Working Draft: Comments welcome

Double for Nothing? Experimental Evidence on the Impact of an Unconditional Teacher Salary Increase on Student Performance in Indonesia

Joppe de Ree Karthik Muralidharan Menno Pradhan Halsey Rogers[†]

22 July 2015

Abstract: How does a large unconditional increase in salary affect employee performance in the public sector? We present the first experimental evidence on this question to date in the context of a unique policy change in Indonesia that led to a *permanent doubling* of base teacher salaries. Using a large-scale randomized experiment across a representative sample of Indonesian schools that affected more than 3,000 teachers and 80,000 students, we find that the doubling of pay significantly improved teacher satisfaction with their income, reduced the incidence of teachers holding outside jobs (and the hours worked on them), and reduced self-reported financial stress. Nevertheless, after two and three years, the doubling in pay led to no improvements in measures of teacher effort or student learning outcomes, suggesting that the salary increase was mostly a transfer to teachers with no discernible impact on student outcomes. While higher salaries may increase teacher quality on the extensive margin in the long run, we find no evidence of meaningful positive impacts of teacher salary increases on student learning from the intensive margin increases in teacher effectiveness that would be predicted by gift-exchange and efficiency wage models of employee behavior, as well as a model where effort on pro-social tasks is a normal good with a positive income elasticity.

JEL Classification: J31, J45, I21, C93, O15

Keywords: efficiency wages, gift exchange, fair wages, target wages, teacher salaries, teacher motivation, teacher performance, education quality, Indonesia, field experiments

[†] Joppe de Ree: World Bank; joppederee@gmail.com
Karthik Muralidharan: UC San Diego, NBER, BREAD, J-PAL; kamurali@ucsd.edu
Menno Pradhan: University of Amsterdam and VU University Amsterdam; m.p.pradhan@vu.nl
Halsey Rogers: World Bank; hrogers@worldbank.org

We thank Nageeb Ali, Julie Cullen, Gordon Dahl, Uri Gneezy, Richard Murphy, Ben Olken, Valerie Ramey, and several seminar participants for comments.

We are grateful to the Indonesian Ministry of Education and Culture for its interest in evaluating its teacher pay reforms, and for supporting this large-scale experiment and data collection. This evaluation would not have been possible without generous financial support from the government of the Kingdom of the Netherlands. The authors are grateful to Dedy Junaedi, Titie Hadiyati, Susiana Iskandar, and Amanda Beatty for their exceptional efforts in managing this evaluation as part of the World Bank BERMUTU project team at various points of time over the course of this project, and to counterparts at the Indonesian Ministry of Education, including Dr. Baedhowi, Ibu Yani, Yendri, and Dian Wahyuni, who supported this experiment and implemented it flawlessly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

1. Introduction

How does a large unconditional increase in salary affect employee performance in the public sector? While unconditional salary increases do not provide a direct incentive for increased effort on the job, there are several classes of models that predict improved worker effort in response to such pay increases. Prominent among these are models of "gift exchange" that posit that employees pay back employers for a wage premium with an effort premium (Akerlof 1982), and models of "efficiency wages" that posit that workers will shirk less in response to wage increases because of the increased cost of losing a job with such a wage premium (Shapiro and Stiglitz 1984).[1] A third mechanism (implicit in policy debates) is that if underpaid public workers take on outside jobs to meet their target incomes, then an increase in their pay would reduce the incidence of outside jobs and increase time and effort on their primary job.

Given the centrality of this question to labor and personnel economics, a large empirical literature has tried to test the gift-exchange model, with varying results (see Esteves-Sorenson and Macera 2014 for a recent review). However, since it is difficult to exogenously change salaries in real employment settings, most of the evidence to date has relied on laboratory experiments and short-term field experiments with researcher-led variation in pay. Thus, despite a large empirical literature on this question, we are not aware of any experimental study of the impact of a *permanent* unconditional salary increase in the context of an *existing* long-term employment contract. This is a critical gap because estimates from the existing literature are often interpreted as providing support for the importance of gift-exchange in real employment contracts (see Levitt and List 2007 for a discussion).[2]

In this paper, we attempt to bridge this gap by providing experimental evidence on the intensive-margin impacts on teacher effort and student learning outcomes of a unique policy change in Indonesia that permanently doubled the base pay of eligible teachers who went

[1] The standard efficiency wage model may not apply to a public-sector setting where workers have a low probability of being fired, but there are variants that may still apply. One mechanism highlighted in developing countries is that communities may be willing to condone shirking (and moonlighting) by teachers if it is widely believed that they are under-paid, whereas they would be less likely to do so if teachers were perceived to be well paid. In other words, there may be a "community enforced" efficiency-wage channel (World Bank 2003). But this is a conjecture based on qualitative studies, and there is little quantitative evidence to support it.

[2] As they note: "Such inference raises at least two relevant issues. First, is real-world on-the-job effort different in nature from that required in lab tasks? Second, does the effort that we observe in the lab manifest itself over longer time periods?" (Levitt and List 2007). This point also helps to explain why there have been so few experimental tests of the efficiency wage hypothesis (which require permanent wage increases over longer time periods).

through a certification process.[3] Given the large fiscal impact of the policy, teacher access to the certification program was phased in over 10 years (from 2006 to 2015), with priority in the certification queue being determined by seniority. Thus, many "eligible" teachers had to wait several years before being allowed to enter the certification process.

Working closely with the Government of Indonesia, we implemented an experimental design that allowed *all* eligible teachers in 120 randomly-selected primary and junior secondary public schools to *immediately* access the certification process and the resulting doubling of pay, while teachers in control schools experienced the "business as usual" access to the certification process through the gradual phase in over time.[4] The experiment thus created a sharp increase in the fraction of teachers in treated schools with a permanent doubling of pay during the three years of our study, which allows us to identify the intensive margin impacts of an unconditional permanent increase in pay on performance. Further, the experiment featured random assignment of 120 treatment and 240 control schools within a near-nationally representative sample of 360 schools across 20 districts and all major regions of Indonesia, thereby providing considerable external validity to our results.[5]

Given the challenges of implementing a randomized experiment at scale with a national government, the experiment worked remarkably well with a strong "first stage". The experiment led to a 29 percentage point increase in the fraction of all teachers in treatment schools who had been certified and paid the salary supplement at the end of two years, and a 23 percentage point increase at the end of three years. Among the teachers affected by the experiment (those who were eligible but not certified at the baseline), there was a 57 (and 44) percentage point increase

[3] The most important pre-requisites for "eligibility" were being a civil-service teacher and either having a four-year university degree or having a high rank in the civil service (rank IV) or having a very long tenure. The policy was designed to reward a process of teacher skill upgrading (signaled by "certification") by providing a professional allowance that was equal to the base pay (thereby doubling base pay). In practice, the skill upgrading component was considerably diluted. Thus, in effect, the certification mainly consisted of the pay increase (see section 2 for more details). To the extent that the certification program also led to improvements in teacher skills, our results will be an upper bound on the intensive-margin impact of an unconditional increase in pay.

[4] Roughly 20% of teachers in both treatment schools were already certified at baseline, and another 30% of teachers were not eligible for certification in any case (due to not being civil-service teachers or college graduates). It is the remaining 50% of teachers who were "eligible but not certified" at the baseline who are affected by the experiment and it is in this population of teachers that the experiment induces a significant increase in pay.

[5] See Heckman and Smith (1995) for a discussion of the threats to external validity of experiments resulting from site-selection bias in experimental studies. Allcott (2015) provides evidence of such bias.

in teachers who were certified and paid their professional allowance at the end of two (and three) years in treatment schools.[6]

The experiment also produced significant impacts on the intermediate mechanisms through which policymakers hoped that the increase in salary would lead to better education quality. At the end of two and three years of the experiment, teachers in treated schools had significantly higher income, were significantly more likely to be satisfied with their income, significantly less likely report financial stress, and significantly less likely to hold a second job.

Yet despite this improvement in teachers' pay and satisfaction, there was no impact on teacher effort towards upgrading their own skills, on teacher effort in the classroom, or on the ultimate outcome of student learning. Teachers in treated schools do not score better on tests of teacher knowledge of either subject matter or pedagogy, and do not report any increase in self-reported measures of effort such as teacher attendance, or the number of teaching hours. Most importantly, we find no difference in student test scores in language, mathematics, and science across treatment and control schools for both primary and junior secondary schools. The test score impact of being in a treated school is not only insignificant, but the point estimates are close to *zero*. The zero effects on learning are also quite precise, allowing us to rule out effects as small as 0.12σ at the 95% level in treated schools.

These are intention-to-treat estimates at the school level and reflect a lack of impact on teacher effort and student outcomes in a setting where the fraction of certified teachers was 29 (and 23) percentage points higher in treated schools over two (and three) years. To estimate the impact of being taught by a certified teacher, we restrict our analysis to students who were taught by teachers who were "eligible but not certified" in either the treatment or control schools over the course of the study (since these were the teachers who were the "targets" of the intervention). We use the school-level random assignment as an instrument for being taught by a certified teacher in a given year, and again find no effect of being taught by a certified teacher (relative to students in control schools taught by similar teachers). The point estimate is again close to zero and we can rule out positive test score effects larger than 0.15σ at the 95% level.

Thus, in contrast to the empirical literature that has found evidence supporting the gift-exchange hypothesis in the lab (Fehr et al. 1993 and 1997) and in short-term field experiments

[6] Note that the "first stage" of the experiment will weaken over time in our setting as teachers in the control schools get certified over time. This also explains why the difference in the fraction of certified teachers across treatment and control schools was not 100% (teachers in the control schools were also getting certified, but at a slower rate).

(Falk 2007), our results are consistent with a growing body of evidence suggesting that increases in worker productivity in response to an unconditional increase in pay, as posited by the gift-exchange model, are either short lived (as in Gneezy and List 2006; and Jayaraman et al. 2014) or non-existent when measured net of other confounding factors (as in Esteves-Sorenson and Macera 2014).

The main contribution of this paper is that it presents the first experimental evidence on the impact of a permanent wage increase on performance in the context of an existing employment contract, as opposed to researcher-led experiments (in both the lab and the field) that have only varied pay in the short run and have typically been for a new employment contract. It is also (to our knowledge) the largest wage increase experiment ever conducted, both in terms of the size of the wage increase studied and the fiscal commitment represented by the policy (which will cost around five billion US dollars *each year* in steady state), and the scale and duration of the experiment (done in a near-representative sample across a country of 200 million people for three years).

Our results also contribute directly to the literature studying the links between teacher pay and performance and are consistent with prior evidence finding no correlation between increases in teacher pay and improved student performance in the US (Hanushek 1986; Betts 1995; Grogger 1996; Ballou and Podgursky 1997). However, these results have been questioned both for not having adequate exogenous variation in teacher pay, and for being based on small changes in pay that may not be enough to detect an impact on outcomes (Dolton 2011). We address both these limitations of the existing literature in our setting.

Our results do not imply that a policy of unconditional salary increases would have no positive impacts on service delivery in developing countries in the long run. Dal Bo et al (2013) show that salary increases for public sector jobs in Mexico increased the observable quality of job applicants, and Ferraz and Finan (2011) find that higher wages for politicians in Brazil led to improved performance through both a selection channel and an efficiency-wage channel. However, Dal Bo et al. (2013) are not able to study the impact of higher public sector wages on performance outcomes or to estimate the cost effectiveness of such a policy, and the results in Ferraz and Finan (2011) are from a setting where non-performing politicians are more likely to lose their jobs (where an efficiency wage channel is more likely to apply).

Our results complement these by showing that even large unconditional wage increases may yield no improvement in performance on the intensive margin in a public sector setting of "permanent" civil-service employment contracts with a low probability of being fired for non-performance. To the extent that several education policy reports advocate increasing teacher pay as a way to improve teacher performance on the intensive margin (UNICEF 2011, UNESCO 2014), our results are highly-policy relevant and suggest that this hypothesis is not supported by the evidence. Thus, policy makers hoping to increase the quality of government service delivery by increasing salaries across the board need to trade off the potential benefits on the extensive margin against the large intensive-margin costs of unconditional increases in public sector pay that may not yield any performance improvement. Further,

The rest of this paper is structured as follows: Section 2 describes the Indonesian education context and the teacher certification policy; section 3 describes our experiment (design, validity, and data collection); section 4 presents results on teacher effort and student outcomes, and section 5 discusses policy implications, caveats, and directions for future research.

2. Background and Policy Context

Indonesia has one of the largest school education systems in the world, serving more than 50 million students spread across 34 provinces and more than 500 districts.[7] The country consists of thousands of islands spanning over 3000 miles from east to west (see Figure 1). Primary education was historically a high priority for Indonesia relative to other developing countries in South Asia and Africa, and Indonesia achieved high rates of primary school enrollment exceeding 90% by the early 1980s. Nevertheless, the performance of the education system on learning outcomes is low compared to that of many other middle-income countries. On the 2012 PISA international mathematics assessment, for example, Indonesian 15-year-olds' average scores were below their peers in all other participating countries except Colombia, Peru, and Qatar, with similar performance on reading (OECD 2013). On the 2011 TIMSS math assessment, Indonesian 8th-graders outscored those from only five other countries (Mullis and other 2012).

