

Heterogenous Teacher Effects of Two Incentive Schemes

Evidence from a Low-Income Country

Felipe Barrera-Osorio

Jacobus Cilliers

Marie-Hélène Cloutier

Deon Filmer



WORLD BANK GROUP

Development Economics
Development Research Group

&

Education Global Practice

May 2021

Abstract

This paper reports on a randomized evaluation of two teacher incentive programs, which were conducted in a nationally representative sample of 420 public primary schools in Guinea. In 140 schools, high-performing teachers were rewarded in-kind, with the value of goods increasing with level of performance. In another 140 schools, high-performing teachers received a certificate and public recognition from the government. After one year, the in-kind program improved learning by 0.24 standard deviations, while the recognition treatment had a smaller

and statistically insignificant impact. After two years, the effect from the in-kind program was smaller (0.16 standard deviations) and not significant; the paper provides suggestive evidence that the reduction may be due to the onset of an Ebola outbreak. The effects of the recognition program remained small and insignificant. The effects differed by teacher gender: for female teachers, both programs were equally effective, while for male teachers, only the in-kind program led to statistically significant effects.

This paper is a product of the Development Research Group, Development Economics and the Education Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://www.worldbank.org/prwp>. The authors may be contacted at ejc93@georgetown.edu, felipe.barrera-osorio@vanderbilt.edu; mcloutier@worldbank.org; or dfilmer@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. 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.

Heterogenous Teacher Effects of Two Incentive Schemes: Evidence from a Low-Income Country*

Felipe Barrera-Osorio[†] Jacobus Cilliers[‡] Marie-Hélène Cloutier[§]
Deon Filmer[§]

Keywords: Student achievement; Teacher incentives; Learning outcomes; Recognition rewards

JEL Codes: I21; I28; J33; J45; O15

*Acknowledgements: We thank a number of colleagues and institutions that have assisted at various stages of the evaluation. Nathalie Lahire, Karine Angles, and Assane Dieng who played an instrumental in support the preparation and implementation of the evaluation. Adama Tiendrebeogo who provide both essential in-country field work coordination and research assistance throughout. Diwakar Kishore, Youn Park, and Ana Alvarez who provided critical research assistance to the project. Faya François Bourouno and Abdoul Aziz Diallo who supported in-country field work and supervision of data collection. This work would not have been possible without the impressive commitment from the Ministry of Pre-University and Civic Education (MEPU-EC), especially the teams of the Inspection Générale de l'Éducation (IGE) and the Cellule Nationale d'Évaluation du Système Éducatif (CNESE). We acknowledge funding from the Strategic Impact Evaluation Fund (SIEF) and the Research Support Budget (RSB) of the Development Economics unit (DEC). All findings and interpretations in this paper are those of the authors, and do not necessarily represent the views of their respective institutions or the Government of Guinea.

[†]Vanderbilt University

[‡]Georgetown University. ejc93@georgetown.edu

[§]World Bank

1 Introduction

Teacher incentive programs are a promising policy tool to improve student learning outcomes in developing countries. This hope stems from the fact that most developing countries currently face a “learning crisis” (World Bank, 2018), and there is evidence of limited teacher accountability (Mbiti, 2016). In fact, multiple studies have demonstrated the success of teacher pay-for-performance schemes in improving student learning outcomes in developing countries. Yet few governments have adopted them, either due to political reasons —teacher trade unions are often opposed to them— or ideological resistance to the idea. In contrast, governments frequently implement policies or have informal systems in place that reward high-performing civil servants through public recognition, such as award ceremonies or certificates. However there is little evidence on the effectiveness of these non-pecuniary incentive schemes. Such schemes activate different sources of motivation —reputational and intrinsic, rather than financial— and so could have very different impacts, depending on the teacher (Bénabou and Tirole, 2006).

This paper presents experimental evidence that allows us to investigate the effects of both types of approaches. We have two main research questions. First, whether a performance-based reward can result in better student learning outcomes, when implemented at a national level by government in a low-income country. Second, whether there is a difference in the effect of in-kind rewards versus rewards that only provide “recognition” for performance. We further explore whether there is heterogeneity across teachers in the effects of these reward schemes.

We analyze a nationally representative sample of 420 public primary schools in Guinea, situated in West Africa. These schools were assigned to one of three treatment arms. In 140 schools, relatively high-performing teachers were rewarded in-kind, with the value of goods increasing with level of performance (rice sacks, radios, phones, televisions, and generators). In another 140 schools, relatively high-performing teachers received a certificate and public recognition from government. The remaining 140 schools were assigned to the control group. Performance was calculated using a weighted average of (i) the annual gain in the average score of the teacher’s students on standardized tests; and (ii) the teacher’s annual improvement in the quality of his/her lessons preparation and delivery, as measured during inspection visits.

We conducted three rounds of data collection: at baseline, prior to the start of the program (May 2012), at midline after the first year of implementation (May 2013), and at endline after two years (May 2014). During these visits, fieldworkers conducted school principal and teacher surveys, as well as student assessments in mathematics and French

literacy. In addition, school inspectors visited schools to assess the quality of teaching and lesson preparation. Steps were taken to assure independence in data collection and minimize risk of collusion between fieldworkers and teachers.

We find that at midline, after one year of program implementation, there is a sizable and statistically significant impact on test scores (about 0.24 standard deviations, sd) for the in-kind treatment arm, and a much smaller and statistically insignificant impact for the recognition treatment arm (about 0.13 sd). At endline, after two years of program implementation, the impact of the in-kind rewards had fallen by about 30 percent in magnitude (to 0.16 sd) and is no longer statistically significantly different from zero. The impact of recognition rewards remains smaller (0.09 sd) and statistically insignificant. We also find that at midline the impacts are larger for girls than boys. We find no evidence of strategic behavior on the number of students who were assessed in each year, in average enrollment rates, or in changes in enrollment. We provide suggestive evidence that the reduction in impacts at endline is linked to the growing Ebola epidemic that Guinea experienced at this time.

Effects differ by teacher gender. The impact of in-kind rewards are very similar across genders, 0.20 sd and 0.18 for female and male teachers respectively. In contrast, the effects of the recognition rewards is 0.23 sd for female teachers, while it is 0.01 sd (and insignificantly different from zero) for male teachers. The findings further suggest that the positive effects of either scheme for female teachers benefit both female and male students. However, the positive effects of the in-kind treatment for male teachers benefit only female students. This is consistent with targeted action since female students have lower test scores at baseline.

Finally, data from the school inspections suggest an improvement in teaching preparation and practice. Teachers in the in-kind treatment arm made better use of resources in the classroom, were better prepared for the class, and implemented better teaching practices. Consistent with the results on student learning, the impacts on teaching practice are no longer statistically significant at endline.

Our paper adds to a growing literature with experimental evidence from teacher pay-for-performance interventions in low- and middle-income countries. This literature documents a range of findings: some studies find no impacts (Barrera-Osorio and Raju, 2017); others document positive impacts (Behrman et al., 2015; Cilliers et al., 2018; Filmer et al., 2020; Leaver et al., 2020; Loyalka et al., 2019; Muralidharan and Sundararaman, 2011); and others have shown impacts only in the presence of complementary resources (Gilligan et al., 2018; Mbiti et al., 2019). Last, one study finds impacts on the incentivized test, but no evidence of impacts on any other learning outcomes (Glewwe et al., 2010). In contrast, null and negative impacts have been typically been found in experiments in high-income settings

(Fryer, 2013; Fryer Jr, 2011). Given these mixed results, it is clear that context and the details of design matter greatly for determining the impacts of performance-based rewards (Bruns et al., 2011). In some contexts teachers may lack the skills and knowledge to respond effectively to incentives, even if they wished to.

Our contribution is both to the design and to the discussion of context specificity of these program. First, we document the contrast in the impacts between in-kind and recognition rewards, and explore heterogeneity in how these effects are mediated by teacher and student gender. To the best of our knowledge, this is the first paper with a clear identification strategy to look at the contrasts between financial and recognition rewards for government civil servants. A similar study experimentally compared financial incentives with providing recognition for pro-social behavior, although the agents were private-sector hairdressers (Ashraf et al., 2014). Second, we document impacts on learning outcomes even in an environment where teacher skills are relatively low.

