Subjective versus Objective Incentives and Employee Productivity. RISE Working Paper Series.

Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets. *The American Economic Review*, *107*(6), 1535–1563.

Badan Pusat Statistik (BPS). (2017). Portret Pendidikan Indonesia. Statistik Pendidikan 2017. Jakarta: Badan Pusat Statistik.

Banerjee, A. V., & Duflo, E. (2008). Mandated Empowerment: Handing Antipoverty Policy Back to the Poor?. *Annals of the New York Academy of Sciences*, *1136*(1), 333–341. https://doi.org/10.1196/annals.1425.019

Banerjee, A. V., Duflo, E., & Glennerster, R. (2008). Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System. *Journal of the European Economic Association*, *6*(2–3), 487–500. https://doi.org/10.1162/JEEA.2008.6.2-3.487

Bano, M. (forthcoming). International Push for SBMCs and the Problem of Isomorphic Mimicry: Evidence from Nigeria. RISE Working Paper Series.

Barrera-Osorio, F., Gertler, P., Nakajima, N., & Patrinos, H. A. (2021). *Promoting Parental Involvement in Schools: Evidence from Two Randomized Experiments*. Research on Improving Systems of Education (RISE). https://doi.org/10.35489/BSG-RISE-WP_2021/060

Beatty, A., Berkhout, E., Bima, L., Coen, T., Pradhan, M., & Suryadarma, D. (2018). *Indonesia Got Schooled: 15 Years of Rising Enrolment and Flat Learning Profiles*. Research on Improving Systems of Education (RISE). https://doi.org/10.35489/BSG-RISE-WP_2018/026

Bjorkman, M. & Svensson, J. (2009). When is community-based monitoring effective? Evidence from a randomized experiment in primary health in Uganda. *Journal of the European Economic Association*, *8*(2-3), 571–581. https://doi.org/10.1111/j.1542-4774.2010.tb00527.x

Broekman, A. (2015). The Effects of Accountability: A Case Study from Indonesia. In J. Evers & R. Kneyber (Eds.), *Flip the System: Changing Education from the Ground Up*. Taylor & Francis Group. https://doi.org/10.4324/9781315678573-7

Cilliers, J., Kasirye, I., Leaver, C., Serneels, P., & Zeitlin, A. (2018). Pay for locally monitored performance? A welfare analysis for teacher attendance in Ugandan primary schools. *Journal of Public Economics*, *167*, 69–90. https://doi.org/10.1016/j.jpubeco.2018.04.010

de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. (2018). Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia. *The Quarterly Journal of Economics*, *133*(2), 993–1039. https://doi.org/10.1093/qje/qjx040

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, *102*(4), 1241–1278. https://doi.org/10.1257/aer.102.4.1241

Gaduh, A., Pradhan, M., Priebe, J., & Susanti, D. (2020). Scores, Camera, Action? Incentivizing Teachers in Remote Areas. RISE Working Paper Series. 20/035. https://doi.org/10.35489/BSG-RISE-WP_2020/035

Gaduh, A. B., Pradhan, M. P., Priebe, J., & Susanti, D. (2021). Scores, Camera, Action: Social Accountability and Teacher Incentives in Remote Areas. In *Policy Research Working Paper Series* (No. 9748; Policy Research Working Paper Series). World Bank. https://openknowledge.worldbank.org/handle/10986/36112

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

Honig, D., & Pritchett, L. (2019). The Limits of Accounting-Based Accountability in Education (and Far Beyond): Why More Accounting Will Rarely Solve Accountability Problems. RISE Working Paper Series, 19/030. https://doi.org/10.35489/BSG-RISE-WP_2019/030

Huang, A. R., Revina, S., Fillaili, R., & Akhmadi. (2020). *The Struggle to Recruit Good Teachers in Indonesia: Institutional and Social Dysfunctions*. RISE Working Paper Series. 20/041. https://doi.org/10.35489/BSG-RISE-WP_2020/041

Islam, A. (2019). Parent–teacher meetings and student outcomes: Evidence from a developing country. *European Economic Review*, *111*, 273–304. https://doi.org/10.1016/j.euroecorev.2018.09.008

Kosec, K., & Wantchekon, L. (2020). Can information improve rural governance and service delivery? *World Development*, *125*, 104376. https://doi.org/10.1016/j.worlddev.2018.07.017

Kurniasih, H., Utari, V., and Akhmadi. (2018). Character Education Policy and Its Implications for Learning in Indonesia's Education System. RISE Insight. https://doi.org/10.35489/BSG-RISE-RI_2018/007

Lumbanraja, S.K., & Prameswari, I.A. (2021). Diagnostic Test to Increase Community Participation in Improving Learning Outcomes in Indonesia's Remote Primary Schools : Quick Test (English). World Bank Group, Brief. http://documents.worldbank.org/curated/en/977701617083603131/

Lumbanraja, S.K., Prameswari, I.A., & Susanti, D. (2021). Community Participation in Measuring Learning Outcomes in Remote Areas of Indonesia: Results from the Development and Implementation of *Tes Cepat* Teacher Performance and Accountability (KIAT Guru). World Bank, Jakarta, Background Paper. https://openknowledge.worldbank.org/handle/10986/35466

Marliyanti, Adelina, U.R., & Susanti, D. (2022). From Facilitation to Participation: Community Empowerment to Improve Education in Remote Areas. Manuscript in preparation.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. *The Quarterly Journal of Economics*, *134*(3), 1627–1673. https://doi.org/10.1093/qje/qjz010

Murnane, R., & Cohen, D. (1986). Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive. *Harvard Educational Review*, *56*(1), 1–18. https://doi.org/10.17763/haer.56.1.l8q2334243271116

Narwana, K. (2015). A global approach to school education and local reality: A case study of community participation in Haryana, India. *Policy Futures in Education*, *13*(2), 219–233. https://doi.org/10.1177/1478210314568242

PAL Network. (2018). PAL Network 2018 Annual Plan: Assessment for Action. Retrieved from https://palnetwork.org/wp-content/uploads/2019/03/2018_PAL_Annual-Plan-Budget_Final.pdf

Pawson, R. (2013). *The Science of Evaluation: A Realist Manifesto*. SAGE Publications. https://doi.org/10.4135/9781473913820

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. SAGE Publications.