Education policy discussions in Indonesia in the years prior to 2005 identified poor teacher quality and motivation as a key limitation in the performance of the Indonesian education

[7] "District" here refers to the 2nd-level administrative subdivision, known as *kabupaten* and *kota* in Indonesian.

system. The ambitious education reforms of 2005 aimed explicitly to address this issue and made a large fiscal commitment to doing so. The highlight of the 2005 Teacher Law was that teachers who met certain eligibility criteria (being a civil-service teacher and either having a four-year university degree or having a high rank in the civil service or having a very long tenure), and who successfully completed a certification process, would receive a "professional allowance" equal to 100% of their base pay (Al-Samarrai et al. 2013; World Bank 2010).[8] The certification process was initially meant to include a high-standards external assessment of teacher subject knowledge and pedagogical practice, with an extensive skill upgrading component for teachers who did not meet these standards that would include up to a year of additional training and tests (Chang et al. 2013, Chapter 1).

Thus, the reform aimed to both increase the average skill level of teachers (by providing a financial incentive for doing so), and to improve teachers' financial situation and hence their ability to focus on their teaching. Stated rationales for the policy included increasing the social respect accorded to teachers, improving their motivation and reducing their financial stress, increasing their ability to focus on teaching by reducing their need to moonlight and hold second jobs, and making teaching more attractive as a career (World Bank 2010). Using representative household survey data, we estimate that the doubling of base pay moved teacher compensation from the 40th percentile of the college-graduate salary distribution to the 80th percentile.[9]

However, this very large salary increase was not conditional on teachers' subsequent effort or effectiveness, but instead depended only on a one-time determination that the teacher met certification certain criteria. Further, by the time the final law got negotiated through the political process (with strong opposition from teacher unions to the idea of taking high-standards certification exams), the quality improvement stipulations for certification were highly diluted and replaced with a much weaker certification requirement that simply required teachers to submit a portfolio of their teaching materials, with two weeks of additional training for those who did not pass an evaluation of this portfolio. In practice the certification process yielded a

[8] Note that the professional allowance was 100% of base pay and not total pre-certification pay. Teachers often receive other allowances based on location of posting and taking on additional tasks, and so the professional allowance increased total pay by 80% on average and 67% for teachers who were eligible for treatment (see Table 5)

[9] This calculation is based on salaries alone and does not include (the likely more generous) pensions and benefits for civil-service teachers, and the value of having much higher levels of employment certainty relative to the private sector. If these were accounted for, it is likely that teacher compensation was higher in the distribution of college graduate compensation both before and after the reform than the numbers presented here.

doubling of base pay with only a modest hurdle to be surmounted.[10] Thus, for practical purposes, the policy can be considered as having resulted in an unconditional salary increase for eligible teachers. To the extent that undergoing the certification process increased teacher human capital in any way, any estimate of the impact of certification will be a lower bound on the intensive margin impacts of an unconditional increase in pay.

Because of the large number of teachers covered, teacher access to the certification process was phased in for budgetary reasons. The budgetary restrictions meant that only around 10% of teachers were allowed to go through the certification process each year since implementation of the certification process began in 2006. Each year, districts were allocated a quota that indicated how many of their teachers could start the certification process. Once in the process, certification was practically guaranteed as describe above. Other eligible teachers therefore had to wait in a certification queue, sometimes for several years, with their position in the queue determined by their educational qualifications and seniority.

3. Experiment Design

3.1. Design, Sampling, and Implementation

Our experimental design takes advantage of the phase-in procedure for teacher access to the certification process. Rather than having teachers wait in the certification queue, the intervention aimed to allow all *eligible but not yet certified teachers* (we define these as "target" teachers) in treatment schools to immediately access the certification process at the start of the experiment (in 2009). Note that the experiment did not change any of the requirements of certification as per the law, but simply allowed teachers in treatment schools to not have to wait for a few more years to access the certification process. The experimental protocol was implemented in close collaboration with the Ministry of National Education of the Government of Indonesia, where senior officials were committed to conducting a high-quality impact evaluation, and provided exemplary support in implementation.

We first identified a near-representative sample of 360 schools across 20 districts of Indonesia to comprise the universe of the study. We started with the 2006 national teacher census, which covered roughly 1,600,000 public primary and junior-secondary teachers across

[10] Field anecdotes suggest that very few teachers entering the certification process failed it (partly due to the emergence of a market in preparing teaching portfolios that would satisfy the certification criteria), and that even those who did were all certified after a two-week training program (World Bank 2010).

454 districts. Districts that were too small, were too dangerous to visit, or that were included in a parallel randomized evaluation were excluded[11], leaving us with 383 districts in the sampling frame. These represented nearly 85% of the districts and over 90% of the population of Indonesia. From these, we randomly sampled 20 districts, stratified across the five major regions of the country, with more districts assigned to regions with a larger population. The list of districts sampled and the stratum they represent are presented in Table A.1. A map of the sampled districts and their representativeness is presented in Figure 1.[12]

Within each district, we stratified schools by the number of teachers, and sampled 12 primary and 6 junior secondary schools (stratified by school size).[13] Thus, the study universe consisted of a near-representative sample of 240 primary and 120 junior secondary schools across 20 districts of Indonesia. 120 of these schools (80 primary and 40 junior secondary) were then randomly assigned to "treatment" status while the other 240 schools (160 primary and 80 junior secondary) were assigned to a "business as usual" control group. Just like the sampling of schools, the randomization was also stratified by district, school-type, and school size, and thus the design was identical across districts, with each district being a microcosm of the overall study.[14]

Teachers in treatment schools who were eligible for certification, but not yet certified, received a personal letter from the Ministry of National Education informing them that they had been granted immediate access to the certification process. Only teachers who worked in the treatment schools at the start of the experiment were eligible for this immediate access to make sure that there would be no incentive for teachers to transfer to treatment schools. The budget for the extra certification "slots" created for the experimental study was provided through

[11] Note that the district sampling for the two parallel sets of randomized evaluations were conducted using the same procedures, and so the 20 districts dropped on account of not wanting spillovers between the studies were also a representative sample. However, the second study ended up not being implemented. Note also that the districts dropped for access and safety reasons had much lower population on average.

[12] The five major regions of Indonesia and the number of districts sampled in each of them (roughly proportional to population) include Java (10), Sumatra (5), Sulawesi (2), Eastern Indonesia (2), and Kalimantan (1). As the scale in Figure 1 shows, the East-West distance spanned by Indonesia is greater than that of the continental United States, and the design imposed considerable logistical complexity. However, the resulting random assignment in a near-representative sample of schools provides greater external validity to our results.

[13] We dropped the strata comprising schools with very large and very small number of teachers. If schools were too large, it would not have been feasible to test all the students in the school during the team that the enumerators would have in the school. If they were too small, they would not provide adequate power. Note that primary schools cover grades 1-6, while junior secondary schools cover grades 7-9.

[14] Specifically, each of the 20 districts had 6 treatment schools (2 junior secondary and 4 primary) and 12 control schools (4 junior secondary and 8 primary). Schools were stratified into "triplets" based on size and one school in each triplet was assigned to treatment status. Note that the intervention was very expensive and thus, optimal sample allocation to maximize power yielded a larger control group than treatment group.

supplementary funds from the National Government, and these slots were provided to districts over and above their regular certification quota. The research design did not create any other change in the schools besides the additional quota allocation to treatment schools, and the personalized letter sent to the "target" teachers (who were eligible but not certified at the start of the 2009-2010 academic year). The teachers in control schools continued business as usual, and those who were eligible but not certified at the start of the study progressed through the certification process at the same rate as the rest of the country. Thus, our identifying variation comes from the sharp increase in the fraction of certified teachers in the treatment schools during the experiment.

The possibilities of spillovers to other schools were minimized by making sure that there was no public announcement of the additional quota and that the eligibility for certification was communicated to teachers only by the personalized letter that they received from the Government. Further, since the teachers who did not receive the certification letter within the treatment schools were not eligible for certification in any case (by virtue of not being a college graduate or a civil-service teacher), the experiment is less likely to have engendered resentment on the part of other teachers in the school relative to a setting where the pay increases may have been seen as arbitrary. Thus, by conducting our study in a setting where the pay increases were in line with pre-announced policy criteria, we minimize the extent to which the intervention may be considered ad hoc or unsustainable.

3.2. Project Timeline and Data

The school year in Indonesia runs from August to May, and the experiment was carried out over three school years from 2009-10 to 2011-12 (and we refer to these three years as Y1, Y2, and Y3 in the paper). The sampling and randomization of schools was conducted during the summer holidays before Y1, and the government sent letters to eligible uncertified teachers announcing their access to the certification process at the start of the school year. The certification process (including preparing and submitting the application and teaching portfolio, having this evaluated, and receiving the certification) typically took one full school year, and teachers typically got "certified" by the end of Y1, and started receiving their certification allowance (equal to 100% of base pay) at the start of Y2.

We collected three waves of data during which we interviewed head-teachers, teachers, and students, and conducted independent tests of both teacher knowledge, and student learning outcomes. The first wave was a baseline collected in October 2009, which we refer to as Y0. The baseline was deliberately conducted a few months into the school year (after the certification eligibility letters were sent to teachers in treatment schools) to be able to verify using teacher-level interviews that they had in fact received these letters and entered the certification process. The second wave of data was collected in April-May 2011 at the end of 2 years of the project (Y2), and the third wave was collected in April-May 2012 at the end of 3 years (Y3).[15] The timeline of the project including both intervention and data collection is shown in Figure 2.

We collected data on school facilities, finances, and other school-level data from head-teacher interviews. Teacher interviews included questions on demographics, experience, pay, outside jobs, income (from teaching and other sources), and job satisfaction. We used a combination of school and teacher interviews to map teachers to specific classrooms and subjects (which will not be needed for the school-level ITT estimates, but will be needed for the IV estimates of the impact of being taught by a certified teacher). Students in all schools were tested on multiple choice tests of math, science, and Indonesian, and students in junior secondary schools were also tested in English. The tests also included a short demographic survey where students filled in basic information on household assets.

3.3 Validity of Experimental Design

The randomization was successful in ensuring that treatment and control schools were similar prior to the experiment. There was no significant difference between treatment and control schools on school-level variables such as the number of students, teachers, or student teacher ratio (Table 1- Panel A). There were also no significant differences in student test scores across treatment and control schools on test scores in any subject (math, science, Indonesian, or English) or in an index of household assets (Table 1 - Panel B).[16]

Similarly, we see no significant difference in teacher characteristics across treatment and control schools either. There were no significant differences on teacher-level variables including their own test scores, their certification status, their base pay, or the incidence of holding an

[15] Since the certification process took one year, the first year in which teachers who entered the certification process as a result of the experimental intervention would have received the additional allowance was in the second year of the project. We therefore felt that it was highly unlikely that there would be any impact at the end of Y1 (since teachers in treatment schools would not have received any additional payments at this point). Thus, given the high costs of surveys across the Indonesian islands, we did not collect data at the end of Y1.

[16] Note that the randomization (and communication to "target" teachers was carried out before the baseline survey) and hence the randomization could not be balanced ex ante on these variables. Thus, it is reassuring to see that treatment and control schools were balanced on observables.

outside job (Table 2 - Panel A). The only difference (which is as expected) is that teachers in treatment schools are 32 percentage points more likely to have entered the certification quota confirming that the intervention successfully led to many more teachers in treatment schools getting access to the certification process.

We see the impact of the treatment even more clearly in Table 2 – Panel B, which is restricted to the "target" teachers who were "eligible but not certified" in either the treatment or control schools at the start of the study. In this group, 76% of teachers in treatment schools were in the certification quota, whereas only 20% of those in the control group were (which is the rate at which eligible but uncertified teachers would have gotten certified in the absence of the experiment). All other teacher characteristics are identical on average as expected. The focus of our analysis will be on school-level ITT (using the sample of all teachers as shown in Panel A) and estimates and IV estimates of being taught by a certified teacher (using the sample of "target" teachers as shown in Panel B).[17]

In addition to balance on initial characteristics across treatment and control schools, we also test for differential attrition and entry of students over the period of the study. Table A.3 shows the different cohorts in our study, the years in which they were tested, and which cohorts are in our estimation sample at different points of the study. We find that there is no differential attrition among students who were in our baseline test and who continue to be in our estimation sample over time (Table A.4), and also find that the treatment does not seem to have induced any compositional changes in incoming student cohorts over time (Table A.5.)

4. Results

4.1 First-Stage

The time path of the fraction of teachers in treatment and control schools who had entered the certification process over the three years of the study is shown in Figure 3. Three points are noteworthy. First, there was no difference in the rate of teacher certification between treatment

[17] For completeness, we also show balance for teachers who were already certified and for those who were not eligible for certification (Table A.2). Teacher characteristics continue to be balanced in both these sub-groups as well. However, we see that 12% of teachers who were classified as non-eligible for certification were in fact in the certification quota in the treatment schools (compared to just 1% in control schools). This is because in practice, there were schools/districts that did not strictly enforce the requirement that teachers had to be both college graduates and civil-service employees to be eligible for the certification. Note that this does not affect the validity of our ITT estimates (which are done at the school level). Similarly, our IV estimates will also not be affected by this, because they will be based only on students taught by the "target" teachers described in Table 2 – Panel B.

and control schools before the start of the experiment in 2009. Second, the intervention introduced a sharp increase in the number of certified teachers in treatment schools in 2009, even as the trend in control schools remained constant. Third, the gap in fraction of teachers certified reduced over time, as the eligible teachers in the control schools gained access to the certification process over time at a "business as usual" rate. Thus, the difference in the fraction of certified teachers across treatment and control schools is higher at the end of Y2 than at the end of Y3.

As described earlier, teachers entered the certification process at the start of each school year, completed the process over the course of the year, got certified by the end of the year, and started receiving their payments at the start of the next year. Thus, at the time of the baseline there was no difference between treatment and control schools in the fraction of teachers who were certified or who had received the extra certification allowance. However, there was a sharp increase in both of these indicators at the end of Y2 and Y3 (Figure 4).