2 Intervention, Sample, and Evaluation Design

Guinea, located on the west coast of Africa, is one of the poorest countries in the world. It is ranked 174/189 on the Human Development Index and, as of 2012, 36 percent of the population lived in extreme poverty.^{1,2} School enrollment at all education levels have been increasing in recent years, but serious challenges remain both in access and quality. For example, an assessment administered in 2017 revealed that only 25 percent of grade two and three children could read a simple text,³ and according to the Human Capital Index a child born today can expect to complete 7 years of schooling, but adjusting for quality suggests that amounts to only 4.5 years of high-quality schooling (World Bank, 2018).

These apparent problems induced a newly elected government in 2011 to seek innovative approaches in personnel and human resource management reforms. Within this context, the Ministry of Pre-University and Civic Education (MEPU-EC) initiated a donor-funded individual performance-based reward program for teachers. The program started in the 2012-2013 academic year and was implemented relatively well (see discussion below), but faced serious challenges in its last year of implementation (2013-14 school year) due to the outbreak of Ebola in December 2013, and especially as the epidemic spread in 2014.⁴

¹<http://hdr.undp.org/en/data>

²<https://data.worldbank.org/country/guinea>

³<https://resenguinee.org/>

⁴Source: <https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak> Accessed September 5th, 2020

2.1 Performance-Based Reward Program

The program was designed in collaboration with government and key stakeholders like teacher unions. While it used principles derived from optimal design, the approach emphasized simplicity in the pursuit of broad understanding. The program targeted grade 3 and 4 teachers, and lasted for two academic years (academic years 2012-13 and 2013-14). At these grades, teachers are assigned to a specific group of students and teach all subjects for that group.⁵

Rewards took the form of either goods (in-kind) or recognition, depending on the treatment arm to which the teacher’s school was assigned. The value of the reward was determined on an absolute scale, without relative performance comparisons to other teachers.⁶

The performance indicator used in both programs was a weighted sum of improvement in learning outcomes and improvements in observed teacher inputs. The indicator measuring learning outcomes was the annual gain (in percentage points) in the average score of the teacher’s students on standardized tests in Math and French. The input indicator was the teacher’s annual improvement in the quality of his/her lessons preparation and delivery, defined as the annual gain (in percentage points) in the inspection scores given to the teacher for two lessons—in French and one in Math—during inspection visits.⁷ The sub-indicators were aggregated using a weight of 0.7 for improvement in learning outcomes and 0.3 for improvement in teacher inputs. In the second year a third indicator, the student attendance rate on the day of the test, was introduced to disincentivize teachers from encouraging some students to stay home on the test day. Weights were then modified to be 0.6, 0.3 and 0.1. Data on these performance sub-indicators were collected through announced visits. Specific effort was made in the design and administration of the assessments and inspections to minimize cheating and collusion. First, the research team used a randomized booklet design, where students in a school were randomly assigned to different test booklets that contained different test items covering the same subject subdomain and some overlapping items. One comparable learning score was constructed for each student *ex post* using Item Response Theory.⁸ Second, test booklets were kept secure prior to administration and answer keys were never distributed, either before or after test administration. Third, tests were administered in the schools by external agents recruited by the education ministry. Fourth,

⁵Except for koranic lessons for which a special teacher is assigned.

⁶Section A.1 provides more details on the types of reward, actual receipt of rewards (Tables A.1 and A.2), implementation of the award ceremonies, and communication strategy to inform schools and teachers about the program.

⁷Both sub-indicators took on a value of zero if the teacher’s performance worsened.

⁸Items from a large pool of potential test items were piloted with students in a set of schools outside of the study sample. Statistical analysis of the piloting data was carried-out to identify items that can be combined into booklets to give accurate and comparable assessment results.

administration conditions, timing, resources and environment were kept as consistent as possible between schools. Fifth, the inspections and classroom observations were conducted by regular inspectors from the ministry, but they were assigned to a region different from their regular assignment to increase objectivity and reduce the risk of collusion between the teachers and inspector.

The program roll out was accompanied by a strong communication campaign (conducted in December 2012) to teachers and principals. This communication campaign was implemented by a specialized consulting firm using pamphlets, posters, and meetings with communication agents. Understanding of the program was also reinforced during the award ceremonies and disbursement of gifts at the end of the first year (December 2013-January 2014). In the second year, ceremony and reward distribution were initially delayed because of the Ebola crisis (which had led to a ban on people gathering in groups), and were ultimately canceled altogether. See Figure A.1 for the timeline of implementation, data collection, and outbreak of Ebola.

Overall, teachers' awareness of the program was high, but knowledge about the specific aspects of program design was weaker (Table A.3). Almost all teachers, whether in the incentive treatment or in the control group, reported being part of the program. We suspect that this is due to their (incorrect) interpretation of the regular visit by inspector and teams from ministry for the student test and interviews. It is thus a possibility that we are underestimating effect sizes if teachers in the control schools also changed their behavior due to incorrect beliefs about their participation in the program. In general, knowledge of key aspects of the program were high, but not perfect. For example, at midline about 75 percent of teachers in the treatment groups—but not all—believed that the program targeted grade 3 and 4 teachers. At midline 91 percent of teachers in the in-kind arm believed that the award had an in-kind component, and 62 percent of teachers in the recognition arm believed that the program included a certificate. Teachers were also relatively well-informed about the inputs to the award function.

2.2 Sample and Experimental Design

The evaluation sample consists of 420 randomly selected primary schools, stratifying by regions and zones (rural-urban), from the population of public French (non-Arabic) speaking schools found in the Education Management Information System (EMIS) database and using the number of teachers in targeted grades to define the selection probability of a given school.⁹

⁹The sampling frame was limited to public schools because they are the schools under direct authority of the MEPU-EC.

The sample is therefore nationally representative for grade 3 and 4 teachers in public French-speaking schools in Guinea.

Out of this sample, 140 schools were randomly assigned to each of the two treatment arms—recognition or in-kind performance rewards—and 140 schools were randomly assigned to the control. Random assignment was carried out publicly in the context of a program launching workshop, after the baseline data collection. Names of schools were put in bowls (per strata) and picked (blindly) by young children to be assigned to one of the groups. The public nature of the randomization process aimed at maximizing trust in the transparent and impartial nature of the allocation process.

2.3 Data

Data were collected on schools, teachers, and students using the following survey instruments: (i) a questionnaire administered to the school’s principal, (ii) a teacher questionnaire administered to targeted teachers (grade 3 and 4), (iii) a Math test and French test (with different parallel booklets) administered to students in the targeted teachers’ classrooms, and (iv) an inspection bulletin administered to targeted teachers in the context of two lessons, one in French and one in Math. Table A.4 presents a more detailed description of the various instruments.

All grade 2 and 3 students in the sampled schools were assessed at baseline (May 2012), and all students in grades 2 to 4 were assessed at midline (May 2013). Because of budget constraints, a random sample of up to 25 students were assessed from each class in grades 3 and 4 at endline (May 2014). The surveys reached a varying number of schools across the different rounds (Table A.5). For example, we have student assessment data for 408, 417, and 391 schools for baseline, midline, and endline rounds of data collection, respectively. It is unclear the source of the variation in number of schools; however, it is likely that, at least for the endline, the outbreak of Ebola might have been a contributing factor.

Most data collection activities were carried out using the existing government systems (no external survey firm was used) and took place simultaneously in each region of the country. Student and teacher tests, as well as principal, teacher and student questionnaires were administered by decentralized government staff selected on a competitive basis and trained as enumerators and supervisors. Quality of the data collected through the above mechanisms was validated through spot check visits from field coordination teams.

The data collection effort was set up as a panel of schools, not of teachers or students. Some teachers—those still teaching in the surveyed grade—appeared in multiple rounds of the data collection; other teachers were new and others left the grade of the school. As such,

we do not have a panel structure for teachers.

Due to budget limitations, the analysis of impacts and the calculation of rewards use the same tests. Given that the program is attached to the same test, there may be problems of degradation and inflation of the metric (Koretz, 2002; Neal, 2013). Nevertheless, we use a measure of learning outcomes that is different from the metric used to calculate the incentive payments. While the latter uses an IRT model to account for differences between grades, our estimation of impacts use a grade fixed-effects specification. In addition, as discussed above, the measures taken to minimize cheating mitigate this concern. Furthermore, it is unlikely that teachers were merely “teaching to the test” in the first year of implementation, since teachers had no experience with the tests.

At the time of baseline data collection, teachers’ mean monthly salaries were USD118, 73 percent of teachers felt that their salary was insufficient, 27 percent of teachers felt that they received insufficient recognition for their work, 67 percent were female, and 70 percent mentioned that they would choose teaching as a profession if they could choose over (Table A.6). There were also large disparities by gender. Female teachers reported to have larger classes, earn less, and were less satisfied with their earnings, but were also 10 percentage points *more* likely to state that they would choose teaching again as a profession, if they had the chance to choose over.

