RBF interventions in education: do they increase inequality of outcomes?

A literature review

This report has been authored by Youdi Schipper and Menno Pradhan. Daniel Rodriguez Segura provided excellent research support. We thank Christina Brown, who responded to our request for additional analysis based on her study. We also thank three anonymous reviewers and Jessica Lee from the World Bank for their comments on an e arlier draft of this report.

aighd | AMSTERDAM INSTITUTE FOR GLOBAL HEALTH & DEVELOPMENT

WORLD BANK GROUP

1818 H Street NW
Washington DC 20433
Internet: www.worldbank.org/reach

Executive Summary

This report reviews high quality impact evaluations of results-based financing interventions (RBF) in education to investigate whether particular subgroups of students benefit systematically more (or less) from RBF interventions than others. Results-based financing in education provides incentives for bureaucrats, teachers and/or students that depend on achieved education results. In this way, they aim to contribute to improving learning for students who are in school. A potential concern with this approach is that actors, while trying to maximize their incentives, focus their attention on high-performing students, resulting in differential impacts of these programs among subgroups. This report thus aims to answer the following research questions: 1) What are the estimated impacts of RBF mechanisms in education on the outcomes for different groups? 2) What design and implementation characteristics explain differences in the impact between subgroups? 3) How can RBF be used to support educational attainment and outcomes of specific (marginalized) target groups?

The review is based on 23 studies with a (quasi) experimental design, ranging in publication date from 2009-2021, covering 12 different countries in 4 regions: 8 studies were conducted in Sub-Saharan Africa, 4 in Latin America, 7 in South Asia, and 4 in South East- and East-Asia). 10 studies evaluate a Government program, 7 a program implemented by an NGO and 6 a program that was designed by researchers for the purpose of the study. Most use student learning as the primary outcome measure (21 studies) and apply randomized controlled trial (RCT) as its main methodological approach (18 studies). The studies apply a wide variety of incentive designs, indicating that there is not yet a best practice for RBF in education which is adopted on a large scale. Most studies (12) estimate effects of incentives that target individual teachers; 8 studies provide estimates for incentive programs that target groups of teachers of bureaucrats; and 6 studies evaluate student incentive programs.

The report finds the following answers to the research questions. Overall, the RBF programs in the studies we reviewed resulted in few statistically significant differential subgroup effects. In other words, most of these programs do not significantly increase or decrease inequality of pre-existing student learning outcomes. We find this absence of significant effect heterogeneity across the three incentive levels in our study: groups, teachers and students.

The absence of differential subgroup effects is noteworthy considering the recent efforts implementors have put into lowering the risk that certain subgroups benefit more from the intervention than others. Consider, for instance, the classic threshold design. Threshold designs base the incentive on the number of students that pass a test. One would expect teachers to devote more attention to students close to the passing threshold. The pay for percentile design is an example of a recent innovation designed to avoid these types of problems. Here the incentive is based on the rank order of the students' test scores when compared with students of similar ability. We would expect teachers to equally share their attention across all students, and as a result no subgroup heterogeneity. This review finds that neither incentive design results in significant subgroup heterogeneity.

We do find several *non-significant* subgroup impact estimates that indicate that students with higher baseline test scores benefit more from RBF interventions than those with low baseline test scores. Across all interventions, we found 37 positive and 18 negative test score subgroup point estimates.

Our review contains a few examples of programs that incentivize teachers to devote particular attention to less advantaged groups, with encouraging results. At group level, Lautharte and co-authors (2021) describe a reform that rewards basic literacy outcomes, with results suggesting a more equitable distribution of learning impacts across grades and subjects. Chang et al. (2020) show that a pay for percentile treatment that increases the rewards for the lowest scoring students yielded more pro-poor learning results. We conclude that RBF programs can offer flexibility to focus incentives on learning outcomes among the weakest performing students. These programs therefore have the potential to redress learning inequalities in addition to improving mean test scores.

Based on the findings in this review, equity of impacts does not appear to be a major concern for RBF programs. Distribution of learning effects across initial levels, gender and wealth should be included in future impact studies, but the empirical evidence so far does not indicate this to be the most urgent policy priority.

Prioritising and scaling incentives in education systems appears to be a more pressing problem. The relatively small number of at scale programs in our review suggests that a first-order question is how to scale RBF that deliver incentives to schools and individual teachers for improving student learning outcomes. This question includes a range of issues, including the organisation of frequent and independent student testing, generating political will and implementation capacity.

Finally, a few studies show it is possible to implement RBF programs that incentivise learning of low-performing students. This evidence points to a pro-poor flexibility of RBF that has not been used much so far and presents an area for future research.

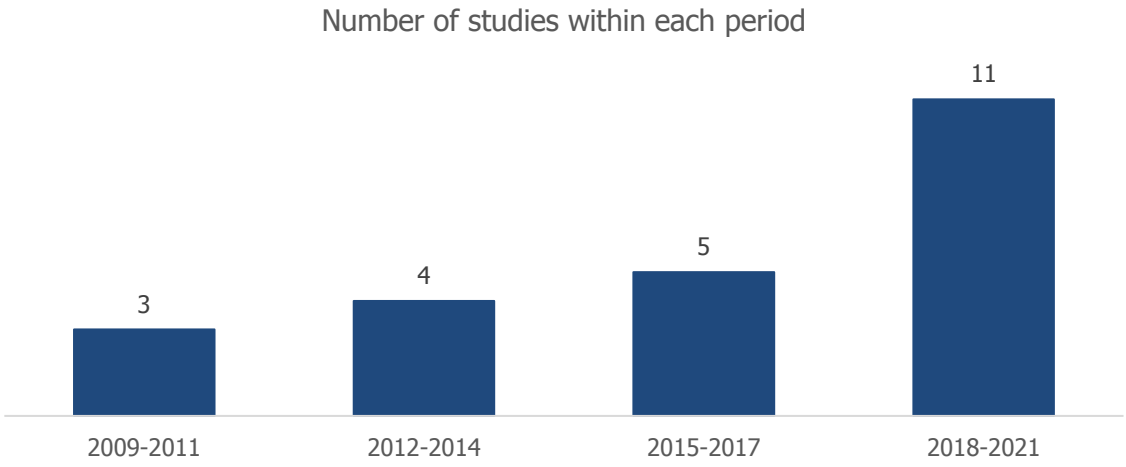
1. Introduction

1.1. Background and motivation

A general definition of results-based financing (RBF) is "... any program or intervention that provides rewards after the credible verification of an achieved result." (World Bank, 2018). A broad range of arrangements fits under this definition, from macro-level programs that define a "result" as having implemented a program (regardless of effect on outcomes) to school or teacher level programs where results are defined strictly in terms of outcomes of interest (desirable education outcomes). The idea of RBF is to formulate outcomes for acceptable goals and create policy support for the goals ex-ante. Financing is conditioned on reaching these goals.

Recently, results-based financing has been championed as a way to increase the efficiency and effectiveness of aid, and as a way to speed up progress towards the sustainable development goals. This has resulted in increased levels of spending being committed to RBF programs. Over the period 2010–2015, RBF has grown to about US$2.5 billion, or 20 percent of the World Bank Group's total investments in education. In 2015, World Bank President Kim announced that the WBG would double results-based financing to approximately US$5 billion over the period 2015–2020.

The growth of donor interest in RBF has been accompanied by an increase in high-quality studies on the effectiveness of a range of interventions under the RBF umbrella. We illustrate this using the publication dates of the studies used in our review:



Many recent reviews on the effectiveness of education interventions in low-income countries discuss (micro level) RBF interventions; see for example Glewwe and Muralidharan (2016). There are detailed analyses of RBF initiatives in studies on school governance and accountability, see for example Bruns, Filmer and Patrinos (2011); and Bruns and Luque (2015). World Bank (2018) provides a more recent review on the effectiveness of RBF instruments.

However, a common concern with the use of results-based financing mechanisms is that they can have detrimental impacts on the equity of education outcomes. For example, in the pursuit of higher financial reward, performance incentives have the potential to encourage teachers to exert more effort on better performing students and neglect disadvantaged students. In fact, equity concerns are a recurring theme in RBF evaluation studies. Most authors are well aware of the such concerns and many studies mention specific design characteristics that were introduced or tweaked in order reduce the possibility that the proposed incentive system will disproportionately benefit stronger students or schools; or may focus teacher effort on a narrow part of the student distribution. These concerns inspired the following research questions.

1.2. Research questions

This review aims to address the three following questions:

1. What are the estimated impacts of RBF mechanisms (in education) on the outcomes for different groups? For example, how do the impacts on student learning outcomes of teacher incentive schemes differ between low and high performing students; or between students from poor and wealthy households?

2. What design and implementation characteristics explain differences in the impact between subgroups? Where similar actors are incentivized, what explains differences in the impact on the equity of outcomes? Is there anything in the design of RBF interventions that can be linked to more pro-poor effects?

3. How can RBF be used to support educational attainment and outcomes of specific (marginalized) target groups (such as girls or poor households)?

Approach

The studies in this review cover a wide variety of contexts, including different school and governance types, learning levels, and a range of RBF instruments. This creates a fundamental challenge for analysts looking to generalize findings across studies, even if these are well-identified. As Glewwe and Muralidharan (2016) argue: "Take the case of teacher performance pay. While there are multiple high-quality studies on the subject, no two studies have the same formula for how teachers will be paid bonuses! Some of the design details that vary include individual versus group incentives, tournaments versus piece rates, linear versus nonlinear bonus formulae, formulae based on students reaching thresholds (such as the fraction who pass a test) versus those that reward improvements for all students."

Variation in context and implementation management, added to this design variation, makes generalizing across findings hard. This issue is compounded by the fact that a) we have a relatively small set of studies and b) we are interested in "second-order" effects in specific parts of the beneficiary population. In addition, studies differ in how they report heterogeneity effects: some use graphs, some use estimates by quantile, some interact quantile indicators with treatment. This means that, different from reviews that focus on the central impact measure, this review lacks a homogenous outcome that can compared across studies. For these reasons, we will largely abstain from meta-analytic (quantitative) summaries of the evidence, except for some counts of estimate types. Our focus is on a more descriptive, narrative review of the evidence.

1.3. Organization

The review is organized as follows. In Section 2, we discuss our study sample selection and narrow down the set of studies that we review in detail. We provide detailed summaries of the final selection of studies in the Appendix B. In Section 3, we describe the shortlisted studies for our review. In Section

4, we analyze the subgroup effects reported in the literature; a list of subgroup estimates is provided in Appendix C. Section 5 concludes.