Table 3 - Panel A shows the differences in Figures 3 and 4 along with tests of equality. In the first year, treatment schools have 31.8 percentage point more teachers who had entered the certification process (more than double that of the control group), while there was not yet any difference in the fraction certified or paid the certification allowance. At the end of Y2 and Y3, the difference in the fraction of teachers who had entered the certification process falls to 16.7 and 7.5 percentage points respectively (since the control schools "catch up"). At the end of Y2 (Y3), the fraction of teachers in treatment schools who report being certified is 23 (14.2) percentage points higher, and the fraction who report being paid the certification allowance is 29 (22.5) percentage points higher. The difference in both indicators falls over time as teachers in the control schools catch up with the certification process.

Note that the difference in fraction of teachers who are paid their certification allowance is higher than the difference in the fraction who are certified (in both Y2 and Y3), because teachers in the control schools who would have entered the certification process at the start of Y2 and Y3, would have gotten certified at the end of Y2 and Y3 respectively, but would only have started getting paid their allowances at the start of the next school year. These teachers will therefore report being certified but will not yet have started getting paid their allowance at the time of the Y2 and Y3 surveys. On the other hand, teachers in treatment schools who gained access to the certification process at the start of Y1 will have completed getting certified by the end of Y1, and started getting paid their allowances in Y2.[18] Since most of the posited mechanisms by which the pay increase would be expected to improve teacher effort and student outcomes are based on teachers actually receiving the extra pay, the most relevant metric of the "effective difference" between treatment and control schools for our study is the difference in the fraction of teachers who have been "paid their certification allowance".

In addition to school-level average differences, we also show the impact of being in a treated school for each of the three categories of teachers: teachers who were "eligible but not certified" and were the "targets" of the intervention, teachers who were "already certified" and teachers who were "not eligible" (because they did not have a college degree or were not civil service teachers). As expected, we see most of the differences in the school-level averages being driven by the target teachers for whom there is a 56 percentage point increase in the probability of entering the certification process. At the end of Y2 (Y3), they are 45 (25) percentage points more likely to be certified, and 57 (44) percentage points more likely to have been paid their certification allowance (Table 3 - Panel B). By definition, there is no impact on teachers who were already certified (Table 3 - Panel C).

For the teachers who were not eligible as per the official norms of the Ministry of National Education, we do see a small impact of being in a treated school, with an 8-9 percentage point increase in the fraction of teachers who are certified and paid at the end of Y2 and Y3. These most likely reflect cases where teachers may have possessed alternative credentials that were acceptable as a basis for certification eligibility in lieu of a college degree (which is the basis on which we classified the eligibility status of teachers), that made them eligible for certification despite our classifying them as ineligible.[19] Since we focus on school-level intention to treat effects, the breakdown presented in Table 3 – Panels A to C are presented more to provide clarity on how the experiment affected the three types of teachers.

[18] Thus, the differences between treatment and control groups across measures reflects variation in the year of entry into the certification process and the time lag in the process. Once we control for year of entry into certification, the difference between treatment and control schools in the fraction of teachers who are certified and the fraction who are "certified and paid" is the same.

[19] The certification rules were also ambiguous with regard to whether teachers had to be civil service teachers to be eligible for certification. The rules were clear that civil service would have the *first priority* for entering the certification process (which was rationed with an annual quota). Thus, all teachers who were certified prior to the experiment were in fact civil service teachers. However, the Government had not yet taken a definitive decision as to whether non-civil service teachers who were college graduates would have access to certification at the end of the queue (since this decision would only have to be made 5-6 years after the start of the certification process, it had not yet been made clearly). Thus, some teachers in our treatment schools who were not civil-service teachers may have entered the certification process, but the government later ruled that only civil-service teachers could get certified.

4.2 Teacher-level Outcomes

We find that the accelerated access to the certification process and the additional allowance had several positive impacts on teachers that persisted both two and three years into the experimental study. At the end of Y2 (Y3), teachers in treatment schools received 96% (54%) more certification pay and 15% (11%) more total pay compared to those in control schools, were 14% (12%) more likely to report being satisfied with their total income, 18% (16%) less likely to report facing financial problems and stress, and were 18% (18%) less likely to be holding a second job (Table 4 – Panel A).[20]

These effects are considerably stronger within the universe of "target" teachers, within whom teachers in teachers in treatment schools received 266% (104%) more certification pay and 30% (23%) more total pay compared to those in control schools. Note that the certification was 100% of base pay for teachers, but that in practice, the increase over their total pre-certification pay was around 70-85% because the total pay (prior to certification) would have included a few allowances in addition to their base pay.[21] hey were 28% (24%) more likely to report being satisfied with their total income, 32% (38%) less likely to report facing financial problems and stress, and were 20% (25%) less likely to be holding a second job at the end of Y2 (Y3) (Table 4 – Panel B). The corresponding changes for teachers who were already certified and for those not eligible for certification are shown in Table A.6.

Thus, the pay increase was successful in achieving its stated objectives regarding teachers' financial situation, job satisfaction, and ability to better focus on teaching by reducing the need to hold outside jobs. However, we find no evidence to suggest that teachers in treatment schools put in greater effort in response to this pay increase. We find no difference between treatment and control schools on teacher test scores or the likelihood of pursuing further education, suggesting that teachers did not upgrade their skills as a result of the program. We also find no difference in self-reported teaching hours per week or absence rates, suggesting that teacher

[20] These figures are presented in percentage changes relative to the mean in the control group. The tables present the changes in percentage points.

[21] It is easy to back this out from the figures in Tables 3 and 4. In the sample with all teachers, we see in Table 3 that 55.8% of teachers in the treatment group had been paid the certification allowance, and see in Table 4 that the mean certification pay received by this group was 1.111million IDR (million Indonesian Rupiah). Thus, average certification pay conditional on receiving it was 1.111M/0.558, which is 2 million IDR. This is an 83% increase over the mean base pay of 2.39 million IDR. The calculation can also be done with the "target" teachers, where we see that the average certification pay conditional on receiving it was 1.5M/0.76, which is again 2 million IDR. But since other allowances for civil service teachers were higher, the pre-certification pay for the "target" teachers was 2.9M. Thus, certified teachers received a 69% increase (2/2.9) in their total pay.

effort was also unchanged. These results hold in both the overall sample of teachers (Panel A) as well as the sample that is restricted to "target" teachers who had an even larger increase in pay. Nevertheless, as per the theoretical framework described in section 2, it is possible that the reduced financial stress, reduced incidence of second jobs, and increase job satisfaction may lead to an improvement in teacher effectiveness as measured by student learning outcomes.

4.3 Student Outcomes

4.3.1 Intention to Treat (ITT) Estimates

Since the randomization was conducted at the school level, we first present school-level intention to treat estimates of the impact on student learning outcomes of being in school that had a sharp increase in the fraction of certified teachers who had received a large unconditional increase in pay. Our main estimating equation takes the form:

$$T_{ijks}(Y_n) = \beta_0 + \beta_j \cdot T_{ijks}(Y_0) + \beta_2 \cdot Treatment_k + \beta_{Z_{ST}} \cdot Z_{ST} + \varepsilon_{ijks} \quad (1)$$

The dependent variable of interest is T_{ijksd}, which is the normalized test score of student *i* on subject s, where *j, k,* denote the grade, and school respectively. $T(Y_0)$ indicates the baseline tests, while $T(Y_n)$ indicates a test at the end of *n* years of the program. Including the normalized baseline test score improves efficiency due to the autocorrelation between test-scores across multiple periods.[22] We also include a set of stratum fixed effects (Z_{ST}), to absorb geographic variation and increase efficiency, and to account for the stratification of the randomization (which was done within district-level "triplets" of schools as described in section 3.1). The main estimate of interest is β_2, which provides an unbiased estimate of the impact of being in a "Treatment" school (the intent-to-treat or ITT estimate) since schools were assigned to "Treatment" status by lottery. We estimate β_2 both with and without controlling for school and household characteristics (X_i) shown in Table 1.

We present these results in Table 5 both combined across school types (Panel A) and also separated by primary schools (Panel B) and junior secondary schools (Panel C). We present results individually for each subject, and also pooled across subjects, and finally present results separately by Y2 and Y3. Overall, we find no evidence that students in treatment schools (with a

[22] As we show in Table A.3, some of the cohorts included in our analysis did not have a baseline test. We set the normalized baseline score to zero for these students (similarly for students who may have been absent at the time of the baseline test but are present in the Y2 and Y3 tests) and include a dummy variable in equation (1) that takes the value 1 when the lagged test score is missing and 0 when it is present. We also allow the coefficient on the lagged test score to vary by grade.

significant increase in the fraction of certified teachers) scored any better than those in control schools. Not a single effect (in any subject, in either type of school, or at either of the two time periods) is significantly different from zero, and the pooled effects across subjects and school types have a point estimate of 0.00σ at the end of Y2 and 0.02σ at the end of Y3. These zero effects are very precisely estimated with standard errors under 0.05σ, which provides us adequate power to detect effects as low as 0.1σ at the 5% level. Thus, not only are the point estimates close to zero, but we can reject effect sizes greater than 0.095σ at the end of Y2 and effect sizes greater than 0.12σ at the end of Y3.

Figure 5 presents quantile treatment effects of being in a treatment school, by plotting student test scores at each percentile of the control and treatment school test score distribution after Y2 and Y3. We see that the treatment effects are not only zero on average, but close to zero at every part of the test score distribution. While quantile treatment effects are based on comparing students at the same percentile in treatment and control *end-line* distributions, a different way of examining the heterogeneity of the results non-parametrically is by plotting the end-line test scores by treatment status at every percentile of the baseline test-score distribution.[23]

We do this in Figure 6, where the left-hand side panel plots the probability of a student at each percentile of baseline test-scores being taught by a "certified and paid" teacher for at least one year during the duration of the study (by treatment and control groups), and the right-hand side plots student test scores non-parametrically at the end of the study period on the same horizontal axis. We see in Figure 6 - Panel A, that students in treatment schools were 25% more likely to have been taught by a teacher who was "certified and paid" in Y2, but that their test scores are unchanged at every percentile of the baseline test score distribution. Figure 6 - Panel B shows the Y3 results, which are even more stark. Students in treatment schools (at every percentile of the baseline test score distribution) had over 0.5 years of being taught by a "certified and paid" teacher during the two year period comprising Y2 and Y3, but their test scores are unchanged at the end of Y3.

One issue in interpreting our school-level ITT estimates is that it is possible that the estimated zero effects result from a combination of positive effects on students taught by teachers who were "targets" of the experimental intervention (who are motivated to increase

[23] These are different because rank order is not preserved across students between baseline and end-line tests. The non-parametric plots as a function of baseline test scores also have fewer data points because we do not include the students/cohorts for whom we do not have a baseline test scores.

effort by the pay raise) and negative effects on students taught by "non-target" teachers (especially those who were not eligible for certification), who may have withdrawn effort in response to the perceived "unfairness" of not receiving the certification allowance.[24] We test for this possibility by decomposing the composite results shown in Table 5 by students taught by "target" teachers and those taught by non-target teachers (across treatment and control schools) and present the results in Table 6.

For the Y2 data, we simply consider whether a student was taught by a target teacher in Y2 (since none of the teachers affected by the treatment would have been paid the certification allowance in Y1), and find no significant difference in the outcomes of these students across treatment and control schools in any subject or in either type of school (Table 6 - Panel A). For the Y3 data, we consider the four possible combinations of teacher type that a student could have had in Y2 and Y3 (target – target; target – non-target; non-target – target; and non-target – non-target) and again find no significant different in test-score outcomes across these categories between treatment and control schools. Focusing on the most extreme comparison of students in treatment schools who were taught by a target teacher in both Y2 and Y3 versus those taught by a non-target teacher in both Y2 and Y3, we still find no evidence that the former did better (if anything, the point estimates on those taught by non-target teachers in both years are slightly higher for all subjects).

4.3.2 Instrumental Variable (IV) Estimates

The ITT estimates presented above are at the school level, and are based on a 29 (23) percentage point increase in the fraction of "certified and paid" teachers in the treatment schools. To estimate the direct impact of being taught by a certified teacher, we restrict ourselves to the students who were taught by a "target" teacher and instrument for being taught by a certified teacher using the random assignment of treatment across schools. Further, to make the most efficient use of our data we pool both Y2 and Y3 data and let the endogenous variable be the number of years a student was taught by a certified teacher. Specifically, we aim to estimate:

$$T_{ijks}(Y_n) = \beta_0 + \beta_j \cdot T_{ijks}(Y_0) + \beta_2 \cdot N \cdot Certified_{jks} + \beta_{Z_{ST}} \cdot Z_{ST} + \varepsilon_{ijks} \text{ (2)}$$

$$T_{ijks}(Y_n) = \beta_0 + \beta_j \cdot T_{ijks}(Y_0) + \beta_2 \cdot \left[Certified_{jkst} + Certified_{jks(t-1)}\right] + \beta_{Z_{ST}} \cdot Z_{ST} + \varepsilon_{ijks}$$

[24] As described earlier, the design of the experiment would have mitigated against this possibility (because the experiment did not change any of the certification norms stipulated in the law). But we still test for it.

where the coefficient of interest is β_2, which estimates the impact on student test-scores for each year (*N*) of being taught by a *Certified* teacher (with the additional pay), and the rest of the variables are defined as in Eq. (1).