Porpora, D. V. (2015). *Reconstructing Sociology: The Critical Realist Approach*. Cambridge University Press.

Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., & Artha, R. P. (2014). Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia. *American Economic Journal: Applied Economics*, *6*(2), 105–126. https://doi.org/10.1257/app.6.2.105

Pritchett, L. (2013). *The rebirth of education: Schooling ain't learning*. Center for Global Development.

Pritchett, L. (2015). Creating Education Systems Coherent for Learning Outcomes. RISE Working Paper Series. 15/005. https://doi.org/10.35489/BSG-RISE-WP_2015/005

Spivack, M. (2021). Applying Systems Thinking to Education: The RISE Systems Framework. RISE Insight Note 2021/028. https://doi.org/10.35489/BSG-RISE-RI_2021/028

Stern, J., & Nordstrum, L. (2014). *Indonesia 2014: The National Early Grade Reading Assessment (EGRA) and Snapshot of School Management Effectiveness (SSME) Survey: Report of Findings*. USAID, EdData II.

UNICEF. (2012). 'We Like Being Taught': A Study on Teacher Absenteeism in Papua and West Papua. https://eric.ed.gov/?id=ED566745

World Bank. (2003). *World Development Report 2004: Making Services Work for Poor People*. The World Bank. https://doi.org/10.1596/0-8213-5468-X

World Bank. (2019). *Primary Education in Remote Indonesia: Survey Results from West Kalimantan and East Nusa Tenggara*. World Bank. https://doi.org/10.1596/33113

World Bank. (2020). Community Participation and Teacher Accountability: Improving Learning Outcomes in Remote Areas of Indonesia. https://openknowledge.worldbank.org/handle/10986/33807

World Bank. (2021). Implementation Completion and Results Report (ICR) Document – Indonesia: Improving Teacher Performance and Accountability (KIAT Guru) Phase 2 - P167216. http://documents.worldbank.org/curated/en/897531625000304453/

9 Appendices

Appendix A Descriptive quantitative analysis of the KIAT Guru surveys and process monitoring

The descriptive analysis below is formulated to quantitatively complement this reanalysis paper as well as the findings of the original qualitative study (World Bank, 2020). The original qualitative study assesses the impacts of KIAT Guru treatments on several categories including (1) teacher presence, (2) teacher performance and school leadership, (3) parent participation, (4) student learning, attitudes, and discipline, (5) village government activities, (6) user committee (UC) effectiveness, and (7) village cadre performance.

In this appendix, we provide descriptive analysis findings for all categories except for (1) and (4), for which the impact evaluation reports provide robust analysis (Gaduh et al., 2020, 2021). A selection of variables provided from various questionnaires were employed to gauge any changes within the respective categories between baseline and endline surveys. In general, the changes we aim to focus on are those related to the coherence between the actions of various stakeholders and the objective of achieving better learning outcomes. Findings are presented in two categories: (a) the overall treatment level, which covers 203 treatment schools; and (b) case study schools, which covers 9 treatment schools that were included in the qualitative study.

Codes in parentheses (e.g., LED05, DPA13) refer to indicator names, as used in the KIAT Guru quantitative survey and process monitoring datasets.

1. Teacher Performance

- *Treatment Level:* Across all treatment groups, at least 99.5% of user committees in 202 schools[28] saw improvements in teachers' performance since they had become active (LED05). In comparison to the baseline, there was an increase of proportion of parents who viewed their children's education to be at least "good" during the endline survey (DPA13; from 86% at baseline to 95% at endline), and higher proportion of parents who viewed the quality of their children's education in better light compared to the previous year (DPA14; from 27.5% to 64.2%).
- *9 Case Study Schools:* Findings corroborated the overall results with very few exceptions. Only two out of the nine case study schools saw a decreased endline proportion of parents who perceived their children's education to be at least good. Meanwhile, UCs in all nine schools agreed that there had been improvement in teachers' performance since the UC had become active. The proportion of parents who thought the quality of education had been better than during the previous year showed consistent increase.

2. School Leadership

- *Treatment Level:* With exception of SAM+Score school leaders, there was an increase in the proportion of school leaders who performed evaluations other than the routine SKP[29] (BEV17; from 69% to 76% in SAM-only schools, and from 74% to 81% in SAM+Cam schools). This was corroborated by teachers, who reported an increase from 79% to 84% in principals conducting non-SKP performance evaluations (CME01). There was also an increase in percentage of school leaders who gave rewards to well-performing teachers (BEV22; from 34% to 45% in SAM-only schools, and 42% to 55% in SAM+Cam schools), with exception of SAM+Score schools

[28] During the 2018 midline survey, enumerators failed to meet any representatives from the user committee in one of the SAM-only schools during the field visit, hence data are available for UCs in 202 out of the 203 schools.
[29] Sasaran Kinerja Pegawai, or Officer Performance Targets, i.e., a mandatory annual performance evaluation for all civil service officers.

(decreasing from 47% to 43%). Out of 202 UCs across all treatment groups, 95.6% of them attested to improvements in the principal's performance since the UC became active.