The analysis in Section 4 is organized by incentive level. We distinguish three levels in our review: (i) Incentives based on group level metrics; (ii) incentives offered to teachers, based on individual performance metrics; (iii) incentives offered to students. A key point is that these levels are defined by the level at which performance is measured. This is not always the level at which the incentive is received, for example, if a teacher is individually paid for the performance of a team of teachers; this case would be under "groups level incentives".[1]

2. Sample Selection

This study documents and analyses mean and sub-group impacts of RBF interventions. As such, we must build a core set of RBF interventions and evaluations for our review. The scope of the RBF literature is wide, in terms of intervention types, methodologies used and geography, so we need a methodical approach to finding and filtering studies.

2.1. Studies search and database

First, we need an extensive database from which we can select the most appropriate studies for this review. Our search for relevant studies took place in four steps: (1) using existing literature reviews; (2) forward and backward tracing of references from/to the relevant papers in step one; (3) adding studies based on keyword searches in digital libraries for published (peer-reviewed) work; and (4) searching in libraries containing working papers.

Our starting point was the reference list in Evans and Popova (2016). Evans and Popova is a review of six earlier, highly cited and comprehensive reviews of impact studies of education interventions that (aim to) improve access to education and learning. The reviews included in Evans and Popova have a much wider scope than our review in terms of interventions, but they use the same country focus (low and middle-income countries, or Africa) and also include only experimental and quasi-experimental studies. Using this list of included studies, we reviewed abstracts to select all studies that fall within the broad definition of RBF, that is: any program or intervention that provides rewards to (groups of) individuals after the credible verification of an achieved result. We cross-referenced this list with other recent reviews of (RBF) studies including Lee et al., (2018), Pham et al. (2020), Alger et al. (2014), Rodriguez-Segura (2021).

As a second step, we look for any additional studies through reference tracing. We backward-traced papers through the literature review sections of these papers, and the papers that they cite. We then

1 We exclude country level RBF, mainly because the existing studies are descriptive, without clear counterfactuals.

forward-traced, i.e., searched other papers that cited these studies, each of these papers through the Google Scholar feature for this process ("Cited by"). After completing this process, we iterated through the process of back- and forward-tracing papers until no additional papers were located.

The third round of searches was within repositories of peer-reviewed journals and databases such as EconLit, EconPapers, and Google Scholar, where multiple combinations of words related to the scope of this review were searched. Finally, we looked in the working paper repositories of well-known organizations that routinely produce education-related research as the World Bank, the Interamerican Development Bank, the EdTech Hub, NBER, the RISE Programme, Annenberg Institute, J-PAL, and IPA. All these papers were then forward- and backward-traced to ensure no other relevant paper was excluded.

In all, these steps yielded a list of 662 potential papers for inclusion. While there is no guarantee that all studies that meet the four main criteria are included in the set of core studies, great lengths were covered to ensure that the review was as extensive as possible.

2.2. Filtering criteria

First, we only include studies for low- and middle-income countries. Second, we only include studies with high-quality evaluation designs or counterfactuals. That is, we only include original studies that use experimental or quasi-experimental designs (RCT, regression-discontinuity, or double difference designs).

Third, we include studies of interventions that target groups, teachers or parents/students with a performance incentive that aim to improve student learning through increased effort of bureaucrats, teachers and/or students. The interventions can be aimed at the education supply-side, schools and teachers, or students and their parents on the demand side. However, we exclude interventions that only incentivize school enrolment or attendance, such as conditional cash transfers (CCT).

We have a few reasons to exclude CCTs: first, a large number of reviews of this literature exist, including García and Saavedra (2017), Bastagli et al. (2016), Baird et al. (2013), Baird et al. (2018), Das et al. (2005). There is also evidence on differential sub-group effects, e.g. in Bastagli et al. (2016). More importantly, CCTs are designed to improve the position of vulnerable groups. In particular, they intend to improve welfare and school participation of children in poor households which means that - by design - the concerns about adverse effects of CCTs on vulnerable groups are less important, compared to other RBF interventions. Indeed, the literature shows that in most cases CCTs improve education opportunities for vulnerable groups. Third, the design of CCTs is relatively simple, compared to teacher incentive schemes. In most cases, every beneficiary receives the same amount based upon verification of (minimal) attendance. For the purposes of this study, it seems a logical choice to focus on the non-CCT interventions and the more complex design choices and empirical evidence these provide, especially in light of the current lack of aggregated evidence on the effectiveness of these.

After applying our filtering criteria on the more extensive database of studies, we are left with 30 potential studies. After reviewing these studies in detail, we reduced our set to a shortlist of 23 studies, as explained in the next section. Both lists are included separately in the Appendix A, where we also discuss the motivation for removing 7 papers from the longlist.

Describing the sample of studies

Our final shortlist consists of 23 RBF studies. Appendix C lists the main characteristics of these studies. These studies range in publication date from 2009-2021, cover 12 different countries and 4 regions: 8 studies were conducted in Sub-Saharan Africa, 4 in Latin America, 7 in South Asia, and 4 in South East- and East-Asia). 18 of the studies have a randomized controlled trial (RCT) as its main methodological approach.

Table 1 and 2 show the distribution of treatments in our core set of RBF studies along the lines of important program attributes. Note that the numbers in the following tables do not necessarily add up to 23 as some studies included multiple treatments.

Table 1: Number of studies by outcome measured

Outcomes measured	Number of studies
Learning	21
Student enrollment, participation and completion (all in combination with learning)	5
Teacher attendance	2
Teacher performance more generally	3
Others (e.g., overall literacy rate, use of a technological platform, level of student-level poverty)	2

Table 2: Number of studies by type of organization implementing the program

Type of organization implementing program	Number of studies
Government /private network of schools	10
NGO/Private org/foundation	7
Research team	6

The studies in our sample represent a broad range of incentive programs. These programs differ in many ways, but a central feature is how the performance metric is structured. We distinguish between four broad types of incentives, by metric:

1. Rewards based on current learning, taking pre-treatment learning into account
2. Rewards based on current learning only
3. Rewards based on learning and other indicators (inputs, enrolment)
4. Rewards based on non-learning indicators: effort and inputs

These four incentive designs cover our set of RBF studies. A second dimension to order our studies is the target level of the incentive; here we distinguish three levels: groups/organisations, individual teachers and students. We thus arrive at a typology of studies in Table 3, where columns represent design type and rows reflect target level. The table shows where the studies in our sample fit in terms of these design characteristics and provides the structure for our discussion of subgroup effects.

Table 3: Studies by RBF level

Performance measured at level of:	Reward based on:			
	Learning		Mixed (Learning, enrolment, inputs/effort, other)	Effort or Inputs
	Dependent on gains with respect to initial level	Dependent on Absolute level		
Groups Schools, Bureaucrats, Teacher groups	Al-Samarrai ea 2018 Behrman ea 2015 Glewwe ea 2010 Muralidharan 2012*		Lautharte ea 2021 Contreras and Rau 2012 Ferraz and Perreira 2016 Barrera-Osorio and Raju 2017	

Teachers As individuals	Gilligan ea 2019 Loyalka ea 2019 Chang ea 2020 Mbiti 2019b Andrabi and Brown 2020 Muralidharan 2012* Behrman ea 2015	Mbiti 2019a Loyalka ea 2019 Mbiti 2019b	Andrabi and Brown 2020	Duflo ea 2012 Gaduh ea 2021
Students	Behrman ea 2015 Berry 2015	Blimpo 2014 Filmer et al. 2020 Hirshleifer 2016 Kremer ea 2009		

Notes: *From Muralidharan 2012 we use estimates for five years that include the estimates from Muralidharan and Sundararaman 2011; we refer to all these results as Muralidharan 2012.

3. Subgroup effects in RBF impact studies

The goal of this study is to document, within studies, whether there are any differential RBF effects by sub-groups relative to mean effects (i.e., "effect heterogeneity"). More specifically, we are interested when estimating:

$Y = a + b_1(\text{Treatment}) + b_2(\text{covariate}) + b_3(\text{Treatment*covariate}) + c\,X + e$

whether b_3 is significant when using covariates like initial achievement, gender and wealth. The aim is to document these subgroup effects and, if we find them, to discuss mechanisms generating these effects, possibly relating them to intervention design features.

The way in which different studies report heterogeneous treatment effects varies. Some estimate regressions like the one above, others report separate estimates by subgroups, other present impact

graphs on a continuous scale. Effects have been categorized as zero if the difference between the subgroups could have been due to rounding error; when more than 2 groups have been distinguished and the pattern in sub-estimates in not monotonous; or when no interaction estimate was provided but the authors in words described the effect to be absent or zero. For the second case, if a higher impact is found for the medium ability group as compared to the low and high ability group, we would classify this as a zero heterogenous treatment effect. When the paper reports separate estimates of subgroups, we conducted a two sample T-test to assess whether the difference between the lowest and highest group was significantly different from zero.

Table 4 below shows for how many of the 23 studies in our sample we were able to analyze heterogeneous treatment effects. In most cases, these were reported in the papers. In other cases, the authors generated the table at our request or we have generated the tables ourselves based on the replication data.

Table 4: Number of studies reporting subgroup effects by gender, SES and baseline score

Sub-group	Number of studies that reports sub-group
By gender	12
By SES	3
By baseline score	22

This table shows that subgroup effects are reported for a minority subset of studies, and mainly for gender and baseline score. Few studies report heterogeneous treatment effects by socio economic status.

To get a bird's-eye overview of the heterogenous treatment effects, we tabulated for each of interventions included in the 23 studies whether the heterogenous treatment effects was negative and significant, negative but not significant, zero, positive and not significant, or positive and significant. We counted each intervention, for each year that impact estimates are available, as a separate intervention. This brings us to 94 interventions.

Figure 1 shows the distribution of heterogeneous treatment effects with respect to gender for all interventions considered. There are few significant effects. Boys benefit slightly more often more than girls. But for girls the effects are more often significantly different from zero. Figure 2 shows the effect by wealth. There is very little we can say as this information is seldomly available from students. Figure 3 shows the distribution by baseline score. This is available for almost all studies. Also, here the effects are mostly insignificant. The point estimates indicate that students with higher ability levels at baseline often benefit more from the interventions when compared to those with low baseline ability.