One technical consideration in estimating Eq. (2) is the issue of test-score decay (or incomplete persistence) over time. Specifically, Eq. (2) assumes that there is no decay in test scores over time (or that persistence is complete). In practice, estimates from several settings suggest that there is considerable annual decay in test scores, with the persistence parameter Υ (estimated as the coefficient on the lagged test score in a standard value-added model) typically being around 0.5 (Andrabi et al. 2013, Muralidharan 2012).

We therefore estimate:

$$T_{ijks}(Y_n) = \beta_0 + \beta_j \cdot T_{ijks}(Y_0) + \beta_2 \cdot [Certified_{jkst} + \Upsilon \cdot Certified_{jks(t-1)}] + \beta_{Z_{ST}} \cdot Z_{ST} + \varepsilon_{ijks} \tag{3}$$

using the sample of students taught by a target teacher (in both treatment and control schools) and instrument for being taught by a certified teacher by the treatment status of the school. Since it is not possible to jointly estimate the persistence parameter and an unbiased experimental treatment effect at the same time (see Andrabi et al. 2013 and Muralidharan 2012 for further discussion), we estimate Eq. (3) for different values of Υ and present estimates of β_2, along with standard errors for a range of values of Υ from 0 to 1 in Table 7. The estimates with $\Upsilon = 0$ correspond to complete decay of any test score gains in a year by the end of the next year, while those with $\Upsilon = 1$ correspond to complete persistence. Based on several prior studies, our preferred estimates assume $\Upsilon = 0.5$.

The main threat to interpreting these estimates as the annual impact of being taught by a certified teacher (at different persistence rates) is the possibility of endogenous re-assignment of certified teachers within treatment schools to potentially weaker students. We test for this in Table A.6 and find that there is no significant difference in the characteristics of students assigned to target teachers across treatment and control schools during the first year of the project (Y0), and also during the second and third year of the project (measured at Y2 and Y3).

Thus, the results in Table 7 use the experiment to credibly show that the causal impact of being taught by a certified teacher on student test score gains is close to zero. Combining the point estimates with the standard errors we see that we can reject a positive effect greater than 0.15σ at the 95% level. Thus, we find that doubling teacher base pay had almost no impact on improving student test scores suggesting that the various posited mechanisms for why such a pay increase may have a positive impact on student learning (as described in Section 2) were not empirically salient in this setting.

5. Cost Effectiveness

Viewed as a program to increase test scores, the certification and salary increase is clearly quite costly. For instance, if we assume a uniform distribution of civil-service teachers between ages 30 and 60, the intensive-margin cost of a policy of doubling teacher pay across the board would be equal to 15 years of the annual teacher wage bill in Indonesia. Discounting at 5% (assuming conservatively that nominal wages increase with inflation, and not with growth rates), the present discounted cost would be over 10 years of the annual teacher wage bill. Since teacher salaries comprise over 10% of the annual Indonesian government budget, the present discounted intensive margin cost of the policy is more than 100% of the annual government budget. Of course, this figure does not entirely represent a social cost, because the salary increase mostly represents a transfer to teachers; the actual social cost would depend on the deadweight loss of raising tax revenue.

For this field experiment, the additional salary costs due to accelerated certification were about 66 US dollars per student in the treatment schools.[25] The cost of implementing the certification program should also be added to this figure, but we have too little information to make a credible estimate. Doing so would require assessing the time costs of teachers, assessors, and trainers--who have to prepare and assess portfolios and possibly attend training--as well as other administrative costs. But even without including those costs, it is clear that other salary-related interventions have been able to achieve substantial positive effects on learning much more cost-effectively. For instance, a multi-year experimental program providing performance-based incentive pay to teachers in India (Muralidharan and Sundararaman 2011) had additional yearly salary costs of only about 2 US dollars per student[26], yet it achieved student learning gains of 0.27σ and 0.17σ in math and language respectively. Over a 5-year period, the performance

[25] Costs were calculated by adding up impacts on monthly certification allowance in Y2 and Y3 (0.543+0.476=1.019mln IDR, Table 4, all teachers), multiplying times 12 and the average number of teachers (9.3, Table 1) and dividing by the average number of children in a school (190, Table 1), using a 9000 IDR/US dollar exchange rate from the duration of the experiment was 2009-2012.

[26] Incentive treatments cost up to Rupees 10,000 per school. Per student costs obtained by dividing by average student in school (113), and using an exchange rate of 44 Rupees to the dollar. Year of experiment 2005-2007.

pay experiment yielded gains of 0.54σ and 0.35σ in math and language for a cohort exposed to the performance-pay intervention for five years (Muralidharan 2012).

These calculations focus only on the intensive margin, benefits could as a result of higher quality professionals entering into the teaching profession. However, there are two considerations to keep in mind in weighing this extensive-margin argument. First, even if the policy led to an improvement in the quality of teachers entering the profession, there would still be a very large intensive margin cost of the policy. It is also worth noting that there is no evidence yet that new entrants in Indonesia are that much more effective than their predecessors. Certainly no effect has yet appeared in the performance of typical lower-secondary students: Indonesia's average PISA scores in math and science fell between 2006 and 2012, while reading scores were stagnant, and average TIMSS scores fell substantially between 2007 and 2011. Second, an alternative policy that connected at least some of the pay increases to performance is likely to be more effective on the extensive margin as well, since increasing the spread of worker pay to more closely reflect their productivity is likely to also be more effective at attracting higher-ability candidates than an across-the-board increase in salaries on a compressed schedule that is not linked to performance (Lazear 2000).

6. Discussion and Conclusion

This paper has offered new evidence on a key question in labor and personnel economics: How does a large, permanent, unconditional increase in salary affect employee job performance. Answering that question has important implications for personnel policy and practice, and it is also central to testing gift exchange and other efficiency wage models that posit that workers may exert effort above the norm in exchange for higher salaries. But while a large empirical literature has tried to test the gift-exchange model, that literature has had to rely on laboratory experiments and short-term field experiments using temporary researcher-generated variation in pay, as opposed to using permanent unconditional salary increases in the context of an existing long-term employment contract.

In this paper, we have provided experimental evidence on the intensive-margin impacts on teacher effort and student learning outcomes of a unique policy change in Indonesia that permanently doubled the base pay of eligible teachers who went through a certification process. Taking advantage of a ten-year phase-in of the policy, the experimental design allowed all eligible teachers in 120 randomly-selected primary and junior secondary public schools to immediately access the certification process and the resulting doubling of pay. The experiment thus created a sharp increase in the fraction of teachers in treated schools with a permanent doubling of pay during the three years of our study, which allows us to identify the intensive-margin impacts of an unconditional permanent increase in pay on performance in a near nationally representative sample of schools.

Given the challenges of implementing a randomized experiment at scale with a national government, the experiment worked remarkably well, with a strong "first stage". Beyond increasing teacher incomes, the experiment also substantially improved the intermediate variables through which policymakers hoped that the increase in salary would lead to better education quality: teachers in treated schools were significantly more likely to be satisfied with their income, significantly less likely to report financial stress, and significantly less likely to hold a second job than teachers in control schools.

Yet despite this improvement in teachers' pay and satisfaction, there was no impact on teacher effort towards upgrading their own skills, on teacher effort in the classroom, or on the ultimate outcome of student learning. Teachers in treated schools do not score better on tests of teacher knowledge of either subject matter or pedagogy, and do not report any increase in self-reported measures of effort such as teacher attendance, or the number of teaching hours. Most importantly, we find no difference in student test scores in language, mathematics, and science across treatment and control schools for either primary or junior secondary schools. The test score impact of being in a treated school is not only insignificant, but the point estimates are close to zero. The zero effects on learning are also quite precise, allowing us to rule out effects as small as 0.12σ at the 95% level in treated schools.

Moreover, when we restrict our analysis to students who were taught by target teachers in either the treatment or control schools at the start of the study, using school-level random assignment as an instrument for being taught by a certified teacher in a given year, we again find no effect of being taught by a teacher who had received a doubling of pay (relative to students in control schools taught by similar teachers). The point estimate is again close to zero, and we can rule out positive test score effects larger than 0.15σ at the 95% level.

While we measured outcomes over a three-year experiment and found no effects on the intensive margin, it is possible that large increases of teacher base pay could have effects on student learning in the even longer run. In theory they could improve learning on the extensive margin, if they led to an increase in the average quality of new applicants and entrants into the teaching profession. But given the ratio of new entrants to incumbents, any such extensive-margin effect would take many years to show significant effects on aggregate learning scores--and in fact, no improvement in students' average performance on international assessments is yet evident. Thus, policy makers hoping to increase the quality of government service delivery by increasing salaries across the board need to trade off these potential benefits on the extensive margin against the large intensive-margin costs of unconditional increases in public sector pay that may not yield any performance improvement.

References:

AKERLOF, G. A. (1982): "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, 97, 543-569.

ALLCOTT, H. (2015): "Site Selection Bias in Program Evaluation," *Quarterly Journal of Economics*, 130, 1117-1165.

ANDRABI, T., J. DAS, A. I. KHWAJA, and T. ZAJONC (2011): "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics," *American Economic Journal: Applied Economics*, 3 3, 29-54.

BALLOU, D., and M. PODGURSKY (1998): "The Case against Teacher Certification," *Public Interest*, 17-29.

BETTS, J. R. (1995): "Does School Quality Matter - Evidence from the National Longitudinal Survey of Youth," *Review of Economics and Statistics*, 77, 231-250.

CHANG, M. C., S. AL-SAMARRAI, A. B. RAGATZ, J. DE REE, S. SHAEFFER, and R. STEVENSON (2013): *Teacher Reform in Indonesia: The Role of Politics and Evidence in Policy Making*. Washington, DC: World Bank.

DAL BÓ, E., F. FINAN, and M. A. ROSSI (2013): "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service*," *The Quarterly Journal of Economics*, 128, 1169-1218.

DOLTON, P., O. D. MARCENARO-GUTIERREZ, L. PISTAFERRI, and Y. ALGAN (2011): "If You Pay Peanuts Do You Get Monkeys? A Cross-Country Analysis of Teacher Pay and Pupil Performance," *Economic Policy*, 5-55.

ESTEVES-SORENSON, C., and R. MACERA (2013): "Revisiting Gift Exchange: Theoretical Considerations and a Field Test," working paper.

FALK, A. (2007): "Gift Exchange in the Field," *Econometrica*, 75, 1501-1511.

FEHR, E., S. GACHTER, and G. KIRCHSTEIGER (1997): "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica*, 65, 833-860.

FEHR, E., G. KIRCHSTEIGER, and A. RIEDL (1993): "Does Fairness Prevent Market Clearing - an Experimental Investigation," *Quarterly Journal of Economics*, 108, 437-459.

FERRAZ, C., and F. FINAN (2011): "Motivating Politicians: The Impacts of Monetary Incentives on Quality and Performance," UC Berkeley.

GNEEZY, U., and J. A. LIST (2006): "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments," *Econometrica*, 74, 1365-1384.

GROGGER, J. (1996): "School Expenditures and Post-Schooling Earnings: Evidence from High School and Beyond," *Review of Economics and Statistics*, 78, 628-637.

HANUSHEK, E. A. (1986): "The Economics of Schooling - Production and Efficiency in Public-Schools," *Journal of Economic Literature*, 24, 1141-1177.

HECKMAN, J. J., and J. A. SMITH (1995): "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, 85-110.

LAZEAR, E. (2000): "Performance Pay and Productivity," *American Economic Review*, 90, 1346-61.

LEVITT, S. D., and J. A. LIST (2007): "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives*, 21, 153-174.

MULLIS, I. V., M. O. MARTIN, P. FOY, and A. ARORA (2012): *Timss 2011 International Results in Mathematics*.

MURALIDHARAN, K. (2012): "Long-Term Effects of Teacher Performance Pay," UC San Diego.

MURALIDHARAN, K., and V. SUNDARARAMAN (2011): "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, 119, 39-77.

OECD (2013): "Pisa 2012 Results in Focus: What 15-Year-Olds Know and What They Can Do with What They Know."

SHAPIRO, C., and J. E. STIGLITZ (1984): "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, 74, 433-444.

SURYAHADI, A., and P. SAMBODHO (2013): "An Assessment of Policies to Improve Teacher Quality and Reduce Teacher Absenteeism," in *Education in Indonesia*, ed. by D. Suryadarma, and G. W. Jones. Singapore:: Institute of Southeast Asian Studies.

UNESCO (2014): "Teaching and Learning: Achieving Quality for All," Paris.

UNICEF (2011): "Teachers: A Regional Study on Recruitment, Development and Salaries of Teachers in the Ceecis Region," Geneva: UNICEF Regional Office for CEECIS.

WORLD BANK (2010): *Transforming Indonesia's Teaching Force*. Jakarta: Human Development Department, World Bank East Asia and Pacific Region.

Table 1: Balance on School and Student level Variables

	[1]	[2]	[3]
	Panel A: Balance on School level variables		
	Treatment	Control	Difference
Number of classes per school	8.892	8.321	0.571
_	(4.883)	(4.485)	[0.517]
Number of students per school	190.850	184.492	6.358
_	(133.797)	(135.322)	[15.073]
Class size	20.598	20.991	-0.394
_	(6.764)	(7.156)	[0.786]
Number of teachers per school	9.350	9.075	0.275
_	(5.198)	(4.591)	[0.537]
	Panel B: Balance on Student level variables		
	Treatment	Control	Difference
Raw math score (fraction correct)	0.408	0.405	0.004
_	(0.229)	(0.232)	[0.020]
Raw science score	0.512	0.515	-0.003
_	(0.214)	(0.210)	[0.015]
Raw Indonesian score	0.584	0.585	-0.002
_	(0.206)	(0.205)	[0.013]
Raw English score	0.398	0.391	0.007
_	(0.176)	(0.172)	[0.023]
Student assets index	0.555	0.540	0.015
_	(0.233)	(0.229)	[0.019]

Notes:
* p<0.1; ** p<0.05; *** p<0.01. Table compares average values between treatment and control schools. Standard errors are clustered at the school level. Standard deviation values reported in parenthesis. Standard error of the estimated difference between treatment and control is reported in square brackets.