In addition to the primary data collection, we also constructed a data set containing the total number of reported deaths due to Ebola for each prefecture by May 2014, the time of endline data collection. We combined the weekly reports on the total number of new cases by prefecture, provided by the WHO. The distribution of known deaths due to Ebola by May 2014 across prefectures is highly varied with the majority of cases were in one prefecture (168 cases in Gueckedou), and just over 65% of the prefectures had no known cases (Figure A.2).

We perform a range of tests to demonstrate the internal validity of the study.¹⁰ The sample is balanced for a range of student and school-level characteristics, including student baseline learning, number of students assessed, and exposure to Ebola (Table A.7). Furthermore, school-level attrition in student assessment data is balanced (Table A.8), and the sample remains balanced if restricted to schools for whom we have midline and endline data (Tables A.9 and A.10, respectively). Finally, there is no *reduction* in the number of students assessed at midline and endline in the respective treatment groups, suggesting that teachers did not strategically keep under-performing students from attending school on the day of assessment to inflate their scores (Tables A.11 and A.13).

¹⁰Section A.2 provides a more detailed discussion.

2.4 Empirical Strategy

Our main estimating equation is the following:

$$Y_{igas} = \beta_0 + \beta_1(\text{In-kind})_s + \beta_2(\text{Recognition})_s + X_{igs}\Pi + \rho_b + \lambda_g + \gamma_a + \epsilon_{igas} \quad (1)$$

where Y_{igas} is learning outcome for student i , grade g , assessment a (Math or French language assessment), and school s . ρ_b refers to strata fixed effects.¹¹ X_{igs} is a vector of school-level and student-level controls: the head-teacher's and students' age and gender, and the school's average baseline performance for grades 2 and 3, respectively.¹² λ_g and γ_a are grade and assessment fixed effects.

For all specifications, we cluster standard errors at the school level. We run two separate regressions for midline and endline observations. Observations are weighted by the inverse of the number of students assessed in the school. This model allows us to directly test the impact of each type of reward approach (i.e. the coefficients β_1 and β_2) as well as test whether the impacts are different for the two approaches (i.e. the test that $\beta_1 = \beta_2$).

For the heterogeneity analysis by teacher gender, we estimate Equation 1 separated for male teachers and female teachers.

When testing whether treatment effects vary by student gender, we estimate the following:

$$Y_{igas} = \alpha_0 + \alpha_1(\text{In-kind})_s + \alpha_2(\text{Recognition})_s + \alpha_3(\text{In-kind} \times \text{Boy})_{is} + \alpha_4(\text{Recognition} \times \text{Boy})_{i,s} + X_{igs}\Pi + \rho_b + \lambda_g + \gamma_a + \epsilon_{igas} \quad (2)$$

This model allows us to test the impact of each type of reward approach for girls (i.e. the coefficients α_1 and α_2), for boys ($\alpha_1 + \alpha_3$ and $\alpha_2 + \alpha_4$), and the difference in effect size between boys and girls (α_3 and α_4). Moreover, we can test whether the impacts for the two treatments are the same, again separately for girls ($\alpha_1 = \alpha_2$) and boys ($\alpha_1 + \alpha_3 = \alpha_2 + \alpha_4$).

3 Results

3.1 Student Test Scores

Table 1 presents our main results. At midline, after one year of program implementation, the in-kind and recognition reward programs improved student learning by 0.24 and 0.13

¹¹The assignment across groups is balanced within each stratum.

¹²If a school does not have baseline data for a specific grade, we assign it to the mean value in the control and create a dummy variable equal to one if data is missing.

standard deviations, respectively. We cannot reject the null of no impact for the recognition program (p-value= 0.205), nor the null for equal effects across arms (p-value 0.195; last row of the table). There is evidence of decay in the effect of the program: at endline, after two years of program implementation, the impact of the in-kind and recognition reward programs were smaller, 0.16 and 0.09 sd respectively, and statistically insignificant.

Table 1 also unpacks effects by grade (Columns (3) to (6)) and by subject (Columns (7) to (10)). The impacts are consistently larger for grade 4 students, relative to grade 3 students, for both treatments and both rounds of data collection.¹³ The effects of the in-kind arm in French and math were similar and positive at midline; the analogous effects at endline are still positive, but only statistically significant for French. These results by grade and subject consistently show positive and higher point estimates of the in-kind treatment than the recognition arm, despite the fact that we cannot statistically rule out equality at conventional levels of significance.

This program is very cost-effective (Table A.12). Assuming that a random sample of 25 students per school are assessed in each round, and that the cost of the midline ceremonies and goods distribution would have carried over to the endline had they occurred, we estimate that the in-kind and recognition programs cost approximately \$11.75 and \$8.29 per student per year respectively (including costs of data collection, award ceremonies, and value of goods awarded). The in-kind program is thus more cost-effective with a 2 sd increase in learning for each \$100 spent per student, relative to 1.48 sd/\$100 for recognition. We note that the biggest cost-driver is the cost of assessing students, which explains why the recognition rewards program was only 29 percent less expensive than in-kind rewards.

3.2 Teacher Heterogeneity

We examine two sources of teacher heterogeneity: teacher gender, and teacher “identity” (defined below). Our analysis is constrained by the fact that we cannot link teachers with students, although we can match at a school-grade level. The data is therefore restricted to school grades where there is only one teacher, or all the teachers in that grade share the same characteristics.

Table 2 examines this heterogeneity. We have 86,840 observations for the whole sample (Table 1, Column 1) and 58,949 observations for the sample in which we can cleanly establish the gender of the teacher. Column (1) replicates the overall result on test scores at midline for the sample with teacher gender. It is reassuring that the point estimates are very similar to the ones in Table 1, Column 1. Columns (2) and (3) presents the estimates disaggregated

¹³In line with fact that the program only targeted teachers in grades 3 and 4, we find no impact on learning for grade 2 students at midline. These results are available upon request.

by teacher gender. The effect of in-kind incentives is very similar for both female and male teachers (point estimates of 0.19 and 0.19 sd respectively). In contrast, the effect of the recognition incentive varies with teacher’s gender. The point estimate of the recognition treatment for female teachers is 0.23 sd (and is statistically significant at conventional levels), while the effect for male teachers is 0.004 sd. We cannot reject the null hypothesis of equality between the impacts of the in-kind and recognition treatments for male teachers, due to weak statistical power.

In short, there is suggestive evidence that the impacts of these performance incentives vary with the gender of the teacher: the effects on test scores of students from female teachers are higher for *both* treatments, while the effects from male teachers are driven only for the in-kind treatment.

We posit two potential reasons why female teachers are more responsive to the recognition program than male teachers that we can empirically investigate. First, it is plausible that a public recognition of one’s professional competence is more valuable to individuals whose personal identity is linked to that profession. Second individuals who feel under-recognized might respond stronger to opportunities to increase the recognition of their work. It is possible that female teachers identify more strongly as teachers, and also receive lower recognition for their work overall.

We find some suggestive evidence for the former hypothesis, but not the latter. As a coarse measure of “identity”, we asked teachers what profession they would choose if they could choose again. At baseline, female teachers are 10 percentage points more likely to indicate that they would choose teaching as a profession again: 77 vs 67 percent. Columns (3) to (6) in Table 2 shows the same pattern of heterogeneous effects based on this question. Teachers who respond that they would choose teaching as a profession respond equally well to both the in-kind and recognition reward programs. However, the impact of the recognition arm for teachers who would *not* choose the profession again is much smaller —0.007 sd— and not statistically significant. Again, we cannot reject the null hypothesis of equality between the impacts of the in-kind and recognition treatments for teachers who would not choose the profession, due to limited power. In contrast, we find no evidence that female teachers are more likely to be unsatisfied at baseline with the level of recognition they receive for their work. Moreover, there is no evidence that teachers who are dissatisfied with the recognition they receive for their work responded more to the recognition program.¹⁴

We further disaggregate the results by the gender of the student, by estimating Equation 2. The first column in Table 3 shows that in both programs girls benefited more than boys. The difference in effect size between boys and girls is largest for the recognition arm (0.12

¹⁴Results available upon request.

sd), and statistically significant at the 5 percent level. As a result, both programs had a positive and statistically significant impact on learning for girls at midline. Note that boys in the control group out-perform girls by 0.18 sd on average. This means that the programs also succeed at reducing the difference in performance between boys and girls, by roughly a third and two thirds in the in-kind and recognition arms respectively.