- *9 Case Study Schools:* Eight of the case study UCs reported an improvement in the school leader's performance since they had become active. At baseline, school leaders performed evaluations other than the routine SKP in only seven schools. However, at endline we found that all school leaders performed teacher evaluations other than the routine one. For schools where school leaders had already conducted non-SKP teacher evaluations, we saw an increase in the number of evaluation criteria. We did not, however, find more school leaders giving rewards to well-performing teachers in these schools.

3. Local Government Support

- *Treatment Level:* Of 202 UCs, 79.2% of them perceived village government performance to have improved since they had become active (LED14). These numbers were slightly lower than the proportion of UCs who thought there had been improvements from the teachers' side (95%). Interestingly, SAM+Score villages had the highest proportion of UCs attesting to village governments taking novel initiatives to honor scorecard indicators (LSR06; 48% out of 68 schools). On the other hand, 93% of 57 SAM+Cam village heads were found to have enthusiasm in attending KIAT Guru meetings, compared to 83% out of 58 SAM-only villages and 90% out of 59 SAM+Score villages (APDD05d). Similar enthusiasm could also be found in all schools' village governments; on average 74.7% of village heads across treatment groups verify that other village government officials had also attended KIAT Guru meetings (APDD05g). With regard to support from school supervisors and district education agencies, between baseline and endline surveys we saw an increase in the proportion of SAM+Cam schools that have been visited by school supervisors from 86.76% to 89.71%, whereas schools in two other treatment groups saw a decrease in the proportion (from 87.76% to 77.94% in SAM-only schools and from 88.06% to 82.09% in SAM+Score schools).
- *9 Case Study Schools:* With the exception of one school, UCs in case study schools saw improvements in village government officials' performance since they had become active. While almost all schools' village governments were involved in KIAT Guru activities such as meetings and proposing education initiatives, SD Usaba Sepotong (SAM+Cam school) did not experience these beneficial changes.

4. Parent Participation

- *Treatment Level:* The introduction of KIAT Guru shifted the belief that school was the only party responsible for children's education. In all schools, we saw a notable increase in the proportion of parents who thought communities were also responsible for children's education (DEO; from 70.5% to 80.0%), with parents in SAM-only schools contributing to the increase. Concurrently, 95% of 202 UCs reported improvements in parents' participation since they had become active (LED08). However, such perceptions were not mirrored by teachers, as the change in proportion of teachers who thought parents had good level of involvement in children's education hovered at 43% (CKP15).

  Yet the data also show minimal behavior change from parents. The proportion of parent respondents who had ever helped with their children's homework decreased by 0.2 percentage points in SAM-only schools and 0.8 percentage points in SAM+Score schools between baseline and endline, while the proportion increased by 3.3 percentage points in SAM+Cam schools (DPR02). Additionally, the proportion of parents and/or guardians who accompanied their children in learning at home within the past one week decreased by 3.28 percentage point in SAM-only schools, 6.18% in SAM+Cam schools, and 3.88% in SAM+Score schools (DPR04). There

also seems to have been a decrease in the average number of days on which parents accompanied children's learning at home per week, from 3.4 days during baseline to 3.2 days during endline (DPR05). However, parents compensated for this reduction by increasing the average daily allocated time to accompany children learning at home, from 31.8 minutes during baseline to 34.7 minutes during endline (DPR06). Additionally, parents also amped up the investment in their children's education by increasing the average education expenditure by IDR 34,269 (11%) between baseline and endline (DPS).

- *9 Case Study Schools:* In four schools, the proportion of parents who believed communities were also responsible for children's education grew between baseline and endline. In the case study schools, UCs saw improvement in parent participation, with sole exception to SDI Konang (a SAM+Score school). This perception is corroborated by the higher proportion of teachers who thought parents had good level of involvement in children's education, with the exception of two schools. Parents in SAM+Cam case study schools exhibited greater positive changes in their level of involvement, with consistently higher proportions of parents who helped with their children's homework, increased their education expenditure, and increased the number of days per week allocated to accompany their children's home learning. However, parents in other treatment groups compensated for it with higher allocated time per day to accompany children learning at home.

5. UC effectiveness

- *Treatment Level:* There is a negligible difference across treatment groups in terms of the efforts that UCs exerted as measured by the average number of members involved in monthly monitoring and monthly evaluation meetings (LMS02; LMM01). According to principals, UCs in SAM-only and SAM+Cam schools exhibited a slightly greater effort in providing suggestions regarding the learning process at schools (BUC05; at 85% and 84% respectively; compared to SAM+Score at 81%). Correspondingly, parents in SAM+Score schools had the lowest awareness of the UC's existence in their village (DUC02; at 50% compared to 59% for SAM-only and 57% for SAM+Cam). Parents also thought that their cooperation with teachers had improved since UCs became active (DUC25). Interestingly, SAM+Score Ucs had the highest average perception scores in self-assessing whether they had the capability to evaluate teachers without feeling intimidated (LMM08; 8.79 compared to 8.58 for SAM-only and 8.66 for SAM+Cam).
- *9 Case Study Schools:* In seven of the case study schools, the average number of UC members involved in monthly monitoring exceeded that of the average in 202 schools. All case study Ucs also exhibited great confidence in their capability to evaluate teachers without feeling intimidated. With the exception of two schools, parents in case study school samples showed good levels of awareness regarding the UC's work in their village. With the exception of SD Simpang Dua (SAM+Score school), at least 80% of parents thought UCs had improved parent-teacher cooperation in their respective schools.

6. Parent–Teacher Dynamics

- *Treatment Level:* Contrary to UCs' self-assessments in the previous point, the highest proportion of UCs that felt intimidated in discussing score results with teachers is exhibited in SAM+Score schools (LMS12; 10% compared to 3% for SAM-only and 5% for SAM+Cam). Additionally, SAM+Score schools had the highest proportion of UCs who felt pressured to give better scores to teachers (LMS16; 21%, compared to 8.5% in SAM-only and 8% in SAM+Cam). In schools with pay-for-performance mechanisms, SAM+Score schools had higher proportion (16%) of UCs who received threats from teachers, in comparison to SAM+Cam schools (7%). Interestingly, the lowest average school leader perception score on the UC's ability to choose important education

indicators is exhibited in SAM+Cam group (7.21), in contrast to SAM+Score (7.39) and SAM-only (7.53) groups.