Table 2: Balance on teacher level variables

	[1]	[2]	[3]	[4]	[5]	[6]
	ALL teachers			Target teachers only		
Fraction of teachers tested	0.876	0.861	0.015	0.906	0.878	0.028
	(0.330)	(0.346)	[0.018]	(0.292)	(0.327)	[0.019]
Fraction of teachers interviewed	0.940	0.937	0.003	1.000	1.000	0.000
	(0.238)	(0.244)	[0.014]	(0.000)	(0.000)	[.]
Raw test score (fraction correct)	0.556	0.556	-0.000	0.562	0.568	-0.006
	(0.165)	(0.163)	[0.014]	(0.167)	(0.166)	[0.015]
Fraction "target" at Y0	0.495	0.517	-0.022	1.000	1.000	0.000
	(0.500)	(0.500)	[0.026]	(0.000)	(0.000)	[.]
Fraction already certified at Y0	0.194	0.181	0.012	0.000	0.000	0.000
	(0.395)	(0.385)	[0.022]	(0.000)	(0.000)	[.]
Fraction not eligible for certification at Y0	0.311	0.302	0.009	0.000	0.000	0.000
	(0.463)	(0.459)	[0.032]	(0.000)	(0.000)	[.]
Fraction with bachelor's degree	0.619	0.590	0.029	0.661	0.620	0.041
	(0.486)	(0.492)	[0.041]	(0.474)	(0.485)	[0.045]
Fraction who started or completed the certification	0.606	0.288	0.318***	0.757	0.200	0.557***
process	(0.489)	(0.453)	[0.034]	(0.429)	(0.400)	[0.033]
Fraction certified	0.194	0.181	0.012	0.000	0.000	0.000
	(0.395)	(0.385)	[0.022]	(0.000)	(0.000)	[.]
Fraction certified and paid the certification	0.115	0.123	-0.009	0.000	0.000	0.000
allowance	(0.319)	(0.329)	[0.017]	(0.000)	(0.000)	[0.000]
Base pay (in MIL IDR)	1.873	1.921	-0.048	2.243	2.247	-0.004
_	(0.830)	(0.798)	[0.058]	(0.421)	(0.421)	[0.036]
Allowances other than certification allowance (in	0.495	0.505	-0.010	0.616	0.646	-0.030*
MIL IDR)	(0.356)	(0.349)	[0.020]	(0.255)	(0.256)	[0.017]
Certification pay (in MIL IDR)	0.210	0.220	-0.010	0.000	0.000	0.000
_	(0.593)	(0.602)	0.030	(0.000)	(0.000)	0.000
Fraction with a second job	0.336	0.336	0.001	0.298	0.315	-0.018
_	(0.473)	(0.472)	0.027	(0.458)	(0.465)	0.030

Notes:

* p<0.1; ** p<0.05; *** p<0.01. Table compares average values between treatment and control schools. Standard errors are clustered at the school level. Standard deviation values reported in parenthesis. Standard error of the estimated difference between treatment and control is reported in square brackets.

Table 3: First stage process -- teacher level

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
		Y0			*Y2*			*Y3*	
					Panel A: All teachers				
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Fraction who had entered or completed the certification process	0.606	0.288	0.318***	0.648	0.480	0.167***	0.713	0.638	0.075**
	(0.489)	(0.453)	[0.034]	(0.478)	(0.500)	[0.034]	(0.452)	(0.481)	[0.032]
Fraction of certified teachers	0.194	0.181	0.012	0.612	0.382	0.230***	0.647	0.505	0.142***
	(0.395)	(0.385)	[0.022]	(0.487)	(0.486)	[0.035]	(0.478)	(0.500)	[0.036]
Fraction of certified teachers who have been paid the certification allowance	0.115	0.123	-0.009	0.558	0.269	0.290***	0.599	0.374	0.225***
	(0.319)	(0.329)	[0.017]	(0.497)	(0.443)	[0.034]	(0.490)	(0.484)	[0.036]
					Panel B: Target teachers only				
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Fraction who had entered or completed the certification process	0.757	0.200	0.557***	0.893	0.576	0.317***	0.949	0.866	0.084***
	(0.429)	(0.400)	[0.033]	(0.309)	(0.494)	[0.033]	(0.219)	(0.341)	[0.024]
Fraction of certified teachers	0.000	0.000	0.000	0.855	0.407	0.448***	0.904	0.654	0.250***
	(0.000)	(0.000)	[.]	(0.352)	(0.491)	[0.033]	(0.295)	(0.476)	[0.033]
Fraction of certified teachers who have been paid the certification allowance	0.000	0.000	0.000	0.760	0.191	0.569***	0.856	0.414	0.442***
	(0.000)	(0.000)	[.]	(0.427)	(0.393)	[0.033]	(0.351)	(0.493)	[0.034]
					Panel C: Teachers who are already certified at Y0				
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Fraction who had entered or completed the certification process	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.993	0.007
	(0.000)	(0.000)	[0.000]	(0.000)	(0.000)	[0.000]	(0.000)	(0.083)	[0.005]
Fraction of certified teachers	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.980	0.020**
	(0.000)	(0.000)	[.]	(0.000)	(0.000)	[.]	(0.000)	(0.141)	[0.009]
Fraction of certified teachers who have been paid the certification allowance	0.591	0.680	-0.089	1.000	1.000	0.000	1.000	0.976	0.024**
	(0.493)	(0.467)	[0.064]	(0.000)	(0.000)	[.]	(0.000)	(0.152)	[0.010]
					Panel D: Teachers who are not eligible at Y0				
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Fraction who had entered or completed the certification process	0.120	0.010	0.110***	0.154	0.068	0.086***	0.279	0.180	0.100**
	(0.325)	(0.098)	[0.025]	(0.361)	(0.252)	[0.028]	(0.449)	(0.384)	[0.041]
Fraction of certified teachers	0.000	0.000	0.000	0.103	0.024	0.079***	0.169	0.080	0.089***
	(0.000)	(0.000)	[.]	(0.305)	(0.154)	[0.025]	(0.375)	(0.271)	[0.032]
Fraction of certified teachers who have been paid the certification allowance	0.000	0.000	0.000	0.091	0.010	0.080***	0.112	0.027	0.085***
	(0.000)	(0.000)	[.]	(0.288)	(0.102)	[0.023]	(0.316)	(0.163)	[0.026]

Notes:

* p<0.1; ** p<0.05; *** p<0.01. Table compares average values between treatment and control schools across different subpopulations of teachers and across the periods of measurement Y0 (November 2009), Y2 (April 2011), and Y3 (April 2012). Standard errors allow for dependence within schools. Standard errors are clustered at the school level. Standard deviation values reported in parenthesis. Standard error of the estimated difference between treatment and control is reported in square brackets.

Table 4: Teacher level impact

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
	ALL teachers						**Target teachers only**					
	Y2			Y3			Y2			Y3		
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Standardized test scores	0.033	0.007	0.025	-0.034	0.007	-0.041	0.043	0.028	0.014	-0.011	0.060	-0.071
	(1.057)	(0.991)	[0.083]	(1.071)	(0.988)	[0.088]	(1.063)	(0.963)	[0.096]	(1.071)	(0.984)	[0.100]
Fraction with a bachelor's degree	0.713	0.677	0.036	0.778	0.730	0.048	0.771	0.726	0.044	0.799	0.752	0.047
	(0.453)	(0.468)	[0.034]	(0.416)	(0.444)	[0.029]	(0.421)	(0.446)	[0.036]	(0.401)	(0.432)	[0.035]
Fraction pursuing further education	0.178	0.184	-0.006	0.140	0.159	-0.019	0.101	0.086	0.014	0.097	0.079	0.018
	(0.383)	(0.388)	[0.022]	(0.347)	(0.366)	[0.021]	(0.301)	(0.281)	[0.020]	(0.296)	(0.269)	[0.022]
Fraction with a second job (self reported)	0.264	0.322	-0.058***	0.218	0.266	-0.048*	0.243	0.305	-0.062**	0.183	0.244	-0.061**
	(0.441)	(0.467)	[0.021]	(0.413)	(0.442)	[0.026]	(0.430)	(0.461)	[0.027]	(0.387)	(0.430)	[0.031]
Teaching hours per week	23.361	22.801	0.560	23.529	22.961	0.568	24.066	23.251	0.816	24.059	23.540	0.520
	(6.304)	(6.523)	[0.492]	(5.631)	(5.979)	[0.442]	(4.776)	(5.887)	[0.599]	(4.030)	(5.272)	[0.532]
Base pay (in MIL IDR)	2.021	2.083	-0.062	2.570	2.592	-0.022	2.469	2.511	-0.042	2.797	2.809	-0.013
	(0.944)	(0.935)	[0.059]	(0.794)	(0.741)	[0.049]	(0.460)	(0.460)	[0.039]	(0.494)	(0.484)	[0.042]
Allowances other than certification allowance (in MIL IDR)	0.622	0.592	0.029	0.511	0.545	-0.034	0.849	0.828	0.021	0.661	0.688	-0.027
	(0.791)	(0.731)	[0.079]	(0.509)	(0.629)	[0.028]	(0.815)	(0.805)	[0.120]	(0.537)	(0.511)	[0.039]
Certification allowance (in MIL IDR)	1.111	0.567	0.543***	1.354	0.878	0.476***	1.498	0.409	1.089***	1.961	0.960	1.001***
	(1.030)	(0.969)	[0.066]	(1.257)	(1.235)	[0.081]	(0.911)	(0.861)	[0.067]	(1.018)	(1.229)	[0.089]
Financial problems (self reported)	0.404	0.495	-0.091***	0.468	0.557	-0.089***	0.302	0.448	-0.146***	0.296	0.475	-0.179***
	(0.491)	(0.500)	[0.028]	(0.499)	(0.497)	[0.033]	(0.460)	(0.498)	[0.030]	(0.457)	(0.500)	[0.035]
Satisfied with total income (self reported)	0.691	0.604	0.087***	0.666	0.596	0.070**	0.821	0.641	0.179***	0.856	0.692	0.164***
	(0.462)	(0.489)	[0.024]	(0.472)	(0.491)	[0.031]	(0.384)	(0.480)	[0.030]	(0.351)	(0.462)	[0.029]
Absent from school at least once in the past week (self reported)	0.134	0.135	-0.001	0.125	0.126	-0.000	0.109	0.120	-0.011	0.119	0.098	0.021
	(0.341)	(0.342)	[0.019]	(0.331)	(0.331)	[0.019]	(0.312)	(0.325)	[0.023]	(0.324)	(0.298)	[0.025]

Notes:

* p<0.1; ** p<0.05; *** p<0.01. Table compares average values between treatment and control schools for ALL teachers (column [1]-[6]) and for target teachers only (columns [7]-[12]) and evaluates these differences separately for the two moments of measurement Y2 (April 2011) and Y3 (April 2012). Standard errors allow for dependence within schools. Standard deviation values reported in parenthesis. Standard error of the estimated difference between treatment and control is reported in square brackets.

Table 5: Intent to treat effects on student test scores

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]
	ALL school types					Primary school only				Junior secondary school only				
	Panel A: Student test score data measured at Y2													
	Math	Science	Indonesian	English	**Pooled**	Math	Science	Indonesian	**Pooled**	Math	Science	Indonesian	English	**Pooled**
Treatment School	0.010	-0.002	-0.017	0.022	**-0.000**	0.025	-0.003	0.009	**0.011**	0.001	-0.003	-0.040	0.022	**-0.005**
	[0.053]	[0.044]	[0.038]	[0.088]	**[0.048]**	[0.042]	[0.043]	[0.045]	**[0.042]**	[0.093]	[0.071]	[0.053]	[0.088]	**[0.074]**
	Panel B: Student test score data measured at Y3													
	Math	Science	Indonesian	English	**Pooled**	Math	Science	Indonesian	**Pooled**	Math	Science	Indonesian	English	**Pooled**
Treatment School	0.034	0.033	0.004	0.028	**0.024**	0.028	0.013	0.019	**0.020**	0.051	0.060	-0.005	0.028	**0.033**
	[0.053]	[0.044]	[0.039]	[0.087]	**[0.048]**	[0.041]	[0.042]	[0.042]	**[0.040]**	[0.093]	[0.070]	[0.060]	[0.087]	**[0.074]**

Notes:

* p<0.1; ** p<0.05; *** p<0.01. Table reports Intent to treat effects. Outcome test scores are standardized so that the mean and standard deviation is 0 and 1 in the control group. The outcome score is then regressed on a dummy variable indicating a treatment school, a full set of 20 district dummy variables, and a standardized Y0 test score, and a dummy variable indicating observations for which the Y0 score is not observed. The standardized Y0 test score is set to 0 for observations for which the Y0 test score is not observed. The parameter on the dummy variable indicating a treatment school is reported in the table as the intent to treat effect. Panel A reports results based on Y2 test score data and panel B reports results based on Y3 test score data. Standard errors allow for dependence within schools. Standard errors are reported in squared brackets.