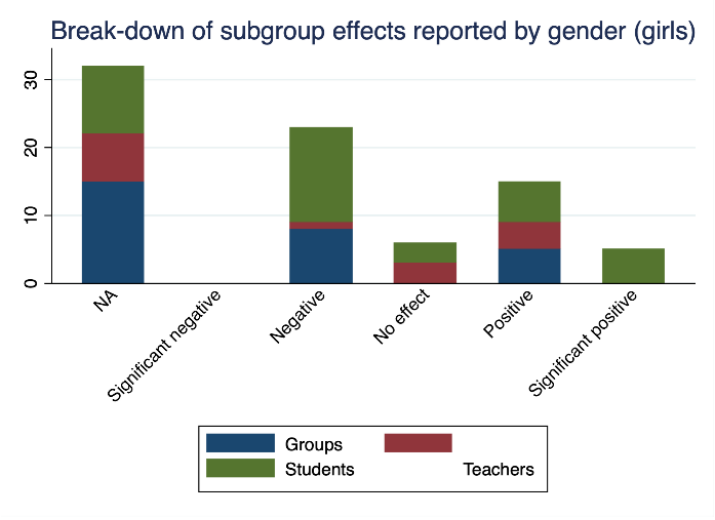


Figure 1: Heterogeneous treatment effects by gender

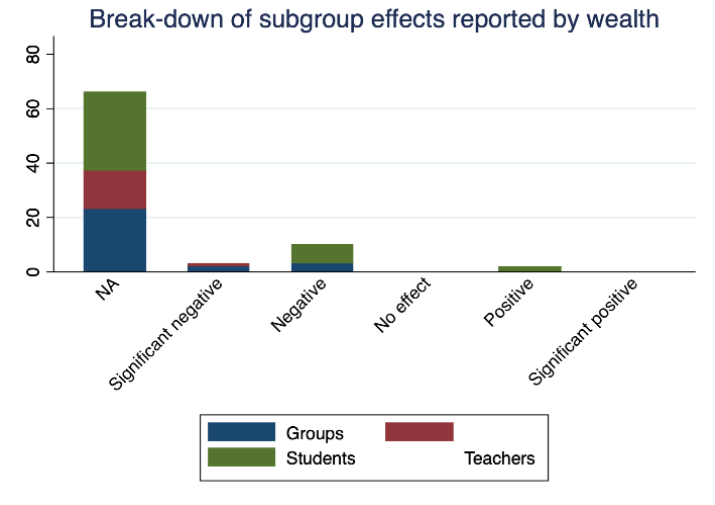


Figure 2 Heterogeneous treatment effects by wealth

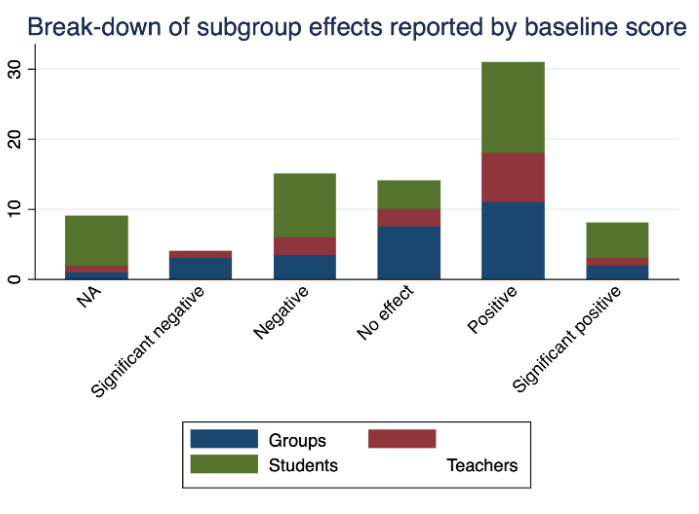


Figure 3 Heterogeneous treatment effects by baseline score

The remainder of this section contains a discussion on how performance incentives could affect different subgroups in different ways. The aim of the discussion is to inform the TOR research question (see Section 1): *Where similar actors are incentivized, what explains differences in the impact on the equity of outcomes?* The RBF program effects are discussed by incentive level: incentives based on group performance, incentives tied to (individual) teacher performance, and incentives linked to student performance.

3.1. Group-level incentives

The eight papers in this subsection study a diverse group of incentive programs; these programs have in common that the incentives are exclusively tied to performance indicators that are measured at the level of a group; for example, programs that offer rewards to all teachers in a school, based on school level learning outcomes. The papers present results from 10 different programs or experiment arms.[2] We provide summaries of these studies in appendix C, with details of the subgroup estimates.[3]

Table 5 provides an overview of the studies included in this section of the review. All studies were published in or after 2010. They study four programs from Latin-America, three from Asia and one from Africa. Some programs were implemented in both secondary and primary schools, some in just one of these two.

[2] We count the primary and secondary school implementation in Al-Samarrai ea (2018) as one program.

[3] Summaries for three studies that also include (experimental arms with) individual teacher incentives can be found under the teacher study summaries: these are Behrman et al., Filmer et al, 2020; and Muralidharan 2012.

Table 5: Group level - Included studies

Study	Country	Incentive type	Level
Al-Samarrai ea 2018	Indonesia	Performance-based grants, competition; top 25% win	Primary and Secondary
Behrman ea 2015	Mexico	Fixed amounts for discrete improvements, (individual and group)	Secondary
Barrera and Raju 2017	Pakistan	Fixed amounts for continuous improvements (test scores, enrolment)	Primary
Contreras and Rau 2012	Chile	Competition on mixed performance formula; top 25% win	Primary and Secondary
Ferraz and Pereira 2016	Brazil	Competition on school target achievement, with 50% threshold	Primary and Secondary
Glewwe ea 2010	Kenya	Teacher group competition, 24/50 win	Primary
Lautharte ea 2020	Brazil	Competition for municipality tax allocations, continuous	Primary and Secondary
Muralidharan 2012	India	Fixed amount for school level test score improvements	Primary

Notes: We refer to a few group level incentive treatments from papers with multiple treatment arms. In this section, we include the third treatment arm (T3) in Behrman ea, where incentives are offered to students, teachers and administrators and the rewards depend partly on peer performance. We discuss the pooled estimates in Barrera-Osorio and Raju (2017), which combine head teacher and teacher group level incentives. We include the teacher group incentive arm from Muralidharan (2012). From Muralidharan 2012 we use estimates for five years that include the estimates from Muralidharan and Sundararaman 2011; we refer to all these results as Muralidharan 2012.

The programs differ in scale along this dimension: the four programs that are implemented in both primary and lower secondary schools - Indonesia, Chile, and both programs in Brazil - are government programs implemented at regional scale, each implemented in populations of 950 schools or more. None of the corresponding evaluation studies has a randomised design. Column three shows that these four (large) programs each have some type of competition in their incentive design. The smaller scale studies are randomized experiments that focus on either primary or secondary schools. Three of these offer fixed amounts for improvements, while only the Kenya experiment (Glewwe et al., 2010) is a competition.

Estimate counts

We summarise the numbers of estimates for these studies in Table 6. The studies vary in the number of years data were collected, the number of impact estimates provided and how detailed both mean effects and heterogeneity were studied and reported.

The first column, effect category, has five labels (apart from Not Available). When direct estimates of significance are available, e.g., for all mean effects and for a number of interaction effects, the coding is straightforward. In some cases, papers model subgroup effects by interacting quantile dummies with treatment, or by estimating regressions separately by quantile. In these instances, we test the difference between the interaction coefficients using a t-test, in order to code the interaction estimate. In a few cases where we had to infer the interaction sign from a graph, we put the result in the non-significant category. (This means that, for these cases, we may underestimate the number of significant (positive) estimates).

To reduce the weight of studies that provide many disaggregated results, we assign a weight of 0.5 to estimates for individual test subjects. All other estimates have a weight of one. The sum of weights for these estimates is 28. The counts in the table are all weighted.

The number of estimates with a significant (at least at the 10 percent level) positive *mean effect* on learning is 18.5. There are two mean impact estimates that are negative (both from the Jakarta school grants study), and one of these is significant. The number of positive but not significant mean impact estimates is 7.5. Most of the interaction estimates relate to baseline test score. Very few of the interaction effects are significantly different from zero. The largest single category is positive interaction effects between treatment and baseline score. Positive test score interactions are more common than negative ones.

Interactions between gender (girls) and treatment are reported less frequently. Of the 13 estimates we have, 8 are negative and 5 are positive. Very few wealth studies reported wealth interactions; all five estimates come from one study, Muralidharan (2012), all of them negative. In the remainder of this section, we focus on the test score interaction estimates.

Table 6: Group level: summary of estimates

Effect category	Mean effect on learning	Interactions		
		Test score	Girls	Wealth
NA		1	15	23
Significant negative	1			2
Negative	1	5.5	8	3
No effect		8.5		

Positive	7.5	13	5	
Significant positive	18.5			
Total	28	28	28	28

Not surprisingly, programs that have no significant positive mean impact are less likely to increase inequality. One program that reports a negative ("progressive") effect is the Jakarta schools grant program in primary schools. This program had a positive impact among quartile 1 schools (in terms of pre-treatment exam scores), but a negative effect in quartiles 3 and 4; and a negative mean effect in one of the three implementation years. Programs with significant and positive mean effects sometimes cause a decrease in inequality: Muralidharan (2012) presents negative wealth interaction effects in all five years. They are significant in two years and in both these years the mean program effects were significantly positive. The sign of the (insignificant) test score interactions alternates over years.

Two program evaluations, Barrera-Osorio and Raju (2017) and Glewwe et al. (2010), find that there is no clear pattern of program impact across the test score distribution. In both cases the mean treatment effect estimates are (partly) insignificant too. Glewwe et al. evaluate an experimental NGO program that provided in-kind incentives to teacher groups based on test scores, in a sample of 100 schools (50 treated) in two districts in Western Kenya. The study finds no treatment effects on non-incentivized (NGO) exams, and positive effects on the incentivized tests (part of the bonus formula).

The other RBF programs in this category combine a significant positive mean impact with a mostly positive (but non-significant) interaction with baseline test scores. There are five of these programs in our selection: the secondary school performance grant competition in Jakarta (Al-Samarrai et al., 2018); the Chilean National System of School Performance Assessment (SNED, Contreras and Rau, 2012); the Pernambuco (Brazil) state teacher bonus program (Ferraz and Pereira, 2016); and the two first versions of the Ceará (Brazil) state tax reform that links tax allocations to municipalities to education outcomes (Lautharte et al., 2021).