In columns (3) and (4) we report the same results as above, but for female and male teachers separately. Again, the comparison between Columns (1) and (2) is important since it confirms that the restricted sample for which we can ascertain a teacher’s gender does not drive the results. The effect for a female teacher on female students is 0.21 sd for the in-kind treatment, and 0.24 sd for the recognition treatment; with both being statistically significantly different from zero at conventional levels. The effects of a female teacher on male students is very similar: the interactions between treatment arm and the boy indicator are small and non-significant. In contrast, there are differential effects for male teachers (Column 4). In this case, the effect for a male teacher on a female student in the in-kind treatment is 0.27 sd, while the effect on male students is 0.12 (0.27-0.15) sd; we note that the coefficient on the interaction term, -0.15, is not statistically significant. The effect for a male teacher on a female student from the recognition treatment is 0.13, which is not statistically significant different from zero, and the effect on male student is -0.08 (0.13-0.21) sd which is negative but not statistically significantly different from zero.

In sum, these estimates present an intriguing perspective of the effects of teacher incentive programs. On one hand, there is supporting evidence that female teachers respond to both pecuniary and non-pecuniary incentives, and do so in ways that benefit all students, male and female. On the other hand, the estimates indicate that male teachers respond mainly to the pecuniary incentive, and do so in a way that benefits only female students. This behavior is compatible with them taking a strategic approach, since female students of male teachers tend to have markedly lower counterfactual test scores than male students.

3.3 Inspectors’ Classroom Observations

We further investigate the ways through which teachers may have responded to incentives by using the data on teacher practices collected through the inspector classroom observations. The inspector assigned each teacher scores on two indicators of classroom quality and five indicators of teaching quality. We standardized each score to have a control group mean of zero and standard deviation of one, and then took the mean of these standardized scores to create an overall index of classroom and teaching quality, respectively. Table 4 reports the effects of the program on these indicators at midline. The specification is similar to that for

Equation 1, but with the outcome variables at the level of the teacher.

Overall, we see positive effects from the in-kind treatment, with effect sizes of 0.14 sd on the overall index of classroom quality, and 0.11 sd on the overall index of teaching quality (with both being statistically significantly different from zero). In contrast, the effects for the recognition arm are not statistically significantly different from zero, and both point estimates are lower than those for the in-kind treatment. Breaking out the sub-components of these overall indices yields systematically positive effects from the in-kind treatment, but varied impacts (both in size and direction) of the recognition treatment.

We take these results with some caution: the evaluation of the inspector is an important input for the provision of the rewards. In the case of the in-kind treatment, there may have been a perception that there was more at stake (since the reward is material), and we cannot rule out that the positive results may be driven by collusion between the inspector and teacher. We do not have evidence of this kind of behavior (or lack thereof). Mitigating this concern is the fact that we actually observe student effects from the in-kind treatment: If inspector observations were driven purely by collusion, we would have expected to find no impacts on student outcomes.

3.4 Exposure to Ebola

We present evidence that the outbreak of Ebola might have been a contributing factor for the reduction in treatment effect sizes at endline, relative to midline. Table 5 presents the results from estimating a model that relates student test scores as a function of the treatment arms, an indicator of Ebola, and the interaction of Ebola and treatments. As a placebo test, column (1) shows that the magnitude of the treatment effects did not vary by exposure to Ebola at midline, prior to the actual outbreak of the disease. Column (2) shows that at endline the treatment effects for both interventions are smaller in magnitude in prefectures with higher exposure to Ebola. For both interventions, the effect sizes decrease by 0.006 standard deviations for every additional death from Ebola. The treatment effects in the prefectures not exposed to Ebola are very similar in magnitude to the average treatment effects at midline. This suggests that all the reduction in effect sizes between midline and endline could be attributed to the Ebola outbreak.¹⁵

¹⁵These results are robust to different transformations of the Ebola variable (e.g. hyperbolic sine, Table A.14), and to the exclusion of five prefectures that had not reported Ebola cases by the time of endline data collection (Table A.15).

4 Discussion and Conclusion

Teacher incentives have the potential to encourage teachers to exert more effort and ultimately lead to increase in student learning, but they could also lead to distortionary behaviors which could be detrimental for learning. The details of program design have been shown to matter for impacts, especially if they interact with teachers' intrinsic motivations differently. In the context of primary schools in Guinea, this impact evaluation shows that teacher incentives led to improvements in student learning outcomes after one year of program implementation—with in-kind rewards having impacts that are roughly twice as large as those of recognition rewards. After two years of the program, the impacts are substantially muted—they roughly halved in size—and no longer generally statistically significantly different from zero. This reduction in effect sizes was likely driven by the Ebola epidemic that was emerging close to the two-year mark of the program. Analysis of enrollment patterns across years and across cohorts does not suggest that there were any perverse enrollment impacts of the program. Data from the school inspections suggest an improvement in teaching preparation and practice: teachers made better use of resources in the classroom, were better prepared for the class, and implemented better teaching practices. Moreover, the results from lesson observations mirror the results on student learning.

Male and female teachers respond differently to the different types of incentives. While there are many potential explanations for this result, we have suggestive evidence that the different types of incentives may be activating different underlying sources of motivation. In particular, the degree to which a teacher identifies themselves with the profession—potentially signaling a greater degree of intrinsic motivation—might be related to their responsiveness to the recognition incentive.

These results have potential implications for policy. They suggest that teachers do indeed respond to performance incentives, and can do so in ways that improve student learning outcomes, even in a low-capacity environment. However, non-pecuniary rewards—which are often preferred by policy makers—might only be effective for some types of teachers.¹⁶ This means that the underlying motivations of teachers will be important in determining the effectiveness of such interventions, and that policy makers should invest in understanding those motivations prior to implementation.

¹⁶This result complements Leaver et al. (2020) who show in Rwanda that there is a range in the degree to which teachers are intrinsically motivated, and that the less intrinsically motivated teachers are more responsive to pecuniary performance-related rewards.

5 Reference

References

- Ashraf, Nava, Oriana Bandiera, and B Kelsey Jack, “No margin, no mission? A field experiment on incentives for public service delivery,” *Journal of public economics*, 2014, 120, 1–17.
- Bank, World, “The Human Capital Project,” *World Bank*, 2018.
- Barrera-Osorio, Felipe and Dhushyanth Raju, “Teacher performance pay: Experimental evidence from Pakistan,” *Journal of Public Economics*, 2017, 148, 75–91.
- Behrman, Jere R, Susan W Parker, Petra E Todd, and Kenneth I Wolpin, “Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools,” *Journal of Political Economy*, 2015, 123 (2), 325–364.
- Bénabou, Roland and Jean Tirole, “Incentives and prosocial behavior,” *American economic review*, 2006, 96 (5), 1652–1678.
- Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos, *Making schools work: New evidence on accountability reforms*, The World Bank, 2011.
- Cilliers, Jacobus, Ibrahim Kasirye, Clare Leaver, Pieter Serneels, and Andrew Zeitlin, “Pay for locally monitored performance? A welfare analysis for teacher attendance in Ugandan primary schools,” *Journal of Public Economics*, 2018, 167, 69–90.
- Filmer, Deon, James Habyarimana, and Shwetlena Sabarwal, “Teacher Performance-Based Incentives and Learning Inequality,” 2020.
- Fryer, Roland G, “Teacher incentives and student achievement: Evidence from New York City public schools,” *Journal of Labor Economics*, 2013, 31 (2), 373–407.
- Gilligan, Daniel O, Naureen Karachiwalla, Ibrahim Kasirye, Adrienne M Lucas, and Derek Neal, “Educator incentives and educational triage in rural primary schools,” Technical Report, National Bureau of Economic Research 2018.
- Glewwe, Paul, Nauman Ilias, and Michael Kremer, “Teacher incentives,” *American Economic Journal: Applied Economics*, 2010, 2 (3), 205–27.
- Jr, Roland G Fryer, “Financial incentives and student achievement: Evidence from randomized trials,” *The Quarterly Journal of Economics*, 2011, 126 (4), 1755–1798.