Table 6: Intent to treat effects on student test scores -- breakdown by "target status"

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]
	ALL school types					primary school only					junior secondary school only			
	Panel A: student test score data measured at Y2 (April 2011)													
	Math	Science	Indonesian	English	Pooled	Math	Science	Indonesian	Pooled	Math	Science	Indonesian	English	Pooled
Target * Treatment	0.023	-0.023	-0.016	0.016	-0.004	0.019	-0.029	-0.016	-0.009	0.022	0.002	-0.013	0.016	0.005
	[0.064]	[0.052]	[0.048]	[0.083]	[0.052]	[0.047]	[0.051]	[0.049]	[0.045]	[0.112]	[0.071]	[0.069]	[0.083]	[0.075]
Non-target * Treatment	-0.007	0.022	-0.005	0.083	0.012	0.042	0.025	0.045	0.038	-0.027	-0.025	-0.066	0.083	-0.005
	[0.055]	[0.054]	[0.044]	[0.111]	[0.051]	[0.056]	[0.056]	[0.060]	[0.054]	[0.085]	[0.096]	[0.052]	[0.111]	[0.077]
test all causal parameters are zero (p-value)	0.568	0.422	0.836	0.452	0.668	0.717	0.410	0.334	0.401	0.510	0.752	0.414	0.452	0.776
	Panel B: student test score data measured at Y3 (April 2012)													
	Math	Science	Indonesian	English	Pooled	Math	Science	Indonesian	Pooled	Math	Science	Indonesian	English	Pooled
Target-Target * Treatment	0.016	-0.017	-0.014	-0.023	-0.018	-0.007	-0.041	-0.011	-0.019	-0.019	0.030	-0.034	-0.023	-0.025
	[0.074]	[0.058]	[0.059]	[0.093]	[0.058]	[0.069]	[0.062]	[0.065]	[0.057]	[0.124]	[0.077]	[0.082]	[0.093]	[0.076]
Target-Nontarget * Treatment	0.026	-0.009	0.057	0.176	0.043	0.059	0.040	0.144**	0.080	0.004	-0.135	-0.044	0.176	0.010
	[0.075]	[0.064]	[0.064]	[0.115]	[0.056]	[0.064]	[0.069]	[0.065]	[0.058]	[0.159]	[0.123]	[0.108]	[0.115]	[0.094]
Non-target-Target * Treatment	0.084	0.056	0.014	0.088	0.061	0.063	-0.028	-0.019	0.006	0.095	0.088	0.063	0.088	0.092
	[0.090]	[0.065]	[0.061]	[0.119]	[0.069]	[0.069]	[0.065]	[0.055]	[0.057]	[0.137]	[0.078]	[0.092]	[0.119]	[0.098]
Non-target - Non-target * Treatment	0.028	0.099*	0.040	-0.007	0.046	0.041	0.070	0.053	0.055	0.062	0.139	0.018	-0.007	0.048
	[0.058]	[0.057]	[0.055]	[0.079]	[0.051]	[0.059]	[0.065]	[0.069]	[0.061]	[0.092]	[0.094]	[0.085]	[0.079]	[0.071]
test all causal parameters are zero (p-value)	0.825	0.161	0.755	0.191	0.334	0.816	0.557	0.125	0.537	0.812	0.064*	0.767	0.191	0.223

Notes:

* p<0.1; ** p<0.05; *** p<0.01. Table reports parameter estimates (equation XX in the main text), where estimated constants are suppressed. The results effectively are intent to treat effects, broken down by type of teacher. The first row in panel B -- target-target -- for example measures the difference in learning outcomes between treatment and control for the subpopulation of students who had a target teacher in Y2 AND in Y3. These are the students most (differentially) affected by our intervention. Outcome test scores are standardized so that the mean and standard deviation is 0 and 1 in the control group. The outcome score is then regressed on a dummy variable indicating a treatment school, a full set of 20 district dummy variables, and a standardized Y0 test score, and a dummy variable indicating observations for which the Y0 score is not observed. The standardized Y0 test score is set to 0 for observations for which the Y0 test score is not observed. The parameter on the dummy variable indicating a treatment school is reported in the table as the intent to treat effect. Panel A reports results based on Y2 test score data and panel B reports results based on Y3 test score data. Standard errors allow for dependence within schools. Standard errors are reported in squared brackets.

Table 7: IV results measuring the causal impact on annual test score gains of being taught by a "certified and paid" teacher

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
	intent to treat estimate on subsample	IV estimates										
Persistence parameter		*0*	*0.1*	*0.2*	*0.3*	***0.4***	***0.5***	***0.6***	*0.7*	*0.8*	*0.9*	*1*
Causal Impact of a Year of being Taught by a "Certified" Teacher	-0.010	-0.019	-0.019	-0.018	-0.018	-0.017	-0.017	-0.016	-0.016	-0.015	-0.015	-0.015
	[0.049]	[0.095]	[0.093]	[0.090]	[0.087]	[0.085]	[0.083]	[0.081]	[0.079]	[0.077]	[0.075]	[0.073]
Number of clusters	320	320	320	320	320	320	320	320	320	320	320	320
Number of observations	172,615	172,615	172,615	172,615	172,615	172,615	172,615	172,615	172,615	172,615	172,615	172,615
effects larger than these are statistically rejected		0.167	0.163	0.158	0.153	**0.150**	**0.146**	**0.143**	0.139	0.136	0.132	0.128

Notes:
* p<0.1; ** p<0.05; *** p<0.01. Columns [2]-[12] report estimates of the parameter beta2 in equation (3) in the main text. The parameter is an estimate of the effect of (approximately) doubling teachers' pay on a year of learning. The estimates depend on fixing the persistence parameter at a value between 0 (column [2]) and 1 (column [12]). Equation (3) is estimated on a subsample of the data. First, Y2 and Y3 outcome data are pooled. Second, only observations which were taught by a target teacher in the last year before measurement (for Y2 data) and the last two years before measurement (for Y3 data) are used in the analysis. For this subsample of the data we have the strongest first stage. Standard errors allow for dependence within schools. Standard errors are reported in squared brackets. The bottom row reports whichever effects are statistically rejected, the value is calculated by adding 1.96 times the standard error to the point estimate.

Figure 1: map of the 20 selected districts in Indonesia

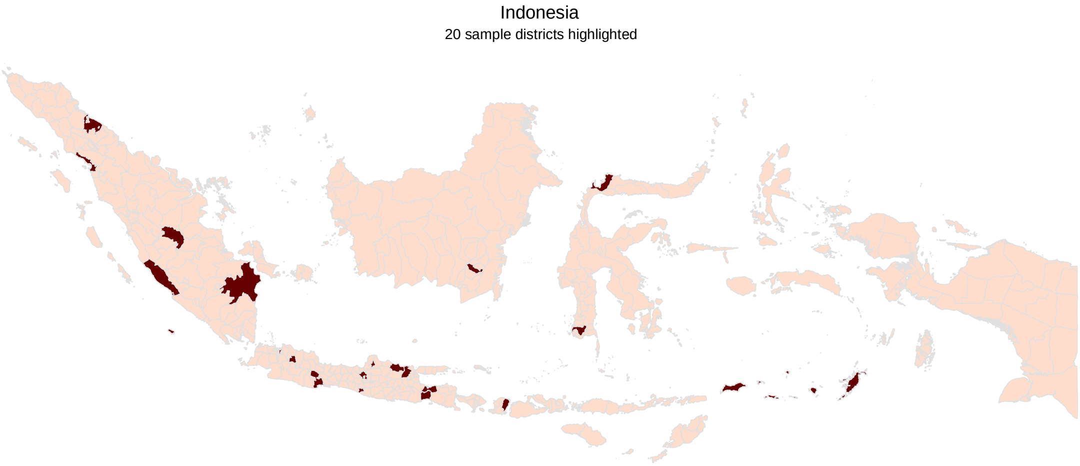


Figure 2: Time line

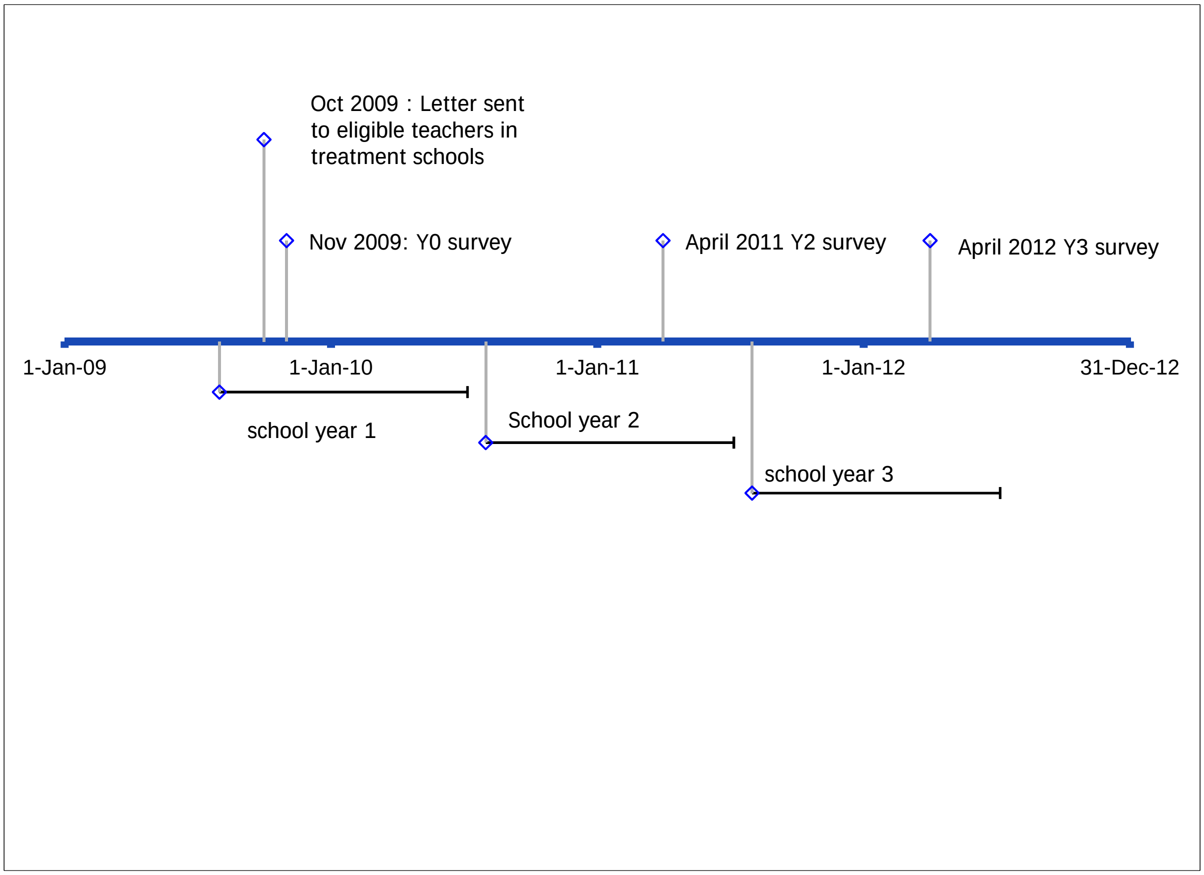
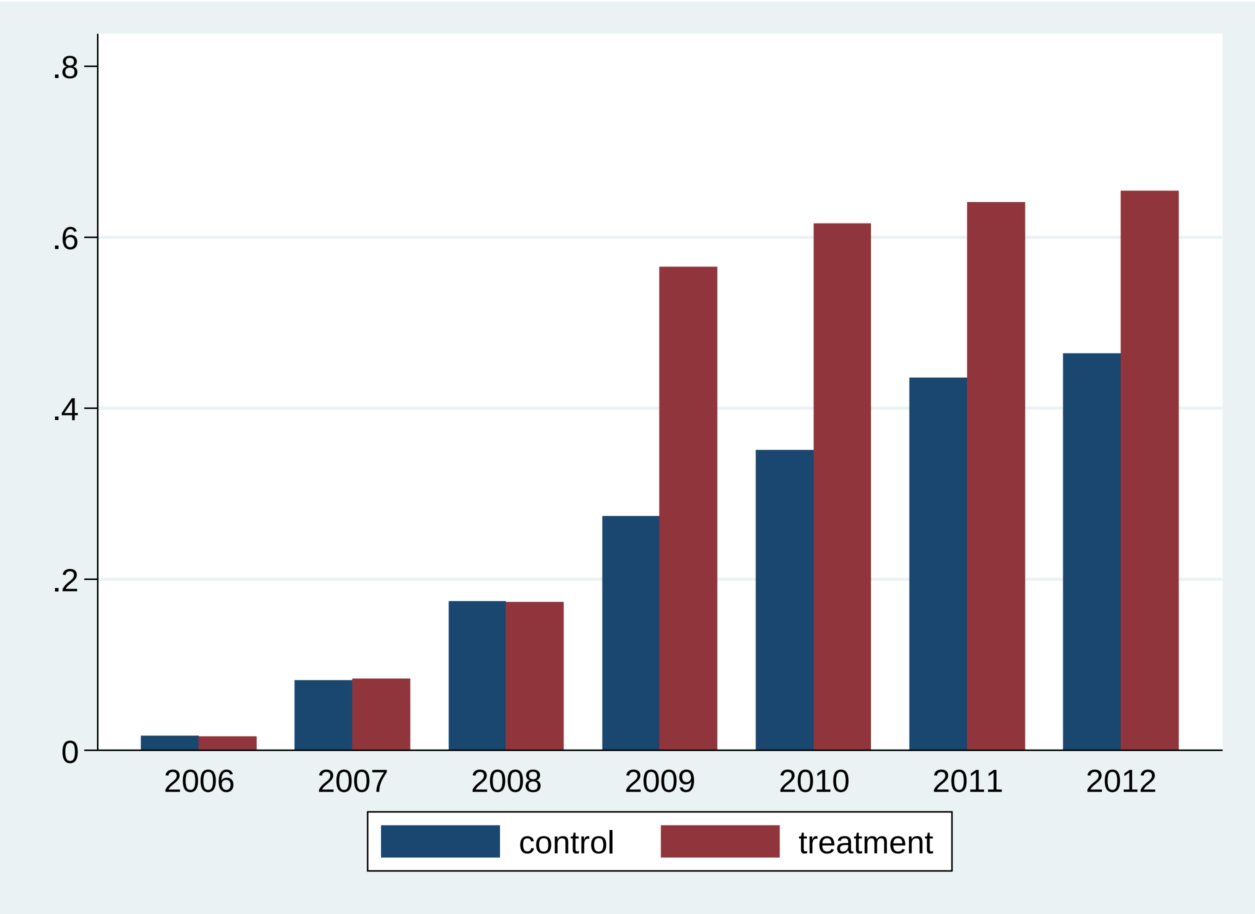
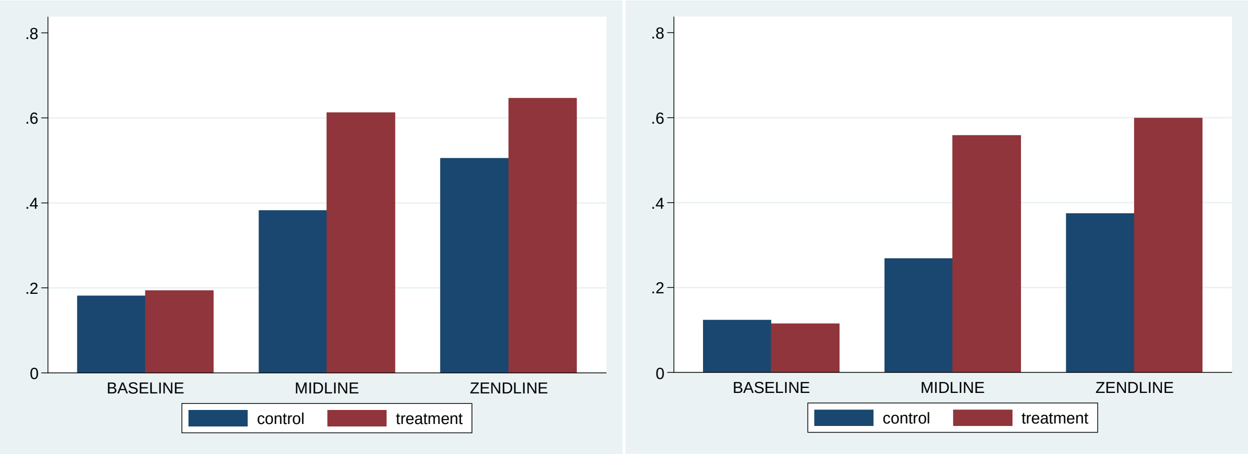