What do these programs have in common? All these programs are government implemented, relatively large-scale programs, typically at the level of a state/province or city. Consistent with this scale, the evaluation studies typically do not have randomized designs and have large sample sizes: for instance, Contreras and Rau study the National System of School Performance Assessment (SNED), a national program covering all public or publicly subsidized schools, and the study uses data on 8044 schools.

The programs all introduce performance incentives based on existing administrative data and within a financial system that allocates resources from central to local institutions. All are hybrid designs that use several variables in the reward formula, typically including both the change in learning and the

level of learning. The large scale of these Government programs requires that performance is measured at an aggregate (school) level rather than at the level of the individual teacher[4].

The Ceará design changes: distributional effects

The most interesting design variation comes from the Lautharte et al. evaluation of the Ceará state tax reform.

The study analyses three RBF policy reforms: the first is the introduction of the basic RBF design in 2008.[5] The basic reform stated that for each municipality, a large share (18/25) of the "Quota Parte" state consumption tax allocation would be determined based on an education performance metric. The second is the addition of a technical assistance package on top of the RBF scheme, (2009 for 5th grade, 2009-2011 for 9th grade); the third is a reformulation of the RBF formula to "penalize municipalities with higher percentages of students performing below pre-defined minimal thresholds on test scores". In effect, the last reform specifically increases incentives in the lower tail part of the distribution by giving more weight to municipalities who manage to increase performance among students who are lagging behind.

The study uses quantile graphs to analyse how the three program versions affect learning across the distribution and provides evidence that suggests striking distributional changes linked to changes in program design. The authors conclude that the basic RBF reform increased learning gaps, with effects close to zero for low performing students, especially in Grade 5 (the lowest grade analysed). Second, they conclude that providing non-mandatory technical assistance substantially increased the mean learning effects and improved learning in the lower tail (see the paper for details on the technical assistance), but did not reverse the widening of learning gaps: indeed, the quantile graphs are still mostly increasing.

The third reform included a change in the incentive formula with an explicit distributional focus. It inserted the proportion of illiterate and partially literate students into the formula, thereby allocating more resources to "(…) municipalities seeing improvements in students who are lagging behind" (page 10). The impact estimates for this last reform show strong mean effects (on top of the previous

[4] Individual performance metrics avoid free-rider problems and are typically linked to student performance under the direct control of the beneficiaries. While such designs may elicit stronger incentives, they have more stringent data collection and management requirements. For this reason, large scale programs tend to provide group level incentives.

[5] As described in the appendix, in the basic design, the education index is a function of two indices, Literacy (weight ⅔) and Learning. The literacy index corrects for literacy level variation (punishing unequal literacy performance) and both the literacy and the learning index reflect the level and the change in learning.

versions, see Table 2) but also a clear change in the distribution of effects. Figure 3 shows much flatter and partly decreasing quantile graphs, especially in grade 9.[6]

The last two reforms of the Ceará RBF instrument are instructive for our review. The addition of technical assistance clearly complemented the RBF instrument. Interestingly, the paper provides evidence that the provision of textbooks accounts for a large part of the mean incremental TA learning effects, while the teacher training does not show significant effects. This evidence is supported by other RBF research; for example, Mbiti et al. (2019a) show that teacher incentives in Tanzania worked especially well in combination with sizable school grants that were spent in large part on textbooks. However, the grants did not have any effect on learning outcomes in the absence of teacher incentives in these schools. Complementarity between RBF instruments and school resources is also shown by Gilligan et al., 2018.

The third version of the program, with its explicit penalty for illiteracy in the bonus formula, combined with the evidence from the other group targeted programs, suggests that policymakers need to be quite explicit about distributional priorities in the design of an RBF instrument. In Ceará, the regressive effects largely disappeared only after illiteracy was introduced explicitly as a disincentive in the RBF instrument.

Anticipating distributional effects

The evaluation studies in this section all describe the trade-offs that decided the respective designs. In most cases the designers anticipated possible regressive distributional effects and tried to avoid them, for example by setting school-specific targets (Pernambuco); by including both gains and levels of learning (Ceará and Jakarta); or by organising competitions between "similar" schools (Chile and Jakarta). The fact that, overall, higher impacts in these programs are registered at the right-hand side of the distribution implies that, even if the distributional concerns were seen as a concern, they did not translate into a design that prioritized learning gains on the left-hand side.

The Ceará example shows that, with sufficient emphasis on basic skills acquisition in the bonus formula, RBF learning effects can be shifted to the left-hand side of the pre-treatment distribution. Possibly the design features intended to address distributional concerns in other programs were not sufficiently explicit; for example, a "fair school competition" does not instruct the teacher to pay extra attention to weaker students.

As illustrated in Table 5, five of the eight papers discussed in this section estimate the impact of a competition, a specific RBF design. Competitions are attractive to designers for a number of reasons,

[6] The paper provides evidence on mechanisms underlying the results, including a higher likelihood of having a formal selection process for school principals after the introduction of RBF; higher training participation by teachers; more teachers working full-time, checking homework and covering more than 80 percent of the curriculum; and fewer schools reporting a lack of textbooks.

including that "players" compete for a fixed bonus fund which makes the budget predictable, a feature that is especially useful for policymakers implementing large scale programs. A more subtle attraction is that "players" know much more about their production function than policy makers do. The competition will reveal (some of) this private information.

Not all the competitions described are the same, however. In our sample, we have seven competitions. Five of these are "rank-threshold competitions": these are the two Jakarta grants competitions, the Chilean SNED, the Pernambuco performance program and the teacher group competition in Western Kenya (i.e. all competitions except the programs in Lautharte et al.). A threshold competition implies that all participants are ranked on the (learning linked) performance indicator and only participants above a minimum threshold win a reward. In four of the threshold competition programs discussed here, the threshold is not absolute but relative: the highest X percent of participants is rewarded.[7]

Threshold designs in test-based accountability programs in education face the difficulty of choosing the proficiency threshold (Neal, 2011). Using data from the Chicago Public Schools, Neal and Schanzenbach (2012) show that a reform that introduced a school accountability system based on proficiency thresholds resulted in an uneven distribution of test score effects, with students at the bottom scoring the same or lower, while ".. students in the middle of the distribution score significantly higher than expected". These papers discuss the tension between setting the standards too low (and risks paying for skills everyone already has while not raising teacher effort) and setting standards too high. The existence of a threshold will make the incentive attractive (feasible) for only a part of the population of students, teachers and/or schools. As a result, effort and learning impact triggered by a threshold design will be concentrated in an area of the distribution not too far from the threshold, while schools or individuals either well below or well above the threshold will not be motivated to exert more effort. Neal and Schanzenbach (2012) show theoretically that " .. raising standards may actually increase the number of low achieving children who are "left behind" by increasing the number for whom the standard is out of reach. "

A practical argument is that the threshold sends a signal about educational standards. Education departments and policy makers can be expected to set proficiency thresholds at aspirational levels, in the upper half of the test score distribution. The programs studied here have thresholds that are roughly in line with this prediction. The Jakarta program rewards schools in the top 25 percent of the ranking; in the SNED, winning schools account for 25 percent of enrolment; in Kenya, teachers in the

[7] The Pernambuco design has a few special features. The state government allocated (at least) one month of teacher salary per year to the bonus pool so there is sufficient budget for all schools to be rewarded and so the percentage of winning schools is not limited: the minimum threshold for winning is set at 50 percent of a school specific target and if all schools meet that target, all can win. If schools do not reach 50 percent of their target, the budget is allocated to schools that do. This means that in advance the maximum absolute amount to be won by a teacher is uncertain. Moreover, once the minimum threshold is reached by a school the reward increases linearly in the percentage of the target reached).

highest scoring 48 percent (24) of 50 schools would be rewarded; and in Pernambuco, schools had to reach 50 percent of their individual target to be rewarded. It is important to note that, to create a more level playing field, in the Chile and Jakarta programs, the competition did not take place across the full population of schools but within school "leagues" based on district (and SES indicators in the SNED). In Pernambuco, targets are school specific and this made for a more level playing field.

Among programs with significant positive mean effects, we see inequality increasing interactions both in competition designs with thresholds and in those without (the first two versions of the Ceará competition). We also see that inequality decreased in the Jakarta primary schools program, a threshold competition. Nevertheless, the evaluations of the threshold competitions that measure increasing inequality refer to the relatively low probability of reaching the threshold for schools in the lower tail.

For example, in Jakarta, despite efforts to create a level playing field, junior secondary schools that performed well before the intervention had to make fewer improvements to win the performance grant. In Chile, a sizable 38 percent of schools that competed in six rounds were never awarded the SNED bonus. The authors show that schools with a low ex-ante probability of winning show much less learning response to the program. In Pernambuco, school targets were set such that schools with lower past performance were required to make bigger improvements. The regression discontinuity graphs show that on the left side of the bonus threshold, the distance from the threshold is generally negatively correlated with test scores. This confirms that lower past performance predicts a lower chance to win the bonus.

Two of the programs in this section were targeted at low performing schools. These are the Punjab program in Barrera-Osorio and Raju (2017), that targeted public primary schools with the lowest mean student exam scores in the province; and the Kenya program in Glewwe et al. (2010) that was implemented in a sample of low performing primary schools in Western Kenya, that were designated by the Ministry of Education as deserving assistance. The other programs in our sample do not target specific subsets of schools. One could argue that selecting these "deserving" school populations lowers the expected overall variation in performance, and, possibly, performance impact heterogeneity. However, the low mean program effects are also consistent with low effect heterogeneity.

In the Pakistan program, the design took into account the common concern in accountability programs that teachers may deselect weaker students from high-stakes exams. To address this, the exam participation rate is part of the bonus formula. However, adding more dimensions to the bonus formula added to "multi-tasking", another well-known design problem for incentive programs (e.g., Neal, 2011; Milgrom and Roberts, 1991). As the authors note, "Given costly effort, teachers may strategically direct their efforts at those incentivized margins where payoffs are more cost-effective". They conclude that, given the overall weakness of the targeted schools, schools may have decided to focus on sitting more students on the exam (a low-cost choice), at the expense of raising mean test scores.