- *9 Case Study Schools:* We found no evidence of pressure or threats from schools to UCs based on the quantitative data for the nine case study schools. Fascinatingly, school leaders in SAM+Score case study schools had the most favorable views UCs ability in choosing important indicators compared to schools in other treatment groups, in contrast the overall trend.

7. UC Effectiveness and Cadre Performance (from process monitoring data)

- *Treatment Level:* Across all treatment groups, more than 60% of village cadres have at least a high school education. This may be attributed to the nature of the role which requires willingness to learn about technology. As of March 2018, the proportion of cadres with at least a high school education is the lowest in SAM+Score schools (61.2%), compared to SAM-only (73.5%) and SAM+Cam (75.0%) schools. On the other hand, the proportion of female cadres, 26% out of 203 cadres, indicates that the role is still male-dominated. SAM-only schools had the lowest representation of female cadres (23.5%), although the difference with SAM+Cam (26.5%) and SAM+Score (26.9%) is quite small, magnitude-wise. Overall, cadres show adequate capability in delivering their tasks with at least 80% of them regularly attending and speaking up in meetings, as well as having facilitation skills, which is their main responsibility as a cadre. In terms of their performance, village facilitators (FM) consistently graded SAM+Cam cadres' performance lower than their SAM-only and SAM+Score counterpart, as suggested in the table below.

Table A1. Profile of KIAT Guru cadres, by treatment group

Treatment	% with at least High School Education	% Female	% Regularly Attends Meetings	% Speaks up in Meetings	% Has facilitation skills
SAM-only	73.5% (68)	23.5% (68)	93.9% (66)	87.9% (66)	90.9% (66)
SAM+Cam	75.0% (68)	26.5% (68)	92.7% (68)	80.6 % (67)	84.9% (66)
SAM+Score	61.2% (67)	26.9% (67)	97.0% (65)	87.9% (66)	89.4% (66)

In comparison to the cadres, the proportion of user committee members with at least high school education is lower across all treatment groups. However, the trends of the between-treatment comparison are reversed, as SAM+Score UCs were found to have the highest proportion of members with at least a high school education (32.89%) compared to SAM-only (24.59%) and SAM+Cam (29.84%). Moreover, the structure of the UC indicates more equal gender representation, with females representing 44.30%, 47.14%, and 43.74% of the total members in SAM, SAM+Cam, and SAM+Score UCs respectively. While almost 80% of UC members regularly attend meetings, only half of them were brave enough to speak up at these meetings. Nearly 20% of them had facilitation skills.

Table A2. Profile of KIAT Guru User Committees, by treatment group

Treatment	% with at least High School Education	% Female	% Regularly Attends Meetings	% Speaks up in Meetings	% Has facilitation skills
SAM	24.59% (614)	44.30% (614)	80.17% (605)	55.54% (605)	23.14% (605)
SAM+Cam	29.84% (630)	47.14% (630)	79.52% (625)	55.36% (616)	17.45% (619)
SAM+Score	32.89% (599)	43.74% (599)	81.93% (592)	51.73% (578)	21.48% (582)

- *9 Case Study Schools:* Eight cadres had at least a high school education and only two of them were female. Due to incomplete data, we do not have information regarding the capacity of SDI Sangka's cadre. All cadres attended meetings regularly. However, two of them rarely spoke up in meetings and one of these two had inadequate facilitation skills.

  The education profile of UC members in the nine case study schools varies from none of the members having a high school education in SDI Konang to almost all members being high-school educated in SDN 10 Engkangin. The level of female representativeness also varies, with six of them achieving more than 50% representation and three of them having less than the average rate of female representation. Among the nine case study schools, UCs in the SAM+Score treatment group had lower willingness to regularly attend meetings. However, when it comes to speaking up in meetings, UCs in SAM+Cam case study schools performed worse than other treatment groups. Intuitively, the provision of KIAT Kamera may have dampened the need to voice out opinions in the meetings.

Appendix B Amendments to the teacher and school leader service agreement indicators

Table B1. Service agreement indicators for teachers in SDK Kondok, pre- and post-amendment

Indicator	**Pre-amendment weight**	**Post-amendment weight**
Indicators directly related to student learning		
Teacher gives homework to students, and makes sure parents are aware of and sign off on students' completed homework	10	10
Teacher creates problem sets for student learning groups to work on, and provides instructions to the problem sets during classroom learning	10	–
Teacher uses varieties of teaching methods, including storytelling, singing, role playing, and question-answer with students, as well as teaching aids	10	–
Teacher supervises student learning groups by conducting regular visits to all groups at least once a month	10	10
Grade 1-3 teachers ask students to rehearse reading letters and numbers daily before classroom lessons begin	–	10
Teachers provides remedial assistance for students who are behind by providing additional lesson 15 minutes before class begins.	–	20
Grade 1-3 teachers use letter and number cards as teaching aids for students to be able to read and count	–	10
Teacher supervises student learning groups by conducting regular visits to all groups at least once a month	–	10
Subtotal: directly linked to learning	*40*	*70*
Indicators not directly related to student learning		
Teacher arrives and leaves school on time. Monday - Thursday: 7.30am-12.20pm. Friday - Saturday: 7.30am-10.55am	20	20
Teacher disciplines students gently with positive discipline. Teachers are not to use harsh words and/or physical punishment when disciplining students.	20	10
Teacher motivates students using positive encouragement and advice.	10	–
Teacher informs parents of students who are facing challenges in school by conducting a visit to their home and writing a formal letter to the parents	10	–
Subtotal: indirectly linked to student learning	*60*	*30*
Total weights	100	100