Figure 3: admitted to the certification process (including those who are already certified)



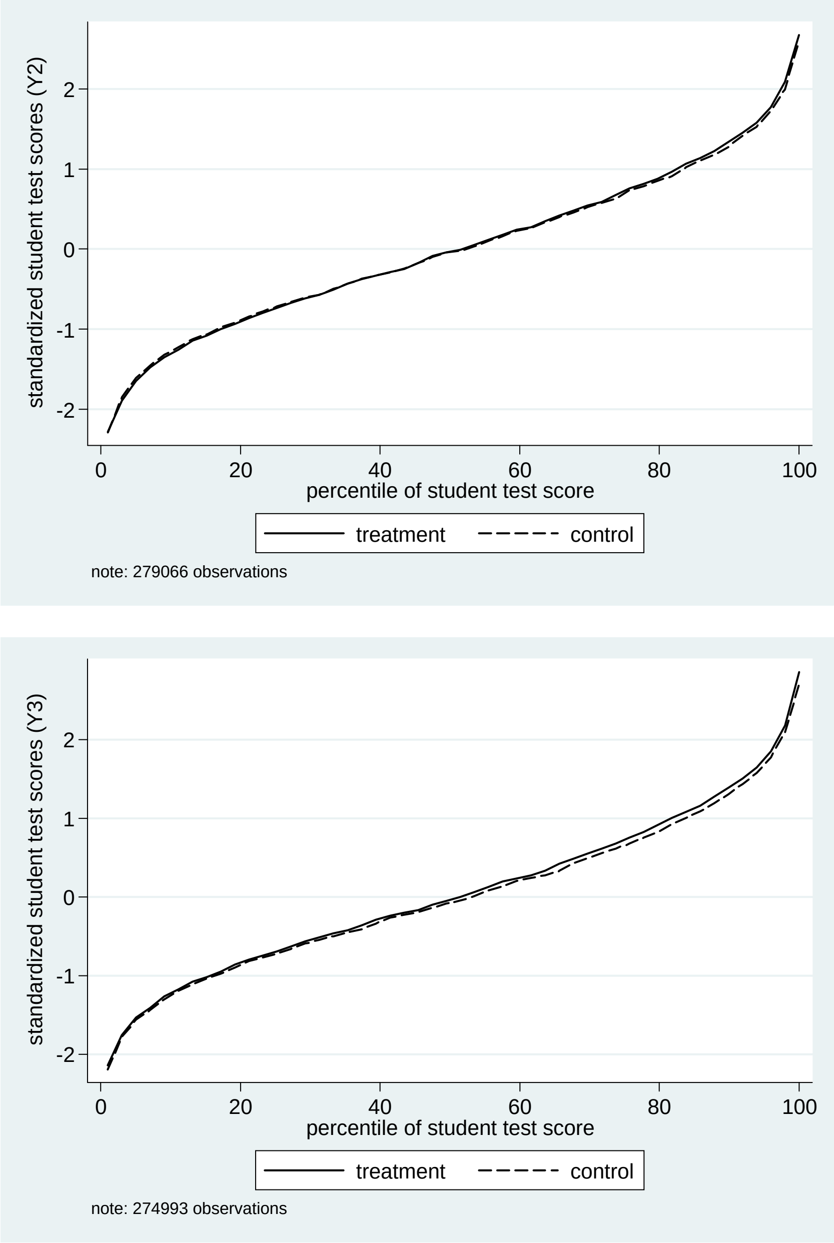
Notes: Teachers have been admitted to the certification process at different times. The first batch of teachers was admitted in 2006. In 2009, the intervention took place, which created a difference between treatment and control schools in terms of the fraction of teachers who were admitted to the certification process.

Figure 4: Completing the certification process and being paid the certification allowance



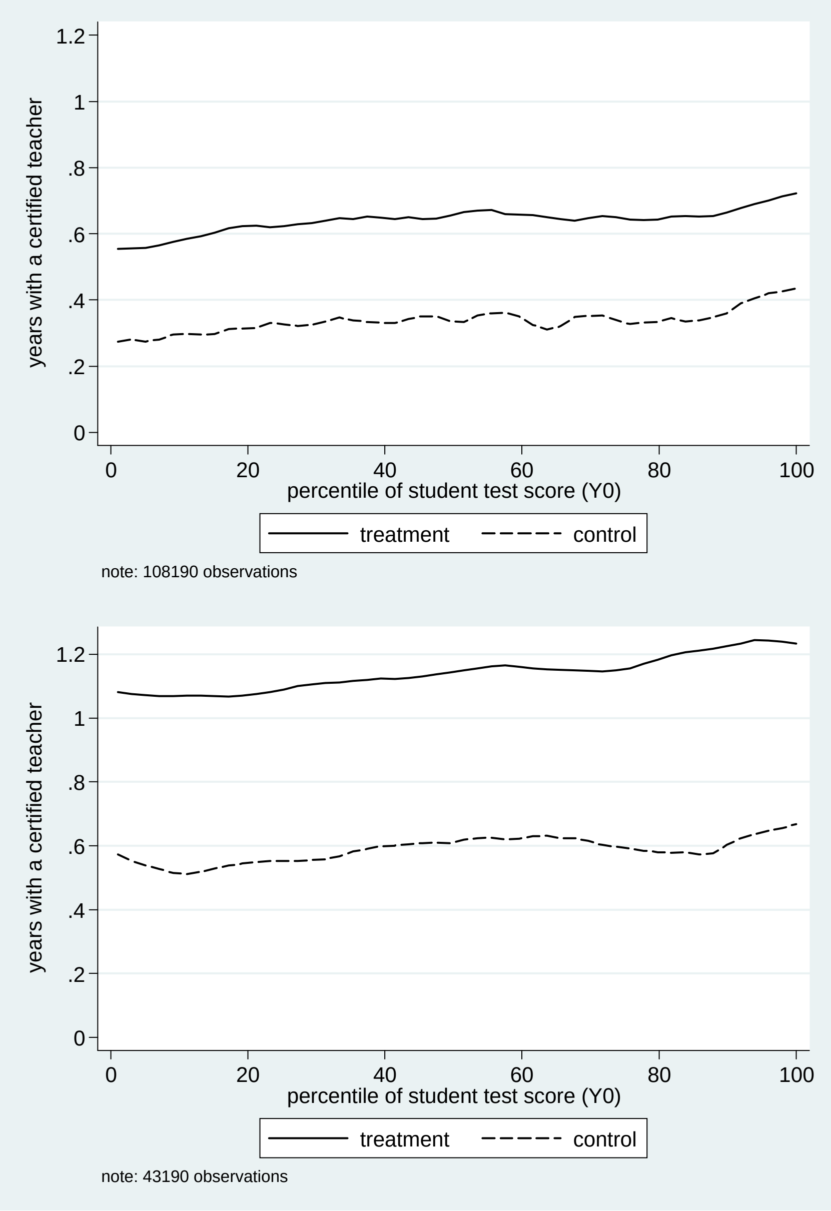
Notes: The left panel presents the fraction of teachers who completed the certification process. The right panel presents the fraction of teachers who completed the certification process and were paid the certification allowance.

Figure 5: Quantile treatment effects (Y2 and Y3)



Notes: Quantile treatment effects for Y2 observations (top panel) and Y3 observations (bottom panel).

Figure 6A: Quantile treatment effects as a function of Y0 test scores – first stage

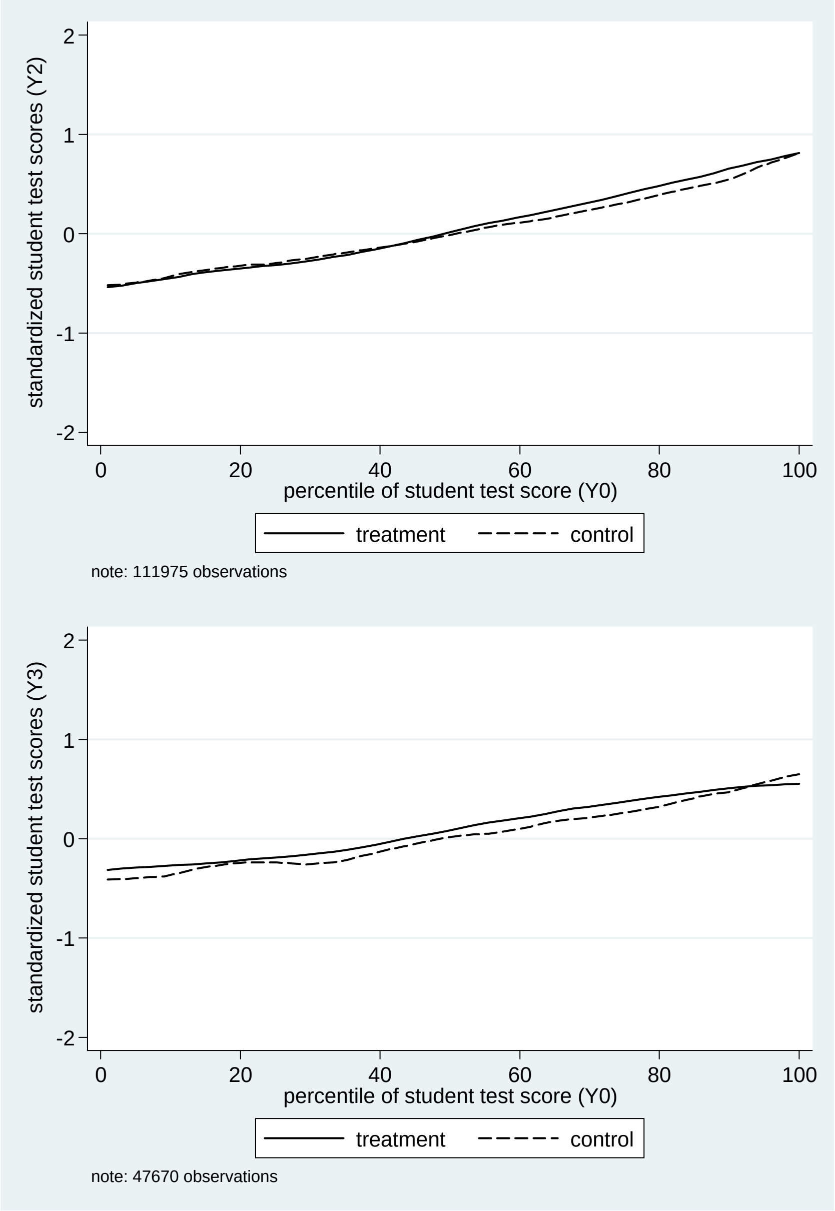


note: 108190 observations

note: 43190 observations

Notes: Figures present differential exposure to certified teachers (measured as the number of full school years with a certified teacher since Y0), as a function of percentiles of Y0 test scores. The top panel measured the differential for Y2 data, and the bottom panel for Y3 data.