- Koretz, Daniel M**, “Limitations in the use of achievement tests as measures of educators’ productivity,” *Journal of human resources*, 2002, pp. 752–777.
- Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin**, “Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants,” 2020.
- Loyalka, Prashant, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi**, “Pay by design: Teacher performance pay design and the distribution of student achievement,” *Journal of Labor Economics*, 2019, 37 (3), 621–662.
- Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani**, “Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania,” *The Quarterly Journal of Economics*, 2019, 134 (3), 1627–1673.
- Mbiti, Isaac M**, “The need for accountability in education in developing countries,” *Journal of Economic Perspectives*, 2016, 30 (3), 109–32.
- Muralidharan, Karthik and Venkatesh Sundararaman**, “Teacher performance pay: Experimental evidence from India,” *Journal of political Economy*, 2011, 119 (1), 39–77.
- Neal, Derek**, “The consequences of using one assessment system to pursue two objectives,” *The Journal of Economic Education*, 2013, 44 (4), 339–352.
- World Bank**, *World Development Report 2018; Learning to Realize Education’s Promise*, Washington, DC: World Bank, 2018.

Table 1: Impacts on learning— by round, grade and subject

	Full sample		Grade 3		Grade 4		French		Math	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Mid	End	Mid	End	Mid	End	Mid	End	Mid	End
In-Kind	0.239*** (0.084)	0.156 (0.099)	0.129 (0.097)	0.121 (0.108)	0.235*** (0.090)	0.184* (0.109)	0.202** (0.082)	0.183* (0.095)	0.274*** (0.093)	0.130 (0.108)
Recognition	0.125 (0.087)	0.088 (0.103)	-0.037 (0.094)	0.053 (0.114)	0.224** (0.094)	0.130 (0.115)	0.102 (0.084)	0.118 (0.098)	0.146 (0.097)	0.060 (0.113)
Control mean	0.031	0.058	0.043	0.044	0.017	0.040	0.035	0.035	-0.004	-0.004
Observations	86840	57808	44105	28278	42735	29530	43421	28904	43419	28904
R-squared	0.097	0.121	0.124	0.121	0.084	0.109	0.109	0.153	0.097	0.102
Test:In-Kind=Recognition	0.205	0.514	0.092	0.577	0.908	0.647	0.256	0.511	0.199	0.543

Notes. Each column represents a separate regression estimated using equation 1. The midline and endline results are in the odd- and even-numbered columns respectively. The first two columns includes data for both grades and subjects, and includes grade and subject fixed effects. Columns (3)-(6) includes data for both subjects; columns (7)-(10) includes data for all grades. The final row reports the p-value of the F-test of equation of coefficients. Standard errors, clustered at a school level, are in parentheses. $*p < 0.1$, $**p < 0.05$, $***p < 0.01$

Table 2: Impacts on learning— by teacher gender and identity

	Teacher gender			Preferred profession		
	(1) Either	(2) Female	(3) Male	(4) Either	(5) Teacher	(6) Other
In-Kind	0.230** (0.100)	0.194* (0.111)	0.189 (0.157)	0.255*** (0.098)	0.284** (0.115)	0.220 (0.159)
Recognition	0.159 (0.105)	0.227* (0.125)	0.004 (0.156)	0.206** (0.104)	0.222* (0.116)	0.007 (0.189)
Observations	58949	47741	11208	46872	36048	10824
R-squared	0.099	0.115	0.183	0.141	0.157	0.225
Test:In-Kind=Recognition	0.501	0.780	0.250	0.633	0.592	0.252

Notes. Each column represents a separate regression on the midline data, estimated using equation 1. Data in columns (1) to (3) are restricted to observations where all teachers in a given grade and school report the same gender. This allows us to unique identify the gender of the teacher. Columns (2) and (3) split this sample to female and male teachers, respectively. Similarly, columns (4) to (6) restrict the sample to schools and grades where all teachers either selected “teacher” or all teachers selected another profession to the question, asked at baseline: “if you could choose again, what profession would you choose?”. Columns (4) and (5) split this sample to teachers who selected “teacher” or not. Standard errors are in parentheses, and clustered at the school level. $*p < 0.1$, $**p < 0.05$, $***p < 0.01$

Table 3: Impacts on learning— by student and teacher gender

	(1) All teachers	(2) Female	(3) Male
In-Kind	0.272*** (0.100)	0.210* (0.116)	0.265* (0.156)
Recognition	0.223** (0.100)	0.237* (0.121)	0.127 (0.149)
Boy	0.182*** (0.041)	0.079*** (0.029)	0.303*** (0.064)
In-Kind x Boy	-0.072 (0.060)	-0.011 (0.043)	-0.146 (0.099)
Recognition x Boy	-0.124** (0.062)	-0.012 (0.060)	-0.211** (0.088)
Observations	58891	47707	11184
R-squared	0.098	0.111	0.184
<i>F-test (p-value): Impact on boys</i>			
In-Kind	0.065	0.074	0.478
Recognition	0.389	0.102	0.614
<i>F-test (p-value): In-Kind vs Recognition</i>			
Girls	0.635	0.812	0.391
Boys	0.372	0.842	0.226

Notes. Each column represents a separate regression on the midline data, estimated using equation 2. Data are restricted to same sample as columns (1) to (3) in Table 2. The bottom four rows report the p-values of the following respective F-tests: $\hat{\alpha}_1 + \hat{\alpha}_3 = 0$; $\hat{\alpha}_2 + \hat{\alpha}_4 = 0$; $\hat{\alpha}_1 = \hat{\alpha}_2$; $\hat{\alpha}_1 + \hat{\alpha}_3 = \hat{\alpha}_2 + \hat{\alpha}_4$. Standard errors are in parentheses. . * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Inspection-level outcomes at midline

	Classroom quality			Teaching quality				
	(1) Overall	(2) Spatial org.	(3) Educ. documents	(4) Overall	(5) Written prep	(6) Teaching material	(7) Class practice	(8) Reflection
In-Kind	0.143** (0.056)	0.149* (0.078)	0.136* (0.071)	0.111** (0.046)	0.095 (0.071)	-0.022 (0.073)	0.115 (0.075)	0.256*** (0.067)
Recognition	0.090 (0.060)	-0.020 (0.087)	0.200*** (0.073)	0.079 (0.051)	0.210*** (0.075)	-0.240*** (0.078)	0.250*** (0.079)	0.096 (0.078)
Control mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations	1056	1056	1056	1056	1056	1056	1056	1056
R-squared	0.050	0.034	0.117	0.080	0.098	0.124	0.075	0.053
Test:In-Kind=Recognition	0.354	0.050	0.355	0.508	0.104	0.004	0.072	0.024

Notes. Each column represents a separate regression using equation 1. Data is at a teacher level. The dependent variables in columns (2)-(3) and (5)-(8) are the scores assigned by inspectors for different domains of classroom and teaching quality, standardized to have a control mean of zero and standard deviation of one. The dependent variable in column (1) is the mean of the dependent variables in columns (2) and (3). The dependent variable in column (4) is a mean of the dependent variables in columns (5) to (8). Standard errors are in parentheses and clustered at the school level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Impacts on learning— by exposure to Ebola

	Midline	Endline		
	(1)	(2)	(3)	(4)
	Full	Full	Grade 3	Grade 4
In-Kind	0.242*** (0.088)	0.227** (0.101)	0.129 (0.124)	0.294** (0.124)
Recognition	0.109 (0.093)	0.171 (0.106)	0.129 (0.136)	0.213* (0.128)
Ebola	-0.001 (0.002)	-0.000 (0.002)	0.000 (0.002)	-0.000 (0.002)
Ebola x In-Kind	0.000 (0.003)	-0.006** (0.002)	-0.004 (0.003)	-0.007** (0.003)
Ebola x Recognition	0.001 (0.002)	-0.006*** (0.002)	-0.005** (0.002)	-0.008*** (0.003)
Observations	86578	57444	28146	29298
R-squared	0.124	0.143	0.149	0.154

Notes. Each column is a separate regression including the same controls as in equation 1. The variable “Ebola” is the total number of deaths reported in a prefecture by the time of endline data collection, May 2014. The WHO only reported cases for 29 out of 33 prefectures. We assume that there were no cases for the four prefectures with no data, but results are the same if instead we treat these prefectures as having missing data (Table A.15).

A ONLINE APPENDIX

A.1 Determination of Rewards and the Reward Ceremonies

Table A.1 shows the mapping between the performance indicator and rewards. The rewards ranged from bronze to platinum. There was one national ceremony (for recipients of the platinum certificate), one ceremony in each of the eight regions (for the gold certificate), one ceremony in each of the 38 prefectures (silver), and one ceremony in each of the 243 sub-prefectures (bronze). The ministry of education organized the national ceremony; the Regional Authorities of Education organized the regional and prefecture-level ceremonies; and the “School Delegation of Elementary Education” (DSEE) organized the sub-prefecture ceremonies. The ceremonies were attended by government officials, trade unions, national or local press, locally elected officials, and school parent associations.