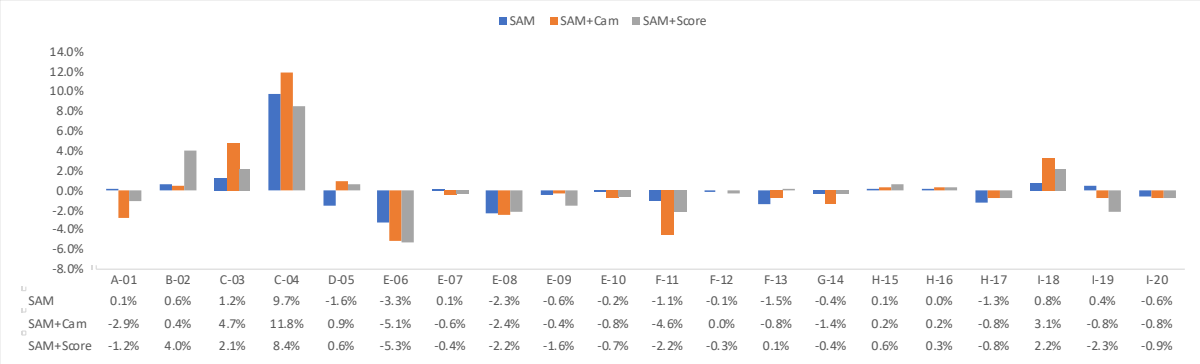
Table B2. Average weighting of teacher service agreement indicator categories, pre- and post-amendment

Legend: *Top 3 categories pre-amendment* | *Top 3 categories post-amendment*

		SAM		SAM+Score		SAM+Cam		Overall	
Code	**Description of indicators**	Pre-	Post-	Pre-	Post-	Pre-	Post-	Pre-	Post-
A-01	Teacher applies fun and motivating learning techniques in classroom.	4.92	4.98	7.43	6.25	7.27	4.41	6.55	5.21
B-02	Teacher develops and applies lesson plan; assists students during classroom learning.	4.47	5.03	2.22	6.24	2.76	3.11	3.14	4.80
C-03	Teacher uses teaching aids.	4.35	5.51	3.06	5.15	3.15	7.85	3.51	6.16
C-04	Teacher works to improve students' literacy and numeracy skills.	2.80	12.50	2.18	10.61	2.32	14.11	2.43	12.40
D-05	Teacher strives to ensure students' learning comprehension, including in providing feedback.	14.59	13.04	11.56	12.18	11.53	12.47	12.55	12.56
E-06	Teacher uses positive discipline with students, and avoids any form of verbal or corporal punishments.	14.13	10.83	14.58	9.25	14.20	9.14	14.30	9.75
E-07	Teacher offers motivation, appreciation and rewarding praises to students.	1.28	1.37	1.75	1.33	1.56	1.00	1.53	1.23
E-08	Teacher instils religious, cultural, and social norms in students.	3.70	1.39	4.61	2.42	4.79	2.35	4.37	2.05
E-09	Teacher inculcates patriotism and values of obedience and orderliness in students.	1.87	1.31	4.01	2.40	3.09	2.69	3.00	2.13
E-10	Teacher does not ask students to work for teacher's personal needs.	0.57	0.38	0.87	0.20	1.20	0.36	0.88	0.31
F-11	Teacher starts and ends class on time.	30.39	29.30	30.42	28.22	31.63	27.05	30.82	28.20
F-12	Teacher requests for permission and provides a legitimate reason and evidence for any absences.	0.08	0.00	0.40	0.12	0.00	0.00	0.16	0.04
F-13	When absent, teacher ensures s/he is replaced by a substitute and complies to administration procedure of taking leave.	2.03	0.54	0.52	0.60	1.60	0.85	1.38	0.66
G-14	Teacher sets good example through their behavior whilst in school.	0.77	0.38	1.51	1.11	1.92	0.56	1.40	0.68
H-15	Teacher promotes the use of Bahasa Indonesia as medium of communication in school.	1.14	1.28	1.27	1.82	1.80	2.02	1.40	1.71
H-16	Teacher informs parents of student learning progress.	1.75	1.75	1.87	2.15	1.76	1.95	1.79	1.95
H-17	Teacher holds and attends meeting with parents and community members; teacher communicates with User Committee.	1.63	0.34	1.19	0.38	1.12	0.27	1.31	0.33
I-18	Teacher provides remedial sessions to improve students' learning comprehension.	6.09	6.85	5.24	7.41	4.47	7.62	5.27	7.29
I-19	Teacher provides assistance to students during physical education, local content, scouting and other extracurricular activities.	1.71	2.12	3.73	1.47	2.52	1.67	2.66	1.76
I-20	Teacher educates students to promote clean and orderly school environment.	1.75	1.11	1.59	0.70	1.32	0.51	1.55	0.78
Total weighting for indicators directly related to learning		**40.1**	**50.9**	**35.1**	**34.8**	**51.8**	**35.1**	**53.5**	**52.1**

Note. The 8 (out of 20) indicator categories directly related to learning are A-01, B-02, C-03, C-04, D-05, H-15, H-16, and I-18, as shown in blue text.