Figure 6B: Quantile treatment effects as a function of Y0 test scores



Notes: Quantile treatment effects as a function of percentiles of Y0 test scores, for Y2 observations (top panel) and Y3 observations (bottom panel). The restriction on the availability of Y0 test score is especially important for Y3 data.

Table A.1: strata and sampled districts

stratum	*district 1*	*district 2*	*district 3*	*district 4*
Eastern Indonesia (Maluku and Papua)	MALUKU TENGGARA BARAT			
Nusa Tenggara	LOMBOK TIMUR			
Western Java	CIAMIS	JAKARTA TIMUR	PURWAKARTA	
Central Java	BANTUL	KUDUS	SEMARANG	
Eastern Java + Bali	LAMONGAN	LUMAJANG	PROBOLINGGO	TUBAN
Kalimantan	HULU SUNGAI SELATAN			
Sulawesi	GOWA	TOLI TOLI		
Northern Sumatra	DELI SERDANG	TAPANULI TENGAH		
Western Sumatra	TEBO			
Southern Sumatra	BENGKULU UTARA	OGAN ILIR		

Notes:

Regions (the strata) are approximate descriptions. Western Java, for example, includes the provinces West Java, Jakarta and Banten, all three located on the western side of the island of Java

Table A.2: Balance on teacher level variables

	[1]	[2]	[3]	[4]	[5]	[6]
	Already certified teachers at Y0			**Not eligible for certification at Y0**		
Fraction of teachers tested	0.851	0.847	0.004	0.829	0.838	-0.009
	(0.357)	(0.361)	[0.030]	(0.377)	(0.368)	[0.040]
Fraction of teachers interviewed	1.000	1.000	0.000	1.000	1.000	0.000
	(0.000)	(0.000)	[0.000]	(0.000)	(0.000)	[0.000]
Raw test score (fraction correct)	0.614	0.596	0.018	0.514	0.520	-0.005
	(0.155)	(0.162)	[0.019]	(0.166)	(0.152)	[0.018]
Fraction "target" at Y0	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	[0.000]	(0.000)	(0.000)	[0.000]
Fraction already certified at Y0	1.000	1.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	[0.000]	(0.000)	(0.000)	[0.000]
Fraction not eligible for certification at Y0	0.000	0.000	0.000	1.000	1.000	0.000
	(0.000)	(0.000)	[0.000]	(0.000)	(0.000)	[0.000]
Fraction with bachelor's degree	0.976	1.000	-0.024**	0.329	0.292	0.037
	(0.154)	(0.000)	[0.010]	(0.471)	(0.455)	[0.052]
Fraction who started or completed the certification process	1.000	1.000	0.000	0.120	0.010	0.110***
	(0.000)	(0.000)	[0.000]	(0.325)	(0.098)	[0.025]
Fraction certified	1.000	1.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	[0.000]	(0.000)	(0.000)	[0.000]
Fraction certified and paid the certification allowance	0.591	0.680	-0.089	0.000	0.000	0.000
	(0.493)	(0.467)	[0.064]	(0.000)	(0.000)	[0.000]
Base pay (in MIL IDR)	2.391	2.406	-0.015	0.968	1.052	-0.085
_	(0.355)	(0.367)	[0.050]	(0.804)	(0.818)	[0.087]
Allowances other than certification allowance (in MIL IDR)	0.721	0.713	0.009	0.265	0.253	0.012
	(0.242)	(0.246)	[0.023]	(0.366)	(0.318)	[0.038]
Certification pay (in MIL IDR)	1.085	1.229	-0.144	0.000	0.000	0.000
_	(0.930)	(0.886)	[0.117]	(0.000)	(0.000)	[0.000]
	0.308	0.280	0.028	0.416	0.405	0.011
_	(0.463)	(0.449)	[0.056]	(0.494)	(0.491)	[0.042]

Notes:

* p<0.1; ** p<0.05; *** p<0.01. Table compares average values between treatment and control schools. Standard errors allow for dependence within schools. Standard deviation values reported in parenthesis. Standard error of the estimated difference between treatment and control is reported in squared brackets.

Table A.3: Estimation sample

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
cohort	grade level observed in Y0	grade level observed in Y2	grade level observed in Y3	cohort used in ITT estimation on the Y2 sample	cohort used in ITT estimation on the Y3 sample	Y0 values available at Y2	Y0 values available at Y3
P1			grade 1	.	1	.	0
P2		grade 1	grade 2	1	1	0	0
P3		grade 2	grade 3	1	1	0	0
P4	grade 2	grade 3	grade 4	1	1	1	1
P5	grade 3	grade 4	grade 5	1	1	1	1
P6	grade 4	grade 5	grade 6	1	1	1	1
P7	grade 5	grade 6		1	.	1	.
P8	grade 6			.	.	.	.
S1			grade 7	.	1	.	0
S2		grade 7	grade 8	1	1	0	0
S3		grade 8	grade 9	1	1	0	0
S4	grade 8	grade 9		1	.	1	.
S5	grade 9			.	.	.	.

Notes:

"1": yes, "0": no, ".": Does Not Apply. The table shows, by cohort, in which grades we observe them throughout the period of measurement (columns [1]-[3]), in which types of analysis we use their test score data (columns [3]-[4]), and whether Y0 test scores are available for the respective cohorts when we observe them in period Y2 and Y3 respectively (columns [6]-[7]). The cohorts P1-P8 are the primary school cohorts and the cohort S1-S5 are the secondary school cohorts in our data.

Table A.4: Testing for differential attrition

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Y2 data, selection on cohort P4, P5, P6, P7, S4			Y2 data, selection on cohort P4, P5, P6			Y3 data, selection on cohort P4, P5, P6		
	Panel A: pooled								
	treatment	control	difference	treatment	control	difference	treatment	control	difference
Fraction of Y0 observations staying in the sample	0.854	0.845	0.009	0.885	0.877	0.008	0.841	0.826	0.016
			[0.024]			[0.011]			[0.032]
	Panel B: breakdown by high and low scoring students								
	treatment	control	difference	treatment	control	difference	treatment	control	difference
Fraction of Y0 observations staying in the sample (Y0 test score above average)	0.876	0.865	0.010	0.907	0.892	0.014	0.863	0.835	0.028
			0.033			0.012			0.036
Fraction of Y0 observations staying in the sample (Y0 test score below average)	0.830	0.824	0.006	0.859	0.859	-0.000	0.815	0.815	-0.000
			[0.020]			[0.014]			[0.033]

Notes:

* p<0.1; ** p<0.05; *** p<0.01. The table presents tests on differential attrition. Different cohorts (defined in table A.3) stay in the sample for multiple rounds of the survey. We have attrition, but these attrition rates do not differ between the treatment and control groups. Standard errors allow for dependence within schools. Standard errors are reported in squared brackets.

Table A.5: testing for differential entry into the sample schools

	[1]	[2]	[3]	[4]	[5]	[6]
	New cohorts in Y2 (cohorts P2, P3, S2, S3)			New cohorts at Y3 (cohort P1, S1)		
	treatment	control	difference	treatment	control	difference
Average HH asset index of entering cohorts	4.694	4.639	0.055	4.811	4.568	0.243
standard error			[0.173]			[0.238]

Notes:

* p<0.1; ** p<0.05; *** p<0.01. New cohorts of students enter our sample schools after the intervention. The table reports tests on whether the socioeconomic backgrounds of students entering are the same between treatment and control. Students were asked 8 simple questions on household assets. Specifically, they were asked whether they have a TV, a fridge, a hand phone, a bicycle, a motor bike, a car, a computer, OR children's books at their home. The asset index we construct is the total number of items and may take on values from 0 to 8. Cohort P1 (first graders entering the sample schools for the first time at Y3) are not considered here, as they were not asked the asset questions for budgetary reasons. The table shows that there is no significant differential entry into the sample schools. Standard errors allow for dependence within schools. Standard errors are reported in squared brackets.

Table A.6: Teacher level impact

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
	Teachers already certified at Y0						**New teachers or teachers who are not eligible for certification at Y0**					
	Y2			Y3			Y2			Y3		
	treatment	control	difference	treatment	control	difference	treatment	control	difference	treatment	control	difference
Standardized test scores	0.198	0.178	0.020	0.172	0.201	-0.029	-0.046	-0.072	0.026	-0.141	-0.096	-0.045
	(0.952)	(0.978)	[0.122]	(0.989)	(0.892)	[0.140]	(1.068)	(0.994)	[0.105]	(1.106)	(1.007)	[0.116]
fraction with a bachelor's degree	0.985	1.000	-0.015*	0.982	0.987	-0.005	0.514	0.479	0.035	0.664	0.612	0.052
	(0.123)	(0.000)	[0.009]	(0.134)	(0.115)	[0.013]	(0.500)	(0.500)	[0.045]	(0.473)	(0.488)	[0.041]
fraction pursuing further education	0.051	0.035	0.016	0.049	0.034	0.014	0.337	0.378	-0.041	0.230	0.296	-0.065*
	(0.221)	(0.183)	[0.023]	(0.216)	(0.182)	[0.027]	(0.473)	(0.485)	[0.038]	(0.422)	(0.457)	[0.034]
fraction with a second job (self reported)	0.265	0.329	-0.063	0.278	0.245	0.033	0.289	0.332	-0.043	0.233	0.297	-0.064*
	(0.443)	(0.470)	[0.047]	(0.449)	(0.431)	[0.052]	(0.454)	(0.471)	[0.032]	(0.424)	(0.457)	[0.037]
Teaching hours per week	22.041	21.867	0.173	22.491	21.966	0.525	23.180	22.611	0.569	23.469	22.684	0.785
	(7.243)	(7.117)	[0.746]	(5.931)	(6.466)	[0.701]	(7.178)	(6.947)	[0.703]	(6.750)	(6.375)	[0.709]
Base pay (in MIL IDR)	2.636	2.612	0.024	3.001	3.003	-0.002	1.131	1.278	-0.146*	1.877	2.022	-0.145
	(0.390)	(0.533)	[0.065]	(0.419)	(0.381)	[0.051]	(0.948)	(0.997)	[0.086]	(0.974)	(0.888)	[0.096]
Allowances other than certification allowance (in MIL IDR)	1.008	0.898	0.111	0.681	0.794	-0.112**	0.332	0.322	0.010	0.341	0.386	-0.045
	(0.947)	(0.839)	[0.133]	(0.354)	(0.777)	[0.056]	(0.563)	(0.464)	[0.049]	(0.474)	(0.584)	[0.039]
Certification allowance (in MIL IDR)	2.089	2.106	-0.017	2.304	2.331	-0.026	0.140	0.081	0.059*	0.265	0.227	0.038
	(0.325)	(0.528)	[0.045]	(0.812)	(0.798)	[0.098]	(0.459)	(0.400)	[0.035]	(0.780)	(0.734)	[0.055]
Financial problems (self reported)	0.194	0.271	-0.077*	0.255	0.303	-0.048	0.635	0.659	-0.023	0.771	0.743	0.028
	(0.396)	(0.445)	[0.044]	(0.437)	(0.460)	[0.047]	(0.482)	(0.474)	[0.037]	(0.421)	(0.437)	[0.034]
Satisfied with total income (self reported)	0.883	0.905	-0.022	0.897	0.851	0.046	0.427	0.414	0.013	0.334	0.384	-0.049
	(0.323)	(0.294)	[0.027]	(0.305)	(0.357)	[0.035]	(0.495)	(0.493)	[0.036]	(0.472)	(0.487)	[0.039]
Absent from school at least once in the past week (self reported)	0.102	0.135	-0.033	0.068	0.118	-0.050*	0.180	0.153	0.027	0.155	0.157	-0.002
	(0.303)	(0.343)	[0.030]	(0.253)	(0.324)	[0.029]	(0.384)	(0.360)	[0.027]	(0.363)	(0.364)	[0.031]

Notes:

* p<0.1; ** p<0.05; *** p<0.01. Table compares average values between treatment and control schools for teachers who were already certified at Y0 (column [1]-[6]) and for teachers who were not eligible for certification at Y0 or who are new teachers in the sample schools (columns [7]-[12]) and evaluates these differences separately for the two moments of measurement Y2 (April 2011) and Y3 (April 2012). Standard errors allow for dependence within schools. Standard deviation values reported in parenthesis. Standard error of the estimated difference between treatment and control is reported in squared brackets.

Table A.7: Test for endogenous matching from students to (target) teachers

	[1]	[2]	[3]	[4]	[5]	[6]
	Y2			Y3		
	Panel A: All students					
	treatment	control	difference	treatment	control	difference
Fraction of students with a target teacher	0.501	0.506	-0.005	0.480	0.478	0.002
			[0.029]			[0.030]
	Panel B: Breakdown by student asset levels					
	treatment	control	difference	treatment	control	difference
Fraction of students with a target teacher -- asset level above average	0.514	0.530	-0.016	0.505	0.503	0.001
			[0.033]			[0.035]
Fraction of students with a target teacher -- asset level below average	0.484	0.476	0.008	0.480	0.466	0.014
			[0.030]			[0.035]

Notes:

* p<0.1; ** p<0.05; *** p<0.01. Panel A tests whether target teachers teach more classes. Panel B tests whether targets are more likely to be matched to students from higher/lower socio economic backgrounds. We do not find there are differences between treatment and control groups. The results suggest that there is no endogenous matching of teachers to students, in response to the intervention. Standard errors allow for dependence within schools. Standard errors are reported in squared brackets.