Table A.2 shows the breakdown of awards received after the first year of the program and the value of the awards as a proportion of the average civil servant teacher salary. Over a third did not receive any reward, and a quarter of teachers received the highest possible reward. The value of the reward was highly convex, with 4% of salary for the lowest tier (rice and radio), and almost 49% of a teacher’s annual salary for the highest tier (rice, TV, cell phone, and generator).

A.2 Tests for Internal Validity

A.2.1 Balance and Attrition

We investigate internal validity by comparing baseline data from Grade 2 and Grade 3 students from the student assessment, as well as school-level data as reported by school principals, across the groups. The sample is balanced on key variables—test scores and gender—collected at the baseline; only age shows a difference between control and treatment arms (Table A.7). School characteristics (number of Grade 2 and Grade 3 students, headteacher gender and age) are likewise never statistically significantly different across the groups, with the exception of headteacher age which is very slightly higher in the control group than in the recognition group. These differences are consistent with random chance. The sample is also balanced by exposure to Ebola. We further test for baseline balance for the sample of schools for which we have midline and endline data, respectively. As before, there are very few systematic differences across the groups in these schools (Tables A.9 and A.10). Two comparisons are statistically significant at the 10 percent level—but overall this small number is still consistent with pure chance.

While balance at baseline establishes the internal validity of the experimental design,

balance at midline and endline also suggest that school attrition is not selective. We further investigate whether there are systematic differences across groups in whether students in specific grades were administered the learning assessment during the midline and endline surveys. We do this by regressing a zero/one variable indicating whether specific data appear in the survey on indicators for the treatment groups, with the control group being the left out/reference category. The results show that there are no systematic differences across groups, again suggesting the results are not likely to be biased by which data are available for the analysis sample (Table A.8).

In sum, the samples we are using in our analysis are indeed balanced across the randomly assigned groups at baseline. In addition, attrition of schools and student assessments from the sample is not selective with respect to the treatment groups. Together these findings, consistent with the intent of the experimental design, allow us to compare indicators and outcomes at midline and endline, and confidently attribute any differences to the intervention itself.

A.2.2 Investigating Gaming by Teachers

High-stakes accountability schemes can produce perverse effects such as reducing the number of students tested in order to increase performance measures. We do not have data that track individual students over time; however, we examine enrollment numbers to explore whether fewer students are being assessed or enrolled, or if more students drop out, in the treatment groups relative to the control. Table A.11 regresses log number of students assessed for the respective grades at midline and endline on treatment dummies, including strata fixed effects. There is no evidence of a strategic holding back of students for either treatment arm, grade, or year of data collection. In fact, slightly more students were assessed for both grades and rounds of data collection for the In-Kind arm ($p = 0.116$ and 0.726 for grades 3 and 4 at midline; and $p = 0.267$ and 0.111 at endline). The results for the Recognition arm are also always statistically insignificant ($p = 0.349$ and 0.762 for grades 3 and 4 respectively at midline; and $p = 0.757$ and 0.87 at endline). It is not clear why the number of students assessed is relative larger in the In-Kind arm; it could be that the learning gains also reduces dropouts. The number of grade 2 students assessed at midline is also larger by a similar magnitude in the In-Kind treatment, even though grade 2 student scores did not determine pay-outs, suggesting that this pattern is not due to gaming. The increase in the number of students assessed is partly driven by outliers. Table A.13 shows that our main results on student learning are robust to dropping these students from the sample.

Other forms of gaming include cheating and “teaching to the test”. We believe that former is unlikely, given the efforts taken to minimize opportunities for cheating or collusion

(see sub-section 2.1). The latter also seems unlikely, since there was no prior knowledge about the universe of potential test items, and different students received different test booklets. Moreover, “teaching to the test” is not necessarily a bad thing provided that the potential test items are a comprehensive reflection of the learning goals required for a given subject/grade.

A.3 Additional Tables and Figures

Table A.1: Rewards and performance indicator level

Performance indicator value	In-Kind Reward	Recognition Reward
$0.4 < x \leq 1$	One bag of rice and a radio	Bronze certificate and ceremony at the community level
$1 < x \leq 1.6$	One bag of rice, a radio, a cell phone, a TV	Silver certificate and ceremony at the community and prefectural level
$1.6 < x \leq 2.4$	One bag of rice, a radio, a cell phone, a TV	Gold certificate and ceremony at the community, prefectural, and regional level
$x > 2.4$	One bag of rice, a radio, a cell phone, a TV, a generator	Platinum certificate and ceremony at the community, prefectural, regional and national level

Table A.2: Distribution of In-Kind rewards

	Proportion of teachers	Proportion of annual salary
None	36%	0%
Rice & radio	18%	4%
Rice & radio & cellphone	10%	8%
Rice & TV & cellphone	11%	32%
Rice & TV & cellphone & generator	25%	49%

Table A.3: Quality of Implementation (teacher data)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Midline				Endline			
	Control	In-Kind	Recog.	Obs	Control	In-Kind	Recog.	Obs
School participates	0.967	0.972	0.964	1055	0.788	0.918	0.832	1100
<i>Which teachers targeted?</i>								
Only grade 2	0.032	0.051	0.065	999	0.104	0.111	0.115	861
Only grade 3	0.388	0.369	0.376	1016	0.296	0.157	0.254	861
Only grade 4	0.386	0.367	0.387	1023	0.052	0.071	0.062	861
Grades 3 and 4	0.704	0.712	0.753	1023	0.341	0.519	0.384	864
<i>Type of award</i>								
Financial	0.484	0.355	0.462	1000	0.129	0.098	0.174	864
In-Kind	0.721	0.916	0.794	1023	0.414	0.739	0.217	864
Certificate	0.455	0.497	0.620	1005	0.169	0.115	0.504	864
Ceremony	0.204	0.179	0.331	1002	0.064	0.056	0.244	864
<i>Performance metric</i>								
Student test scores	0.756	0.840	0.794	1019	0.863	0.932	0.790	863
Inspection score	0.943	0.956	0.929	1022	0.860	0.921	0.795	865
French evaluated	0.943	0.947	0.969	1017	0.976	0.986	0.964	897
Math evaluated	0.931	0.924	0.959	1017	0.960	0.989	0.957	895
Science evaluated	0.334	0.337	0.376	993	0.195	0.123	0.148	779
<i>Abs vs relative performance</i>								
Relative—Other teachers in school	0.238	0.181	0.184	1027	0.383	0.295	0.304	906
Relative—Other schools	0.158	0.164	0.146	1027	0.199	0.187	0.204	899
<i>Receipt of award</i>								
Teacher received					0.066	0.458	0.504	889
Money					0.012	0.067	0.126	845
In-Kind					0.008	0.387	0.077	868
Certificate					0.008	0.027	0.386	851
Ceremony					0.008	0.030	0.125	847
School received					0.058	0.817	0.754	936

Table A.4: Data collection instruments

Instruments	Sections
Principal questionnaire	A. School and principal identification B. Demographics C. Education and professional training D. Work experience and training needs E. Pedagogical practices and languages F. School basic characteristics G. School environment H. Interaction with colleagues (subordinate, supervisors, etc.) I. Support and monitoring of teachers J. Motivation
Teacher questionnaire	A. Class and teacher identification B. Demographics C. Class characteristics D. Education and professional training E. Work experience and training needs F. Pedagogical practices and languages G. Interaction with colleagues H. Motivation I. Absenteeism and event disturbing teaching J. Remuneration K. Perception of key factors influencing student learning L. Performance recognition or punishment
Student tests	A. Identification of school, class, and teachers B. School-related student characteristics C. Student environmental and familial backgrounds D. French test questions E. Math test questions
Inspection bulletin	A. Class and inspector identification B. Teacher identification C. Summary of scores D. General material and spatial classroom arrangement 1. Lesson 1 – Identification of the lesson 2. Lesson 1 – Teaching and learning material preparation 3. Lesson 1 – Lesson planning (Competency-based approach) 4. Lesson 1 – Delivery of the lesson 5. Lesson 1 – Analysis of own performance 1. Lesson 2 – Identification of the lesson 2. Lesson 2 – Teaching and learning material preparation 3. Lesson 2 – Lesson planning (Competency-based approach) 4. Lesson 2 – Delivery of the lesson 5. Lesson 2 – Analysis of own performance