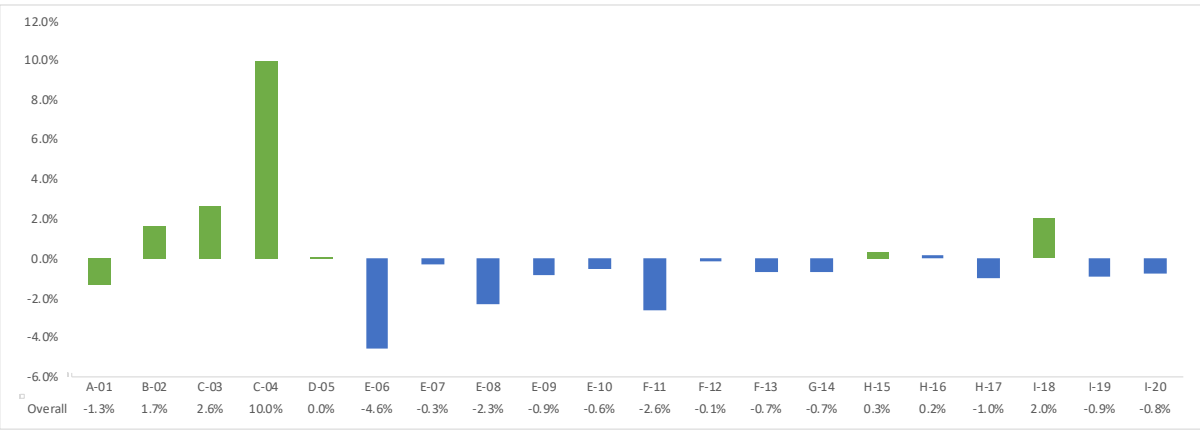
Figure B1. Changes in teacher service agreement indicators, by treatment group



	A-01	B-02	C-03	C-04	D-05	E-06	E-07	E-08	E-09	E-10	F-11	F-12	F-13	G-14	H-15	H-16	H-17	I-18	I-19	I-20
SAM	0.1%	0.6%	1.2%	9.7%	-1.6%	-3.3%	0.1%	-2.3%	-0.6%	-0.2%	-1.1%	-0.1%	-1.5%	-0.4%	0.1%	0.0%	-1.3%	0.8%	0.4%	-0.6%
SAM+Cam	-2.9%	0.4%	4.7%	11.8%	0.9%	-5.1%	-0.6%	-2.4%	-0.4%	-0.8%	-4.6%	0.0%	-0.8%	-1.4%	0.2%	0.2%	-0.8%	3.1%	-0.8%	-0.8%
SAM+Score	-1.2%	4.0%	2.1%	8.4%	0.6%	-5.3%	-0.4%	-2.2%	-1.6%	-0.7%	-2.2%	-0.3%	0.1%	-0.4%	0.6%	0.3%	-0.8%	2.2%	-2.3%	-0.9%

Notes. All values are within-treatment-group average changes, in percentage points. See Table B2 for full descriptions of each indicator category.

Figure B2. Changes in teacher service agreement indicators, average across treatment groups



	A-01	B-02	C-03	C-04	D-05	E-06	E-07	E-08	E-09	E-10	F-11	F-12	F-13	G-14	H-15	H-16	H-17	I-18	I-19	I-20
Overall	-1.3%	1.7%	2.6%	10.0%	0.0%	-4.6%	-0.3%	-2.3%	-0.9%	-0.6%	-2.6%	-0.1%	-0.7%	-0.7%	0.3%	0.2%	-1.0%	2.0%	-0.9%	-0.8%

Notes. All values are overall average changes, in percentage points. See Table B21 for full descriptions of each indicator category.

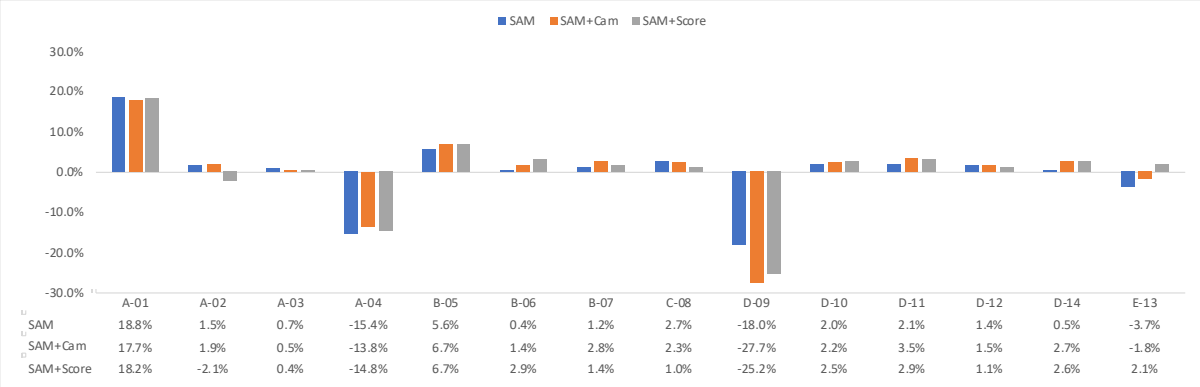
Table B3. Average weighting of school leader service agreement indicator categories, pre- and post-amendment