3.2. Teachers

This review includes 12 studies that incentivized teachers for performance. Table 7 shows the performance metric that is used to reward teachers. Most studies reward teachers for the learning of students in their classroom, as measured by a test at the end of the school year. The way the score of the test translates into incentives differs across studies. In the threshold design, used in 1 study, the teacher gets paid if the student passes the test. In the level treatment, used in 2 studies, the teacher gets paid based on the grade of the student at the end of the school year. Four studies use the absolute change in test scores as the performance metric and 5 use the percentile rank of the student at the end of the year when compared to students with similar baseline ability (pay for percentile). Other performance metrics, not directly related to student grades, are teacher presence (2 studies) and performance rating of principals (2 studies), sometimes used in combination.

Table 7: Performance metric used in teacher incentive studies

Performance metric	Study	Size of reward*
Student learning: threshold	Mbiti et al. (2019a)	0.38
Student learning: continuous level	Loyalka et al. (2019) Mbiti et al. (2019b)	0.61 / 1.22 0.42
Student learning: absolute gains over year	Behrman et al. (2015) Filmer et al. (2020) Loyalka et al. (2019) Muralidharan (2012) and Muralidharan and Sundararaman (2011)	0.36 0.11 0.61 / 1.22 0.36
Student learning: pay for percentile	Andrabi and Brown (2020) Chang (2020) Gilligan et al. (2018) Loyalka et al. (2019) Mbiti et al. (2019b)	0.00 1.40 1.00 0.61 / 1.22 0.50
Teacher presence	Duflo et al. (2012)	-2.4
Rating of principle	Andrabi and Brown (2020) Gaduh et al. (2020)	0.00 -0.61
Absolute gains / rating of principal	Barrera-Osorio et al. (2021)	2.04
Rating of principal / teacher presence	Gaduh et al. (2020)	-0.41

Notes: * Mean rewards (penalties) as share of monthly teacher salary

Together, these studies test the impact of 23 different interventions. Most of these, 19, used financial incentives. Usually these consisted of bonuses on top of regular salaries for performance. Only three interventions (Duflo 2012 and Gaduh 2020 two times) used penalties. Two studies (Barrera-Osorio et al., 2021, Filmer et al., 2020) provided in-kind incentives, such as prizes, and another 2 studies used social incentives, such as social pressure or recognition (Gaduh et al. (2020), Barrera-Osorio et al. (2021).)

We focus on the impact that incentives have on learning as measured by a standardized test. If possible, we use a low stakes test (not used to calculate the incentive), but many studies do not report such a test. In those cases we report on the test that was used to calculate the incentive. Of the 38 estimates included, 20 showed significant positive mean effects, 17 reported a positive effect which was not significantly different from zero and 1 reported an insignificant negative effect. The mean effect size across all teacher studies is 0.11. std dev increase in learning. Not surprisingly, we find more significant heterogeneous treatment effects among studies which showed significant average impacts on learning. Table 8 provides a summary of the heterogenous treatment effects of teacher level incentive studies. The large majority of heterogenous treatment effects are not significantly different from zero. Looking at the point estimates, we observe that students with higher baseline scores benefit more often from the treatment. The same holds for boys. We also tabulated heterogenous treatment effects with respect to wealth, but only very few studies report these.

Table 8 Heterogenous treatment effects of teacher

Interaction	Test score	Girls	Wealth
NA	7	10	27
Significant negative	1	-	-
Negative	8	14	7
No effect	4	3	-
Positive	13	6	2
Significant positive	3	3	-

Perhaps the most important conclusion to note for Table 8 is that students who scored higher at baseline generally benefit more from teacher incentives. There are various potential explanations for this. One possibility is that teachers who teach higher-performing students respond more to the incentive treatment. Another possibility is that teachers direct their attention to particular students, in order to maximize their gains. A third possibility is that higher-performing students are equipped to take advantage of the additional effort of teachers, even though this attention is not particularly directed at them. We will review the studies selectively to investigate the validity of alternative hypotheses in the subsequent paragraphs.

The first hypothesis, whether teachers who teach higher-performing students are more responsive to incentive treatments, is easy to test. The empirical evidence reviewed does not support this hypothesis.

In Duflo et al. (2012), teachers were incentivized to be present at schools, and thus did not face any incentive to direct their attention to any particular type of student once present. The study finds that students who were literate at baseline, as indicated by the fact that they took the baseline test on paper, benefit more from the treatment (0.16 vs 0.25 std dev). On the other hand, the study also showed that teacher presence went up more in schools with lower scoring students. Teachers teaching in schools with test scores above the median were 0.15 percent point more likely to be present whereas those teaching classes with below median test scores were 0.24 percent more likely to be present (see table B discussion). Students with lower baseline scores thus on average received more teacher inputs, yet they benefited less from the program.

Gaduh (2020) also reports heterogenous treatment results by baseline ability level of students, and by whether the classroom was above or below the median. A similar pattern occurs. For the most successful intervention (Social accountability combined with absence monitoring with camera), the point estimates indicate that students with above median test scores benefited more from the intervention. Yet, they also find that children attending schools with above median test scores at baseline, had lower learning impacts in the second year.

The second hypothesis is that teachers direct their attention to students in order to maximize their gains. Testing this hypothesis is more complicated, as it is not a priori clear whether this would imply that teachers would direct their attention to students who score low or high at baseline. Various factors could be at play, and different studies shed light on different aspects.

Some children may benefit more from teacher attention because there are low hanging fruits, that is, with a little more attention these children can make large steps forward. Loyalka (2019) tests for this hypothesis by asking teachers directly which students they think would benefit most from an additional hour of private tutoring by the teacher. Students who the teacher classified as having "high benefit" benefited 0.21 to 0.33 std deviations more from an incentive where the teacher was paid for levels or gains student test scores (see table in summary). The evidence suggests that indeed teachers in these treatment groups ensured that the students they identified as needing additional attention were able to make those gains.

It is interesting to note that this pattern was not observed for teachers who were in the pay for percentile treatment in Loyalka (2019). For this treatment, the comparable point estimate was much lower (0.05 std dev) and insignificant. Possibly, this treatment reduces the incentive for teachers to focus on these students in need. One possible hypothesis is that the type of skills these students lack are similar across classrooms, for example, they lack some foundational skills, which can be easily taught. Because the pay for percentile treatment is organized as a rank order competition among students with similar starting levels, it would not be advantageous to focus more on these students, as the competition has the same advantage. In the other schemes, where the payment is based on test scores directly, gains in skills for these students will translate in high increases in test scores, and thus high payments for teachers.

Threshold designs, where the teacher incentive depends on whether the student succeeds or not, are criticized because they would provide an incentive to focus on those students for whom succeeding depends on whether they receive additional assistance or not. Success could be passing a test or making it to the next school level. In the case of passing a test, the hypothesis is that teachers focus their attention on those students who would be close to the passing threshold. Only one study (Mbiti 2019a) in our review uses a threshold design where teachers get paid depending on the number of students that pass a test. The authors test whether there is any heterogeneity with respect to the distance from the threshold. They find mainly negative interaction effects (larger distances from the threshold associated with higher impact), but these are not significant.

Test design could also cause teachers to devote particular attention to a particular group of students. If a test rewards a lot of points for particular skills, it is relatively easy to increase test scores by focusing on students who still lack these skills and can learn them easily. Note that this is not necessarily a bad thing, the fact that these skills received a high weight in the test is likely also related to educational objectives. The pay for percentile treatment removes this potential concern, as the payment is based on percentile rankings and not test scores directly.

One would thus expect less heterogenous treatment effects in pay for percentile studies as compared to other pay for performance studies. Loyalka (2019) and Mbiti (2019b) allow for an immediate comparison of incentives that pay on the basis of test score versus one that pays for percentile. The evidence does not support the hypothesis that the pay for percentile treatment yields less heterogenous treatment effects with respect to baseline ability. The point estimates in Loyalka (2019) suggest that lower scoring students benefited more from the interventions across all treatment arms. The absolute values of the heterogeneity point estimates are even slightly higher for the pay for percentile treatment, which is opposite of what one would expect. Also in Mbiti (2019b), the point estimates suggest more heterogenous treatment effects with pay for percentile, but the difference is not significant.

An across study comparison of pay for percentile versus other treatments which pay teachers based on test scores of students also does not yield a clear pattern on which type of incentive has more heterogeneous treatment effects. We observe for the pay for percentile interventions one study which shows significantly positive interaction effects with baseline ability (Mbiti 2019b, math), two that show a positive insignificant effect (Andrabi and Brown 2020, Mbiti 2019b, language) and 3 that show negative insignificant effects (Loyalka 2019, Chang 2020, and Gilligan 2018).

Looking at the other (non pay for percentile) interventions which provide rewards for student test scores, we find 3 that show a positive insignificant interaction effect (Filmer 2020, Mbiti 2019b, Muralidharan (2012), and three that show a negative insignificant effect (Loyalka 2018, levels and gains, Mbiti 2019a, incentives arm). For the combination treatment of threshold incentives and school grants, Mbiti et al. (2019a) estimate a significantly *negative* interaction effect with baseline test score.

The third hypothesis is that higher-performing students are better equipped to take advantage of the additional effort of teachers, and this explains the positive interactions we see between teacher incentives and baseline ability of students. Students with higher baseline test scores have been able to learn more in the past. This could be because they have mastered pre-requisite foundational skills, or because they have (had) more resources, such as books and supportive parents, which helped them to learn. Having these resources also makes them able to take advantage of additional teacher effort. The empirical evidence reviewed is supportive towards this hypothesis.

Mbiti (2019a) is the only study that experimentally tests the complementarity between resources and teacher incentives. They find clear evidence that teacher incentives have more impact if they are combined with a school grant. Whereas the teacher incentive yields an insignificant positive impact on test scores of 0.06 std dev in the first year and 0.03 std dev in the second year, the combined program yields significant effects of 0.12 std dev and 0.23 std dev respectively. That this is not a direct effect of the additional resources the grant brought becomes clear when we look at the effect of the grant alone, which yielded impacts of -0.03 std dev and 0.01 std dev respectively, both insignificant (Mbitit 2019a, original paper). This indicates that at the school level, there is a complementarity between resources and teacher incentives. As mentioned, for this treatment arm the study finds a significantly negative interaction term with baseline test scores. This finding contrasts with the standard hypothesis that threshold designs favor higher performing students that are close to the passing threshold before the treatment. The authors do not find the negative interaction for the incentives-only arm. In addition, Mbiti et al. (2019b) find non-significant positive interactions with baseline test score for a modified threshold design that makes it easier for low-performing students to pass tests that will pay a bonus to their teacher. Overall, the evidence does not suggest that threshold designs lead to inequality effects.