Table A.5: Sample sizes for different rounds and instruments

Variable	2012 Baseline	2013 Midline	2014 Endline
No of Schools (principal data)	416	403	390
No of Schools (teacher data)	406	387	360
No of Schools (student data)	408	417	391
No of Schools (inspection data)		387	388
No of Teachers (teacher data)	1,165	1,068	1,120
No of Teachers (inspection data)		1,057	1,179
No of Grade 2 Students	19,621	12,327	—
No of Grade 3 Students	18,663	22,053	21,368
No of Grade 4 Students	—	14,139	14,765

Table A.6: Teacher characteristics

Variable	N	(1) Male Mean/SE	N	(2) Female Mean/SE	N	(3) Total Mean/SE	T-test Difference (1)-(2)
Class size	377	43.363 (0.900)	755	52.375 (0.558)	1132	49.374 (0.494)	-9.011***
Permanent	386	0.902 (0.015)	784	0.955 (0.007)	1170	0.938 (0.007)	-0.054***
Insufficient— recognition	385	0.249 (0.022)	784	0.281 (0.016)	1169	0.270 (0.013)	-0.031
Insufficient— salary	385	0.675 (0.024)	784	0.759 (0.015)	1169	0.731 (0.013)	-0.084***
Monthly salary (USD)	340	125.268 (1.254)	743	115.562 (0.598)	1083	118.610 (0.585)	9.706***
Choose teaching	381	0.635 (0.025)	775	0.732 (0.016)	1156	0.700 (0.013)	-0.096***

Notes: The value displayed for t-tests are the differences in the means between male and female teachers, using baseline data. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.7: Baseline balance on key variables

Variable	(1) Control		(2) In-Kind		(3) Recognition		T-test Difference	
	N	Mean	N	Mean	N	Mean	(1)-(2)	(1)-(3)
<i>Student-level (grade 2)</i>								
Math	6466 [112]	0.027 (0.075)	6643 [109]	0.078 (0.067)	6512 [115]	-0.089 (0.078)	-0.051	0.117
Language	6466 [112]	0.032 (0.069)	6643 [109]	0.057 (0.067)	6512 [115]	-0.013 (0.064)	-0.025	0.045
Female	6454 [112]	0.474 (0.008)	6639 [109]	0.477 (0.009)	6495 [115]	0.471 (0.008)	-0.003	0.003
Age	6262 [112]	9.411 (0.067)	6398 [109]	9.561 (0.092)	6299 [115]	9.585 (0.079)	-0.150**	-0.175**
<i>Student-level (grade 3)</i>								
Math	6089 [119]	-0.004 (0.077)	6280 [120]	-0.055 (0.067)	6294 [117]	-0.115 (0.077)	0.051	0.112
Language	6089 [119]	0.035 (0.077)	6280 [120]	-0.007 (0.063)	6294 [117]	-0.091 (0.084)	0.041	0.126
Female	6083 [119]	0.473 (0.009)	6267 [120]	0.472 (0.009)	6290 [117]	0.480 (0.008)	0.001	-0.007
Age	5964 [118]	10.949 (0.071)	6143 [120]	11.071 (0.078)	6173 [117]	11.037 (0.067)	-0.122	-0.088
<i>School-level</i>								
No. grade 2 students	112	57.732 (3.545)	109	60.945 (4.444)	115	56.626 (3.600)	-3.213	1.106
No. grade 3 students	119	51.168 (3.130)	120	52.333 (4.200)	117	53.795 (3.812)	-1.165	-2.627
Headteacher female	138	0.181 (0.033)	139	0.165 (0.032)	139	0.230 (0.036)	0.016	-0.049
Headteacher age	132	48.455 (0.782)	136	47.882 (0.846)	135	46.281 (0.913)	0.572	2.173**

Notes: The value displayed for t-tests are the differences in the means across the groups. Standard errors, in parenthesis are clustered at the school level. Observations with square brackets indicate number of clusters. Strata fixed effects included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.8: Attrition

	Midline			Endline	
	(1) Grade 2	(2) Grade 3	(3) Grade 4	(4) Grade 3	(5) Grade 4
In-Kind	-0.006 (0.053)	-0.042 (0.045)	0.022 (0.044)	-0.031 (0.049)	0.069 (0.044)
Recognition	0.011 (0.054)	-0.051 (0.047)	-0.022 (0.044)	-0.022 (0.048)	-0.011 (0.048)
Control mean	0.691	0.831	0.838	0.772	0.794
Observations	408	408	408	408	408
R-squared	0.129	0.109	0.116	0.170	0.097
Test:In-Kind=Recognition	0.759	0.856	0.291	0.859	0.078

Notes. Each column represents a separate regression, including strata fixed effects. Data is at a schools level, and the dependent variable is an indicator variable equal to one if student learning data is available for a given school in a given grade and a given round of data collection. Standard errors are in parentheses. . $*p < 0.1$, $**p < 0.05$, $***p < 0.01$

Table A.9: Baseline balance, restricted to sample with midline student assessment data

Variable	(1) Control		(2) In-Kind		(3) Recognition		T-test Difference	
	N	Mean	N	Mean	N	Mean	(1)-(2)	(1)-(3)
<i>Student-level (grade 2)</i>								
Math	5528 [84]	0.049 (0.084)	5771 [80]	0.093 (0.075)	5409 [86]	-0.100 (0.088)	-0.044	0.149
Language	5528 [84]	0.050 (0.076)	5771 [80]	0.063 (0.074)	5409 [86]	-0.025 (0.074)	-0.013	0.075
Female	5517 [84]	0.482 (0.008)	5767 [80]	0.484 (0.010)	5400 [86]	0.479 (0.009)	-0.003	0.003
Age	5349 [84]	9.485 (0.071)	5549 [80]	9.618 (0.101)	5255 [86]	9.644 (0.085)	-0.133*	-0.159**
<i>Student-level (grade 3)</i>								
Math	5473 [96]	0.048 (0.082)	5642 [93]	-0.032 (0.071)	5482 [91]	-0.122 (0.085)	0.080	0.171
Language	5473 [96]	0.067 (0.083)	5642 [93]	0.017 (0.067)	5482 [91]	-0.098 (0.094)	0.050	0.165
Female	5468 [96]	0.479 (0.009)	5633 [93]	0.477 (0.010)	5478 [91]	0.478 (0.009)	0.003	0.001
Age	5364 [95]	10.977 (0.075)	5545 [93]	11.112 (0.082)	5373 [91]	11.096 (0.069)	-0.136*	-0.119
<i>School-level</i>								
No. grade 2 students	112	57.732 (3.545)	109	60.945 (4.444)	115	56.626 (3.600)	-3.213	1.106
No. grade 3 students	119	51.168 (3.130)	118	52.907 (4.251)	116	54.112 (3.832)	-1.739	-2.944
Headteacher female	138	0.181 (0.033)	137	0.168 (0.032)	138	0.232 (0.036)	0.013	-0.051
Headteacher age	132	48.455 (0.782)	134	47.754 (0.854)	134	46.388 (0.913)	0.701	2.066**
No. deaths, Ebola	107	14.523 (4.002)	105	11.638 (3.497)	108	10.787 (3.111)	2.885	3.736

Notes: The value displayed for t-tests are the differences in the means across the groups. Standard errors, in parenthesis are clustered at the school level. Observations with square brackets indicate number of clusters. Strata fixed effects included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.10: Baseline balance, restricted to sample with endline student assessment data

Variable	(1) Control		(2) In-Kind		(3) Recognition		T-test Difference	
	N	Mean	N	Mean	N	Mean	(1)-(2)	(1)-(3)
<i>Student-level (grade 2)</i>								
Math	6346 [109]	0.019 (0.076)	6643 [109]	0.078 (0.067)	6311 [112]	-0.109 (0.078)	-0.059	0.129
Language	6346 [109]	0.024 (0.070)	6643 [109]	0.057 (0.067)	6311 [112]	-0.030 (0.064)	-0.033	0.055
Female	6334 [109]	0.476 (0.008)	6639 [109]	0.477 (0.009)	6294 [112]	0.469 (0.008)	-0.001	0.007
Age	6153 [109]	9.423 (0.067)	6398 [109]	9.561 (0.092)	6101 [112]	9.563 (0.075)	-0.138*	-0.140**
<i>Student-level (grade 3)</i>								
Math	5846 [111]	0.018 (0.079)	6129 [113]	-0.050 (0.068)	5932 [107]	-0.117 (0.077)	0.069	0.136
Language	5846 [111]	0.051 (0.079)	6129 [113]	-0.003 (0.063)	5932 [107]	-0.106 (0.085)	0.054	0.157
Female	5840 [111]	0.479 (0.009)	6116 [113]	0.473 (0.009)	5928 [107]	0.479 (0.008)	0.006	0.000
Age	5727 [110]	10.968 (0.071)	5996 [113]	11.082 (0.079)	5815 [107]	11.044 (0.067)	-0.115	-0.076
<i>School-level</i>								
No. grade 2 students	109	58.220 (3.631)	109	60.945 (4.444)	112	56.348 (3.563)	-2.725	1.872
No. grade 3 students	111	52.667 (3.303)	113	54.239 (4.396)	107	55.439 (3.945)	-1.572	-2.773
Headteacher female	129	0.194 (0.035)	132	0.174 (0.033)	127	0.244 (0.038)	0.020	-0.050
Headteacher age	123	49.008 (0.787)	129	47.961 (0.877)	123	46.976 (0.922)	1.047	2.033*
No. deaths, Ebola	98	15.224 (4.347)	101	11.446 (3.624)	99	11.606 (3.381)	3.779	3.618