Legend: | *Top 3 categories pre-amendment* | *Top 3 categories post-amendment*

		SAM		SAM+Score		SAM+Cam		Overall	
Code	**Description of indicators**	Pre-	Post-	Pre-	Post-	Pre-	Post-	Pre-	Post-
A-01	Principal sets a good example through their attendance and behavior whilst in school.	11.00	29.76	9.92	28.14	10.18	27.87	10.36	28.59
A-02	Principal uses positive discipline with students.	6.40	7.95	7.78	5.72	5.62	7.51	6.61	7.07
A-03	Principal offers motivational, appreciation and rewarding praises to students and teachers.	0.16	0.87	0.56	0.95	0.16	0.71	0.30	0.84
A-04	Principal tends to the needs of the community, attends meeting with parents and communicates with school committee and user committee.	17.57	2.20	16.27	1.51	14.71	0.94	16.18	1.55
B-05	Principal develops curriculum and ensures that teachers develop lesson plans and deliver fun learning activities.	2.30	7.87	2.70	9.38	1.62	8.31	2.21	8.52
B-06	Principal communicates as well as ensures other teachers to communicate student learning progress to parents.	0.99	1.42	1.19	4.13	1.13	2.52	1.10	2.68
B-07	Principal instills patriotism and values of compliance, cleanliness, and orderliness in students.	4.02	5.20	4.60	6.04	4.12	6.89	4.25	6.04
C-08	Principal develops and reports school's work and budget plan.	0.16	2.91	0.16	1.19	0.48	2.75	0.27	2.29
D-09	Principal optimally manages and uses school facilities to support learning activities.	27.59	9.61	31.51	6.28	35.68	8.03	31.61	7.98
D-10	Principal manages school environment that promotes security, safety and health.	1.89	3.94	2.06	4.61	1.21	3.43	1.72	3.99
D-11	Principal manages the schedule and supervises teacher's additional assignments as well as extracurricular and cleaning activities in school.	3.69	5.83	3.25	6.20	3.88	7.40	3.61	6.48
D-12	Principal manages school administration.	1.89	3.31	1.19	2.31	1.13	2.68	1.40	2.76
D-14	Principal holds meetings with teachers.	0.82	1.34	1.51	4.13	1.05	3.78	1.13	3.08
E-13	Principal develops teaching schedule for teachers, conducts supervision over teacher's discipline, and ensures learning activities are implemented.	21.51	17.80	17.30	19.40	19.02	17.20	19.25	18.13
Total weighting for indicators directly related to learning		**24.8**	**27.1**	**21.2**	**32.9**	**21.8**	**28.0**	**22.6**	**29.3**

Note. The 3 (out of 14) indicator categories directly related to learning are B-05, B-06, and E-13, as shown in blue text.

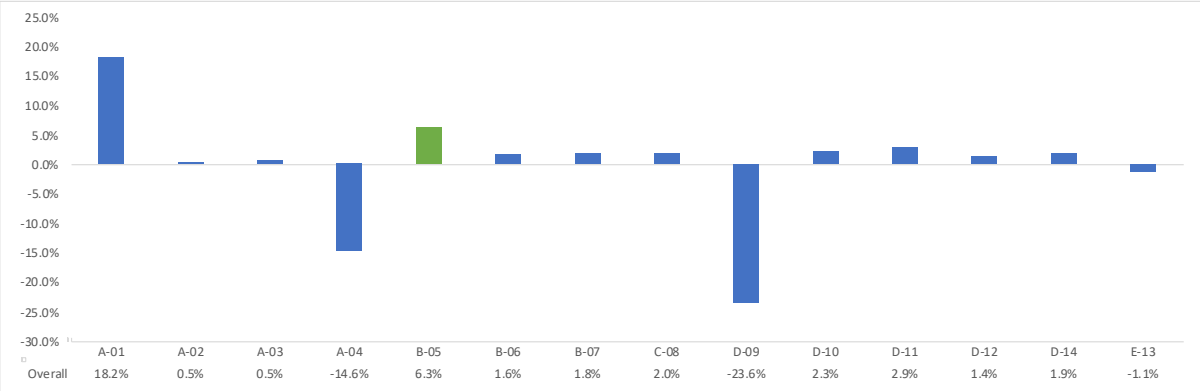
Figure B3. Changes in school leader service agreement indicators, by treatment group



	A-01	A-02	A-03	A-04	B-05	B-06	B-07	C-08	D-09	D-10	D-11	D-12	D-14	E-13
SAM	18.8%	1.5%	0.7%	-15.4%	5.6%	0.4%	1.2%	2.7%	-18.0%	2.0%	2.1%	1.4%	0.5%	-3.7%
SAM+Cam	17.7%	1.9%	0.5%	-13.8%	6.7%	1.4%	2.8%	2.3%	-27.7%	2.2%	3.5%	1.5%	2.7%	-1.8%
SAM+Score	18.2%	-2.1%	0.4%	-14.8%	6.7%	2.9%	1.4%	1.0%	-25.2%	2.5%	2.9%	1.1%	2.6%	2.1%

Notes. All values are within-treatment-group average changes, in percentage points. See Table B3 for full descriptions of each indicator category.

Figure B4. Changes in school leader service agreement indicators, average across treatment groups



	A-01	A-02	A-03	A-04	B-05	B-06	B-07	C-08	D-09	D-10	D-11	D-12	D-14	E-13
Overall	18.2%	0.5%	0.5%	-14.6%	6.3%	1.6%	1.8%	2.0%	-23.6%	2.3%	2.9%	1.4%	1.9%	-1.1%

Notes. All values are overall average changes, in percentage points. See Table B3 for full descriptions of each indicator category.

Appendix C Detailed analysis of coherence across relationships in case study schools

Figures C1, C2, and C3 categorize each stakeholder group's overall opinions of other village-level stakeholders, as expressed in interviews and focus groups during the field visits, for each of the three case study schools in each treatment group. These figures are summarized in Figure 5 in the main body of the paper, which shows the proportion of expressed views (excluding stakeholders' views of themselves) that are positive.

Stakeholder groups' overall opinions of other village-level stakeholders are classified as follows:

(a) *negative* (e.g., teachers are tardy; the UC doesn't bother checking school documentation; parents don't help their kids with homework);
(b) *neutral* or *don't know* (e.g., saying that they don't know what the school committee's duties are; or listing the committee's duties without stating whether committee members perform their duties effectively);
(c) *mixed* (i.e., having positive views about some actions but negative views about others; or some members of the stakeholder group having positive views and others having negative ones);
(d) *positive*, which can include some mildly negative views if on the whole the opinion is obviously positive (e.g., one teacher still disciplines children harshly, but all of the teachers are now punctual and hardworking); and
(e) *positive and explicitly oriented toward student learning* (e.g., students now study diligently; the quality of teachers' lessons has improved; because of parents' support more children now can read).