Non-experimental evidence pointing in the same direction comes from Gilligan (2018). Their study showed that a pay for percentile experiment in Uganda had higher impacts in schools with books. The impact was -0.03 std dev in schools without books and 0.072 std dev in schools with books, and was significantly different at the 10 percent level.

Size of reward

Another dimension on which teacher incentives differ is the size of the reward that is given to teachers. Table 7 presents the average size of the reward that was paid to teachers, on a yearly basis, expressed in monthly salary. There is a wide variation across studies ranging from penalties which could run up to 2.4 months' salary to bonuses equivalent to 2 months' salary.

One would expect a higher reward to induce greater additional effort by teachers and thus higher impacts. And those studies which have higher impacts also more often have impacts that vary by subgroups. But a priori it is not clear what type of students would benefit from higher rewards. One possibility is that the teacher, when exerting low effort, direct their attention to the low hanging fruits, that is, those students that can make a lot of progress with just a little bit of effort from the teacher, and thus generate rewards for teachers. When the reward increases, and teachers exert higher effort,

they may direct their attention to other students as well. This reasoning suggests that heterogenous treatment effects would reduce as the reward increases.

The only study that tests the impact of the size of the reward experimentally is Loyalka (2019). They find that doubling the reward does not increase the average treatment effect. Averaging across three different incentive payment mechanisms, the low reward yielded an impact of 0.081 std dev, and a high reward 0.067 std dev (difference is not significantly different from zero), running against the hypothesis that teachers would increase effort if the award amount increases.

Chang (2020) provides an interesting case study which indicates that it is possible to incentivize teachers to focus on low performing students by adding additional rewards for this group. In this case, the Loyalka study was extended with an adjusted pay for percentile treatment, where higher rewards were given for students with lower baseline ability. Limiting ourselves to schools with the pay for percentile treatment, we find that in Loyalka (2019), it was the middle tercile that benefited most. For the Chang (2020) study, on the other hand, it is the bottom quintile which benefits most. This is an important finding for our review, as it suggests that distributional concerns can be addressed using the flexibility of RBF instruments. The Chang (2020) study is the only teacher level incentive experiment that studies this possibility, and, given global concerns around persistent illiteracy in primary schools, the finding suggests an important area for future research.

Gender effects

Across studies, differential treatment effects by gender are generally very small. Perhaps the most interesting gender effects are found in Barrera (2021) who compare an in-kind award scheme with a scheme that focuses on giving public recognition for achievements. The in-kind reward scheme showed no differential effects by gender. For the recognition award scheme, girls benefited substantially more (interaction term -0.12, significant). The paper also showed that female teachers were motivated more by recognition awards.

3.3. Students

The six papers in this subsection study programs in which incentives are tied to performance of students, either as individuals or as small teams. The studies are listed below and summaries are provided in the appendix C. These studies provide impact estimates for eleven different student level incentive programs, where different experimental arms in one study count as different programs. We provide an overview in

Table 9.

These studies all have an experimental design and their relative size is limited, especially compared to some studies in the section on group targeted incentives. The studies span a range in study sample

sizes: 8 schools (Berry), 18 schools (Hirshleifer), 88 schools (Behrman), 100 schools (Blimpo), 127 schools (Kremer), 400 schools (Filmer).

Three of the papers study experiments that provide (or include) cash rewards; two provide in-kind rewards; and in one the reward is a scholarship.

Table 9: Student level: included studies

Study	Country	Incentive type	School type	Estimates
Behrman et al. (2015)	Mexico	Cash, for improvements	Secondary	3
Berry (2015)	India	Cash/in-kind, individual targets	Primary	1
Blimpo (2014)	Benin	Cash, thresholds and tournament	Secondary	3
Filmer et al. (2020)	Tanzania	In-kind, competition	Secondary	2
Hirshleifer (2016)	India	In-kind, value linear in correct answers	Primary	2
Kremer et al. (2009)	Kenya	Scholarship for girls, competition	Primary	2

The papers provide 13 sets of impact estimates, summarized in Table 10 below. Each year for which a study reported a subgroup estimate for a student incentive treatment is included separately. The coding of the studies follows the standard approach in this review.

We obtained 13 mean impact estimates; 13 estimates of the interaction of treatment with baseline test score; and eight and two estimates of the interactions with, respectively, gender (girls) and household wealth.

Table 10: Student level estimates

Effect category	Mean effect on learning	Interactions		
		Test score	Girls	Wealth
NA			5	11
Significant negative		1		
Negative		2	1	
No effect		2	3	2

Positive	4	7	4	
Significant positive	9	1		
Total	13	13	13	13

Of the 13 result sets, nine have mean learning impact estimates that are positive and significant (at least at the 10 percent level); and four have a positive but not significant effect. Of the six student incentive programs that were effective in raising student learning on average, three are from Blimpo (2014); one is from Behrman et al. (2015); one is from Kremer et al. (2009); and one from Hirshleifer (2016). The program offering incentives to both students and teachers (Filmer et al., 2020) also had a positive significant treatment effect. The highest mean impacts are found in Hirshleifer, Behrman and Blimpo.

Of the baseline score interaction estimates, two are significant. One is a positive interaction effect in the team tournament treatment in Blimpo (2014). Berry (2015) has a significant negative interaction effect with the baseline test score in an analysis that compares the incentive effects of toys versus cash rewards (so not with a pure control group). Of the eight positive baseline score interaction estimates, seven are from programs with a significant positive mean treatment effect.

For wealth we have no treatment interaction coefficient estimates. Kremer et al. (2009) note that they tested and did not find a treatment wealth interaction in the two regressions we include, but do not table the interaction coefficients. For gender we have a few interaction estimates. Blimpo (2014) notes that there are no impact differences between boys and girls in any of the incentive treatments. Hirshleifer (2016) provides (insignificant) interaction estimates for the two arms in the experiment, one negative and one positive. Behrman et al. (2015) find insignificant positive effects for girls in all three years of the student incentive experiment.

Incentive designs

In our discussion, we focus on the baseline test score interaction estimates. Overall, the student level RBF studies provide some evidence that RBF interventions increase test score inequality, but the interaction effects are generally not significant. This is not an obvious outcome since most of the interventions in this section do not provide special mechanisms to level the reward probability across the distribution of baseline proficiency.

There is one program that provides uniform in-kind rewards for individualized student targets, based on a pre-treatment test (Berry, 2015), and thus explicitly accounts for baseline proficiency. This study reports a significant negative interaction effect. Note, however, that the estimate is relative to students in a cash reward treatment arm (not relative to pure control).

Most of the other programs in these papers have designs that, in theory, could provide stronger incentives among high performing students. For example, the two target experiments (team and

individual) in Blimpo (2014) provide rewards for reaching absolute exam standards, with pay-offs that are sharply increasing in exam performance. For these treatments, mean impact estimates are robust. The interaction with baseline score is small and negative for the individual target treatment, and it is positive and sizable for the team target - but not statistically significant.

The team tournament design in this paper only has rewards for three out of 84 teams, meaning teams need to perform above the 96th percentile to win. This is the program with the only positive significant baseline interaction effect, suggesting that indeed incentives were concentrated among students who were high performers at baseline. The small reward probability may have discouraged lower performing student teams.

Kremer et al. (2009) is a competition for a scholarship for girls that pays for fees and inputs in the final two years of primary school in Western Kenya. The scholarship is assigned to the highest scoring 15 percent, based on a government exam, across all grade six female students in treatment schools. Students are thus competing across the full distribution of student proficiency, without consideration for initial test scores. Nevertheless, the authors test for effect heterogeneity and do not find a consistent pattern of interactions with baseline test scores. The authors do find spill-over effects on the performance of boys, who were not eligible for the scholarship.

Filmer et al. (2020) study a different student competition design as one arm of an incentive experiment in secondary schools in Tanzania. Here, students compete with other students in the same class, which takes away competition across schools. The top three students in each class received an in-kind reward. The student incentives do seem to motivate the students with higher initial test scores (as shown in the quantile estimates).

The last three studies discussed (Blimpo team tournament, Kremer et al., Filmer et al.) all estimate tournament treatment effects, with small reward probabilities even if prizes would be assigned randomly. These designs provide quite steep reward thresholds that do not provide much incentive in the left-hand tail of the test score distribution. Indeed, in the Benin and Tanzania studies, this incentive design feature may have discouraged students who consider themselves to be too far from the reward threshold when the bonus rules are explained. However, we note that the team target treatment in Benin resulted in estimates that are quite similar to those in the tournament treatment. The common characteristic is that both these treatments contain a social element, although the emphasis in these two experimental arms is very different (collaboration versus competition). Furthermore, it is interesting to see that discouragement did not happen in Kenya, where even formally non-eligible students - boys - were motivated and improved their test scores relative to the control group.

The Behrman et al. (2015) study finds two consistent results for the student incentive arm across three years. First, mean treatment effects are large and highly significant. Second, the pattern of quantile interactions suggest that better off students benefited more. Often their point estimates were double that of the lowest ability level. The facts that higher payments were provided for higher proficiency

levels may have motivated students who were close to that level more. Nevertheless, the difference between the highest and lowest quantile estimates is not significant (large standard errors).

Finally, the design in Hirshleifer (2016) provides students with credit that can be exchanged for rewards. There are no clear thresholds in this case, as the rewards are increasing in the number of correct answers provided on exams. This design is quite different from the "few winners take all" competitions described earlier, as rewards are attainable across the distribution. Nevertheless, the design could potentially focus rewards on students performing highly before the treatment, e.g. if more valuable items in the store are more salient and only attainable by these students. The interaction coefficient is positive in one arm that measures and incentivizes continuously ("input design"); and negative in the "output design", where students are only assessed and rewarded at the end of the unit. None of the interactions is statistically significant.

4. Conclusion

We offer some concluding remarks, under the three research questions of the study.

1. What are the estimated impacts of RBF mechanisms (in education) on the outcomes for different groups?

Overall, the RBF programs in the studies we reviewed resulted in few statistically significant subgroup effects. In other words, most of these programs do not increase or decrease inequality of pre-existing student learning outcomes. We find this absence of significant effect heterogeneity across the three incentive levels in our study: groups, teachers and students.