Notes: The value displayed for t-tests are the differences in the means across the groups. Standard errors, in parenthesis are clustered at the school level. Observations with square brackets indicate number of clusters. Strata fixed effects included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.11: Log number of students assessed— by grade and round of data collection

	Baseline		Midline			Endline	
	(1) Grade 2	(2) Grade 3	(3) Grade 2	(4) Grade 3	(5) Grade 4	(6) Grade 3	(7) Grade 4
In-Kind	0.030 (0.076)	-0.051 (0.072)	0.148 (0.134)	0.145 (0.092)	0.113 (0.101)	0.026 (0.073)	0.106 (0.066)
Recognition	0.004 (0.070)	0.002 (0.068)	0.172 (0.126)	0.093 (0.099)	-0.034 (0.109)	-0.020 (0.064)	0.012 (0.071)
Control mean	57.732	51.168	39.835	57.888	52.508	43.269	41.387
Observations	336	356	289	336	350	317	340
R-squared	0.340	0.399	0.330	0.406	0.388	0.412	0.433
Test:In-Kind=Recognition	0.720	0.458	0.846	0.601	0.196	0.539	0.168

Notes. Each column represents a separate regression, including strata fixed effects. The dependent variable is the log of the the total number of students assessed for a given school and grade, and for a given round of data collection. Data is at a school level. Standard errors are in parentheses. . $*p < 0.1$, $**p < 0.05$, $***p < 0.01$

Table A.12: Breakdown of per student costs by treatment arm

	In-kind	Recognition
Data Collection	\$ 5.57	\$ 5.57
Award Ceremonies	N/A	\$ 2.72
In-kind Rewards	\$ 6.18	N/A
Total	\$ 11.75	\$ 8.29

Notes. Costs are in USD, 2014 prices. Data collection costs are from 2014 (endline), when a random sample of students in every school were assessed . Implementation costs data are from 2013 (midline) but adjusting for one year of inflation (the award ceremonies did not take place in 2014). We do not include the costs of the communication campaign conducted prior to the start of the program, since this is a fixed cost and future communication can take place in combination with the award ceremonies. The schools in our sample have on average 77 students enrolled in both grade 3 and grade 4.

Table A.13: Main impacts on learning, dropping outliers

	Full sample		Grade 3		Grade 4		French		Math	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Mid	End	Mid	End	Mid	End	Mid	End	Mid	End
In-Kind	0.239*** (0.084)	0.156 (0.099)	0.129 (0.097)	0.121 (0.108)	0.235*** (0.090)	0.184* (0.109)	0.202** (0.082)	0.183* (0.095)	0.274*** (0.093)	0.130 (0.108)
Recognition	0.125 (0.087)	0.088 (0.103)	-0.037 (0.094)	0.053 (0.114)	0.224** (0.094)	0.130 (0.115)	0.102 (0.084)	0.118 (0.098)	0.146 (0.097)	0.060 (0.113)
Control mean	0.031	0.058	0.043	0.044	0.017	0.040	0.035	0.035	-0.004	-0.004
Observations	86840	57808	44105	28278	42735	29530	43421	28904	43419	28904
R-squared	0.097	0.121	0.124	0.121	0.084	0.109	0.109	0.153	0.097	0.102
Test:In-Kind=Recognition	0.205	0.514	0.092	0.577	0.908	0.647	0.256	0.511	0.199	0.543

Notes. See Table 1. Students in grades with more than 300 students are dropped from the sample.

Table A.14: Impacts on learning— by exposure to (log) Ebola

	Midline	Endline		
	(1) Full	(2) Full	(3) Grade 3	(4) Grade 4
In-Kind	0.258*** (0.099)	0.285** (0.117)	0.132 (0.144)	0.386*** (0.141)
Recognition	0.066 (0.104)	0.223* (0.124)	0.142 (0.157)	0.290* (0.150)
Ebola (invhs)	-0.053 (0.038)	-0.023 (0.042)	-0.017 (0.048)	-0.047 (0.052)
Ebola (invhs) x In-Kind	-0.013 (0.052)	-0.112** (0.052)	-0.050 (0.062)	-0.170** (0.066)
Ebola (invhs) x Recognition	0.060 (0.050)	-0.103* (0.054)	-0.066 (0.061)	-0.140** (0.067)
Observations	86578	57444	28146	29298
R-squared	0.128	0.144	0.144	0.168

Notes. See Table 5. The variable “Ebola (invs)” is the Inverse Hyperbolic Sin transformation of the total number of Ebola cases, by prefecture. This transformation approximates the natural logarithm.

Figure A.1: Timeline

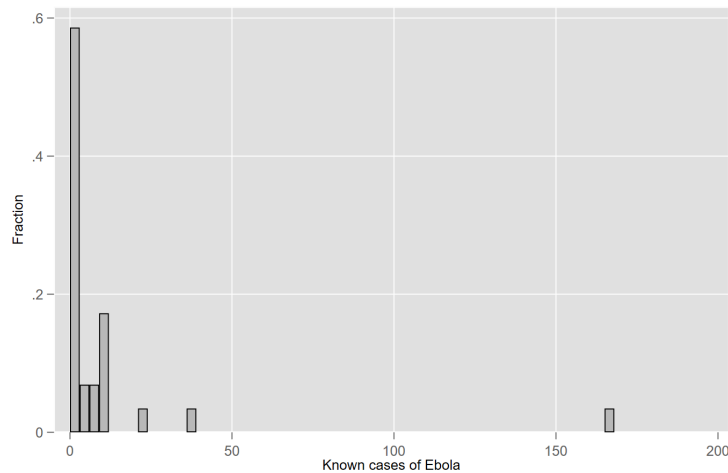
	2012						2013					2014		
	Jan-Apr	May	Jun-Sep	Oct	Nov	Dec	Jan-Apr	May	Aug-Sept	Oct-Nov	Dec	Jan	Feb-Apr	May
Data collection														
Pilot														
Student assessment		2,3						2,3,4						3,4
Inspection														
Teacher questionnaire														
Implementation														
Communication campaign														
Reward distribution														
School year				Start				End		Start				End
Ebola														

Table A.15: Impacts on learning— by exposure by Ebola (reduced sample)

	Midline		Endline		Endline Grade 3		Endline Grade 4	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
In-Kind	0.208** (0.097)	0.198* (0.102)	0.124 (0.107)	0.195* (0.114)	0.014 (0.118)	0.072 (0.125)	0.207 (0.139)	0.263* (0.144)
Recognition	0.182* (0.099)	0.171 (0.107)	0.020 (0.112)	0.103 (0.121)	-0.003 (0.133)	0.068 (0.143)	0.087 (0.140)	0.162 (0.148)
Ebola	-0.000 (0.001)	-0.001 (0.002)	-0.003*** (0.001)	-0.000 (0.002)		-0.000 (0.002)		-0.001 (0.002)
Ebola x In-Kind		0.001 (0.003)		-0.005** (0.002)		-0.004 (0.003)		-0.007** (0.003)
Ebola x Recognition		0.001 (0.002)		-0.006*** (0.002)		-0.005** (0.002)		-0.008** (0.003)
Observations	72452	72452	47592	47592	23470	23470	24122	24122
R-squared	0.128	0.129	0.173	0.181	0.196	0.211	0.172	0.190

Notes. See Table 5. Prefectures that did not report any cases of Ebola by the time of endline data collection (May 2014) are excluded from the sample.

Figure A.2: Histogram of known cases of Ebola by May 2014



Note. Data at a prefecture level for the total reported number of known cases by May 2014. The data are compiled from the World Health Organization Africa's situation reports. Data for total number of cases by prefecture in Guinea can be found [here](#):