Figure C1. Stakeholders' views of each other in SAM-only case study schools

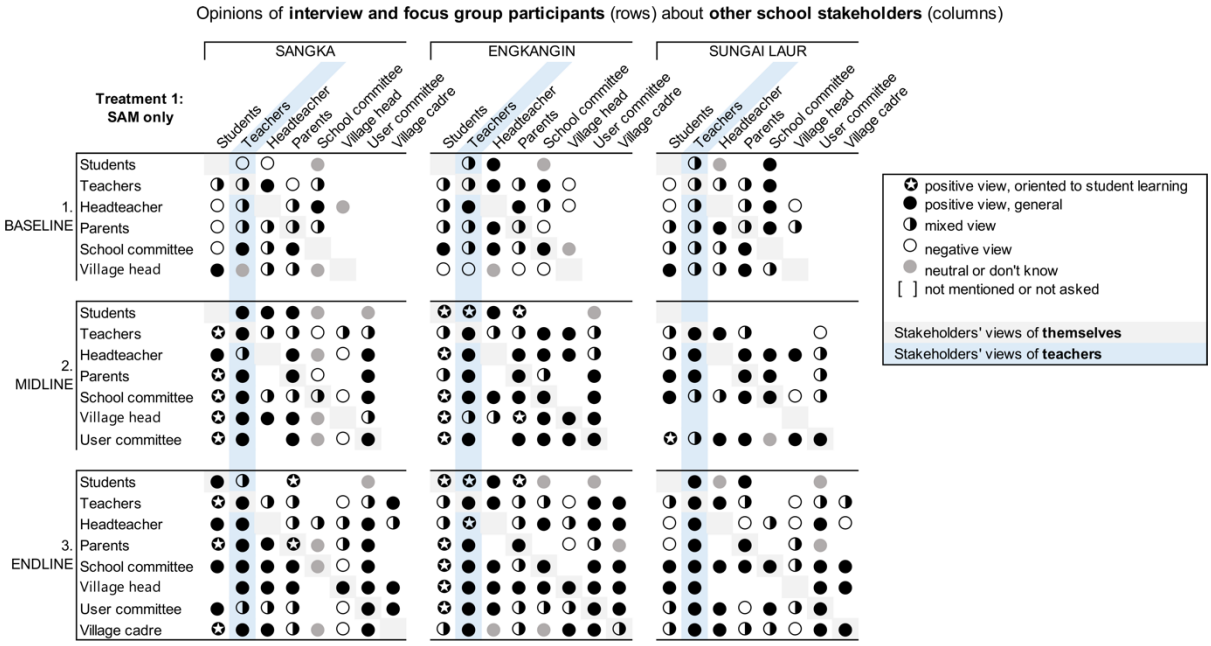


Figure C2. Stakeholders' views of each other in SAM+Score case study schools

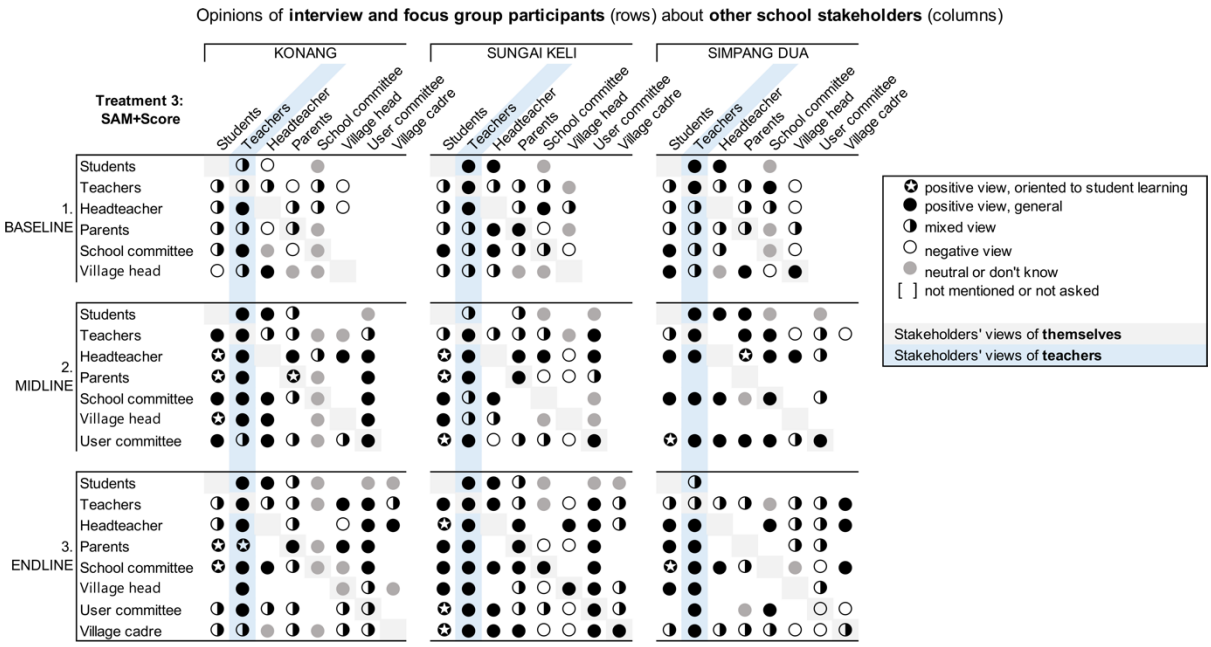


Figure C3. Stakeholders' views of each other in SAM+Cam case study schools