We summarise the literature findings using estimate counts; most of the heterogeneity estimates relate to (pre-treatment) test scores, a few to gender and wealth. For test scores, we find zero significant positive interaction effects for group level incentives (out of 28 mean effect estimates), and zero significant negative interactions; three significant positive interaction effects for teacher level incentives (out of 36), and one significant negative interaction; one significant positive interaction effect for student level incentives (out of 13), and one significant negative interaction. The individual teacher level incentive studies provide most significant positive test score subgroup effects (three), and most significant positive subgroup effects for girls (three).

We do find several *non-significant* subgroup impact estimates and there are more positive than negative *test score* subgroup estimates, implying that students performing higher before the treatment benefitted more from the RBF program. Across all interventions, we found 37 positive and 18 negative test score subgroup point estimates – but the impact difference with lower performing students is typically not significant.

We find non-significant subgroup estimates in both small (experimental) studies and in evaluations of larger scale programs. In particular, in a subset of four large scale group incentive programs that introduce performance incentives based on administrative data, implemented by (sub-national) governments, researchers find, in most versions studied, a significant positive mean impact with a mostly positive but non-significant interaction with baseline test scores.

A possible explanation for the overall absence of significant heterogeneity effects is that the incentive design anticipated such effects and took steps to avoid or reduce them. As discussed in more detail under the third research question, most studies in this review acknowledge the risk of increasing outcome inequality and chose design features that potentially mitigate these effects. However, even in studies without explicit mitigating factors (e.g. most of the student level incentive designs) we do not find strong evidence of effect heterogeneity.

2. What design and implementation characteristics explain differences in the impact between subgroups?

As noted, there are few significant subgroup effects of RBF education programs. The literature, however, shows clearly that implementors have been aware of the possibility that the design of the incentive would cause certain subgroups to benefit more from the interventions than other. For the simplest design, which is based on the fraction of students which pass a threshold level, there is a risk that teachers will focus their efforts on students close to the threshold, to maximize their pay-out with minimum effort. The most complicated design, pay for percentile, overcomes this problem, and addresses perverse incentives which could result from test design. In this design, students of similar ability compete against one another, and pay-outs are based on the rank order in de competition.

A number of incentive studies provide evidence on the importance of threshold incentive designs for distributional effects. We do not find strong evidence that threshold designs result in more learning inequality. Among group level programs, we see (non-significant) positive test score interactions both in competition designs with thresholds and in those without. Similarly, and in contrast with the standard hypothesis that threshold designs favor students that are close to the passing threshold, we do not find such interactions in teacher incentive studies with a threshold design. Loyalka et al (2019) and Mbiti et al. (2019a) do not find impact heterogeneity (with respect to the baseline test distribution or distance from the threshold, respectively). Overall, the evidence does not suggest that threshold designs lead to inequality effects.

The review provides some evidence on the distributional dimensions of the pay for percentile design (in individual teacher level programs). The Loyalka et al. (2019) study shows that teachers focus less on students who would most benefit from their attention (according to the teacher) in the pay for percentile arm, compared to treatments that pay for levels or gains. This is consistent with the rank order competition design, but not necessarily desirable. Mbiti et al. (2019b) study the heterogeneity of learning effects in pay for percentile and find that it does not result in more equal outcomes than a threshold design. Taken together, the evidence does not support the hypothesis that the pay for percentile treatment yields less heterogenous treatment effects with respect to baseline ability.

3. How can RBF be used to support educational attainment and outcomes of specific (marginalized) target groups (such as girls or poor households)?

The studies do show, however, that students with high test scores at baseline often benefit more than those with low test scores at baseline, although, again, these differences are generally not significant. This could be because they have more resources (as in Gilligan et al., 2018). A few studies have shown that RBF design can be tailored to counter this. Programs could add resources for learning, which could help to overcome the resource advantage that better off students have at baseline. Another option is to add extra incentives for learning gains of students who are at the lower end of the distribution.

Adding resources to an incentive program could reduce this advantage. Incentive programs tend to more effective if combined with a program that provides resources, such as school grants. The findings

indicate that adding school resources (finance, books, teachers) to incentive programs makes them more effective (Lautharte et al., 2021), and it may reduce inequity (Mbiti et. al., 2019a).

RBF programs could also be designed with the explicit purpose of achieving more *progressive* distributional effects. The empirical literature on how to achieve this is limited. All studies in our review are designed to precisely measure average treatment effects and they are often designed to compare the mean effects of different experimental arms. The effects on subgroups (inequality effects) are in many cases reported but are typically not the primary concern of these studies. The programmatic differences between intervention arms are usually driven by expectations concerning the mean effects (or cost-effectiveness based on mean treatment effects), not by questions of distribution.

Nevertheless, our review contains a few examples of programs with specific incentive features designed to incentivize teachers to devote particular attention to less advantaged groups, with encouraging results. At group level, Lautharte and co-authors (2021) describe a revision of the reward formula in the Brazil (Ceará) tax allocation incentive program. The formula was revised to emphasise basic literacy outcomes, with results suggesting a more equitable distribution of learning impacts across grades and subjects. A second example, at teacher level, is the study for China by Chang et al. (2020), that increases the rewards in a pay for percentile treatment for the lowest scoring students in a sample where previously a pay for percentile treatment was implemented without any differentiation across students. The Chang study yielded more pro-poor learning results.

These two examples suggest that RBF incentive design can be structured to support educational outcomes for low performing and disadvantaged students. We consider this to be a promising area for future research.

Appendix A: Studies included in the review with link to summary

Appendix Table 1: Studies reviewed

Full citation for each study	Page
1. Al-Samarrai, S., Shrestha, U., Hasan, A., Nakajima, N., Santoso, S., & Wijoyo, W. H. A. (2018). Introducing a performance-based component into Jakarta's school grants: What do we know about its impact after three years?. *Economics of Education Review, 67,* 110-136.	41
2. Andrabi, T., & Brown, C. (2021). *Subjective Versus Objective Incentives and Teacher Productivity.* (Working Paper as of March 11 2021).	58
3. Barrera-Osorio, F., & Raju, D., (2017). Teacher performance pay: Experimental evidence from Pakistan. *Journal of Public Economics, 148,* 75–91. https://doi.org/10.1016/j.jpubeco.2017.02.001	49
4. Barrera-Osorio, F., Cilliers, J., Cloutier, M. H., & Filmer, D. (2021). *Heterogenous Teacher Effects of Two Incentive Schemes: Evidence from a Low-Income Country.* (Policy Research Working Paper 9652). https://openknowledge.worldbank.org/handle/10986/35565	68
5. Behrman, J.R., Parker, S.W., Todd, P.E., & Wolpin, K.I. (2015). Aligning learning incentives of students and teachers: results from a social experiment in Mexican high schools. *Journal of Political Economy, 123* (2), 325-364. https://doi.org/10.1086/675910	69
6. Berry, J. (2015). Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India. *Journal of Human Resources, 50* (4), 1051-1080. http://jhr.uwpress.org/content/50/4/1051.abstract	75
7. Blimpo, M. P. (2014). Team incentives for education in developing countries: A randomized field experiment in Benin. *American Economic Journal: Applied Economics*, *6* (4), 90-109. DOI: 10.1257/app.6.4.90	72
8. Chang, F., Wang, H., Qu, Y., Zheng, Q., Loyalka, P., Sylvia, S., Shi, Y., Dill, S. E., & Rozelle, S. (2020). The Impact of Pay-for-Percentile Incentive on Low-Achieving Students in Rural China. *Economics of Education Review. 75* : 101954. https://doi.org/10.1016/j.econedurev.2020.101954	52
9. Contreras, D., & Rau, T. (2012) Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile. *Economic Development and Cultural Change*, *61* (1), 219-246. https://doi.org/10.1086/666955	45

10. Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, *102* (4), 1241–1278. https://doi.org/10.1257/aer.102.4.1241	55
11. Ferraz, C., & Pereira, V. (2016) Can Students Benefit it if Teachers Lose their Bonus? Behavioral Biases Inside the Classroom. (Working paper as of August 5, 2016).	51
12. Filmer, D. P., Habyarimana, J. P., & Sabarwal, S. (2020). *Teacher Performance-Based Incentives and Learning Inequality.* (Policy Research Working Paper No. 9382). World Bank. https://openknowledge.worldbank.org/handle/10986/34468	65
13. Gaduh, A., Pradhan, M., Priebe, J., & Susanti, D. (2021). *Scores, Camera, Action: Social Accountability and Teacher Incentives in Remote Areas.* (Policy Research Working Paper No. 9748). World Bank. https://openknowledge.worldbank.org/handle/10986/36112	57
14. Gilligan, D.O., Karachiwalla, N., Kasirye, I., Lucas, A., & Neal, D.A. (2018). Educator Incentives and Educational Triage in Rural Primary Schools. *Journal of Human Resources*. 10.3368/jhr.57.1.1118-9871R2	59
15. Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2* (3), 205-27. https://economics.ucr.edu/pacdev/pacdev-papers/incentives_for_effort.pdf	46
16. Hirshleifer, S.R. (2016). *Incentives for effort of outputs? A field experiment to improve student performance.* (Working Paper No. 201701), University of California at Riverside, Department of Economics.	74
17. Kremer, M., Miguel, E., Thornton, R. (2009). Incentives to Learn. *The Review of Economics and Statistics*. *91* (3), 437-456. https://doi.org/10.1162/rest.91.3.437	73
18. Lautharte, I., de Oliveira, V. H., Loureiro, A. (2021). Incentives for mayors to improve learning: evidence from state reforms in Ceará, Brazil. (Policy Research Working Paper No. 6694). World Bank. https://openknowledge.worldbank.org/handle/10986/35024	42
19. Loyalka, P., Sylvia, S., Liu, C., Chu, J., Shi, Y., (2019). Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement. *Journal of Labor Economics* 37, 621–662. https://doi.org/10.1086/702625	52
20. Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., Rajani, R. (2019a). Inputs, Incentives, and Complementarities in Primary Education:	64

Experimental Evidence from Tanzania. *The Quarterly Journal of Economics, 134* (3), 1627-1673. https://doi.org/10.1093/qje/qjz010	
21. Mbiti, I., Romero, M., & Schipper, Y. (2019b). Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania. (NBER Working Paper No. 25903). National Bureau of Economic Research. https://www.nber.org/papers/w25903	63
22. Muralidharan, K. & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119* (1), 39-77. https://doi.org/10.1086/659655	61
23. Muralidharan, K. (2012). Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India. (Working paper for Society for Research on Educational Effectiveness). https://eric.ed.gov/?id=ED530172	61

Studies excluded from the shortlist

1. Barrera-Osorio, F., & Filmer, D. (2016). Incentivizing Schooling for Learning: Evidence on the Impact of Alternative Targeting Approaches. *Journal Human Resources, 51*, 461-499
2. Castro, J., & Esposito, B. (2018). *The effect of bonuses on teacher behavior: a story with spillovers.* (Peruvian Economic Association Working Paper No. 104). https://ideas.repec.org/p/apc/wpaper/2017-104.html
3. Cilliers, J., Kasirye, I., Leaver, C., Serneels, P., & Zeitlin, A., (2018). Pay for locally monitored performance? A welfare analysis for teacher attendance in Ugandan primary schools. *Journal of Public Economics 167*, 69–90. hhttps://doi.org/10.1016/j.jpubeco.2018.04.010
4. Hong, S. Y., Cao, X., & Mupuwaliywa, M. (2020). Impact of financial incentives and the role of information and communication in last-mile delivery of textbooks in Zambia. (Policy Research Working Paper No. 9305). World Bank. https://openknowledge.worldbank.org/handle/10986/34020
5. Leaver, C., Ozier, O., Serneels, P., & Zeitlin, A. (2021). Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools. *The American Economic Review*, *111* (7), 2213-2246. DOI: 10.1257/aer.20191972
6. McEwan, P. J., Santibanez, L. (2005). Teacher and Principal Incentives in Mexico., In E. Vegas, (Ed.). *Incentives to Improve Teaching*. The World Bank. Washington DC. http://hdl.handle.net/10986/7265
7. Pugatch, T., Schroeder, E. (2018) Teacher pay and student performance: evidence from the Gambian hardship allowance, *Journal of Development Effectiveness*, *10* (2), 249-276, DOI: 10.1080/19439342.2018.1452778

Studies not included

In this sub-section we describe the motivation for removing these 7 papers from our shortlist. Three studies were excluded because the interventions studied operate on the extensive margin only, typically through a one-off bonus for accepting a teaching position. These interventions do not offer incentives for increased effort that depend on outcomes achieved on the job and so *intensive margin* effects (effort, learning outcomes) are less likely in these studies.

Pugatch and Schroeder (2018) were both removed because they evaluate the impact of a remote area allowance, aimed at attracting teachers to remote areas. Once the teacher accepted the position, this bonus was provided without conditions. We do not think this a performance incentive to improve education, as it works on the extensive margin only, and is similar to an unconditional salary increase once on the job.

For similar reasons we excluded Barrera-Osorio and Filmer (2016), which is an experiment that studied the participation and learning effects of a merit scholarship program in Cambodia. The key question was whether targeting the scholarship to well performing students, and framing it as a merit-based scholarship, would affect both learning and participation. In a second experimental arm a poverty targeted scholarship was offered. We decided not to include the study because the program does not offer incentives on the intensive margin, based on prospective learning results during the program, but rather a one-off incentivize on the extensive margin based on a pre-test.

The remaining four papers do operate on the intensive margin, and thus met the criteria for entering the shortlist, but did not yield any insights on heterogenous treatment effects of performance incentives for various reasons.

Cilliers et al. (2018) is a study where teacher presence was incentivized in Uganda. Presence is recorded by the headmaster. In one treatment arm, the reported presence was relayed and made public in the community. In the second, a salary incentive was added. The study shows no impact of the information treatment, but strong impacts on teacher and student attendance if a salary incentive is added. The authors use Lee bounds to correct for differential attrition in the control and treatment groups. Unfortunately, the Lee bounds are very wide and become uninformative (include zero), so no statement can be made with regards to the impact on learning. While this is an interesting study, the study contributes little to our understanding of heterogenous treatment effects. The learning data are not useful because of the attrition problem. We also cannot calculate heterogenous treatment effects for the attendance results, as none of the covariates we are interested in (gender, baseline scores, wealth) is included in the replication data.

Hong et al. (2020) studies an incentive for book collection/delivery to remote schools in Zambia. The counterfactual is non incentivised delivery to schools in control districts and the paper uses a diff-in-diff regression framework. Data for the evaluation are at school level, but do not allow subgroup analysis over the margins of interest for this review, that is, for test scores, gender, or wealth.

Leaver et al (2021) is a high-quality paper, published in the American Economic Review, that reports on a study that randomized offered contracts across 16 labor markets in Rwanda and then rerandomized some teachers into a pay for performance (P4P) contract, and some into a fixed wage contract. The P4P contract included a pay for percentile element, but also elements related to pedagogy and presence. The P4P contract paid 15 percent extra salary to the 20th percentile top scoring teachers in a labor market.

We focus on the impact of pay for performance on the job. Of the 164 schools, 85 were assigned to P4P and 79 were to the fixed wage contract. Teachers were given a retention bonus if they participated in the randomization, which was always higher than the maximum bonus they could receive under the program they expected to participate in. This was sufficient to have no one object to being rerandomized, the study reports. The study has a baseline, and collected follow up data in end of 2016 and end of 2017. The impact on learning was 0.06 std deviations in year 1 (confidence band −0.03, 0.15], and 0.16 std dev in year 2 [confidence band 0.04, 0.28]. The paper does not report heterogenous treatment effects and we have not been able to obtain these effects from the authors or calculate them ourselves[8,9].

McEwan and Santibanez (2005) is a paper on the teacher promotion scheme in Mexico. Teachers are promoted based on a composite score, which is mostly not related to performance in the classroom. However, student test scores are 20 percent of the total score on which the promotion is based. The authors compare learning in schools with teachers who are close to the threshold, and for whom the students' test scores could make the difference between getting promoted or not, with those who are far away, for whom promotion (or not) was a sure thing. They find no impacts.

This paper is interesting because it reports on a national program that conditions teacher promotion on student test scores, and for this reason it was included in the shortlist. But the incentive is quite weak, 80 percent of the promotion is based on other criteria than learning, and the identification assumes that teachers will respond differently depending on whether they are close to the threshold or not (rather than test it). The paper does not report heterogenous treatment effects and replication data are not available. For this reason, it was excluded from the analysis.

Other literature

1. Adelman, M., Blimpo, M. P., Evans, D., Simbou, A., & Yarrow, N. (2015). Can Information Technology Improve School Effectiveness in Haiti? Evidence from a Field Experiment. Working Paper, World Bank, Washington, DC.

[8] Attempts to reach the authors for guidance on reproducing the results were unsuccessful.
[9] We have not been able conduct additional analysis using the replication files made available on the AER website as the program is written in a non-standard language and requires extreme computing power. We have tried to reach out to the authors for guidance, but have received no response so far.

2. Baird, S., McKenzie, D. & Özler, B. (2018). The effects of cash transfers on adult labor market outcomes. *IZA Journal of Development and Migration 8*, (22) https://doi.org/10.1186/s40176-018-0131-9
3. Baird, S., Ferreira, F. G. H., Özler, B., Woolcock, M. (2013). Relative Effectiveness of Conditional and Unconditional Cash Transfers for Schooling Outcomes in Developing Countries: A Systematic Review. *Campbell Systematic Reviews, (9),* 1, 1-124. https://doi.org/10.4073/csr.2013.8
4. Bastagli, F., Hagen-Zanker, J., Harman, L., Barca, V., Sturge, G., & Schmidt, T., & Pellerano, L., (2016). Cash transfers: what does the evidence say? A rigorous review of programme impact and of the role of design and implementation features. Overseas Development Institute, London, UK. https://odi.org/en/publications/cash-transfers-what-does-the-evidence-say-a-rigorous-review-of-impacts-and-the-role-of-design-and-implementation-features/
5. Brown, C. Andrabi, T. (2021). Inducing positive sorting through performance pay: experimental evidence from Pakistani schools. Working paper. https://christinalbrown.github.io/Christina_Brown_JMP_Teacher_Sorting.pdf
6. Bruns, B., Filmer, D., Patrinos, H.A., (2011). *Making schools work: New evidence on accountability reforms*. World Bank. https://openknowledge.worldbank.org/handle/10986/2270
7. Bruns, B., Luque, J., (2015). *Great teachers: How to raise student learning in Latin America and the Caribbean*. World Bank. https://openknowledge.worldbank.org/handle/10986/20488
8. Das, J., Do, Q. T., & Özler, B. (2005). Reassessing Conditional Cash Transfer Programs. *The World Bank Research Observer*. *20 (1),* 57–80. http://www.jstor.org/stable/41261409.
9. Ferraz, C., Bruns, B. (2012). Paying teachers to perform: the impact of bonus pay in Pernambuco, Brazil. Unpublished manuscript.
10. García, S., Saavedra, J.E. (2017). Educational Impacts and Cost-Effectiveness of Conditional Cash Transfer Programs in Developing Countries: A Meta-Analysis. *Review of Educational Research. 87* (5), 921-965. doi:10.3102/0034654317723008
11. Lee, L., Diana, J., Medina Pedreira, O., (2019). *Results-Based Financing in Education: Learning from What Works*. World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/915061548222619389/Results-Based-Financing-in-Education-Learning-from-What-Works.
12. World Bank. (2015). Snapshot: The Rise of Results-Based Financing in Education. Education Global Practice.

Appendix B: Summaries

Group Incentive summaries

Study 1: Al-Samarrai, S., Shrestha, U., Hasan, A., Nakajima, N., Santoso, S., & Wijoyo, W. H. A. (2018). Introducing a performance-based component into Jakarta's school grants: What do we know about its impact after three years?. Economics of Education Review, 67, 110-136.

Program and evaluation

Al-Samarrai and co-authors (2018) study a performance-based grant component that was introduced into an existing grants scheme for primary and junior secondary public schools in Jakarta, Indonesia, in 2014. The performance-based component awarded winning schools with an additional per student allocation equivalent to 20% of the existing basic grant allocation: 12 USD for primary schools (basic grant 60 USD) and 22 USD for junior secondary schools (basic grant 110 USD, Table 1). The RBF rewards in this case are school grants, the use of which is decided by school principals. Examples of expenditure mentioned in the paper are infrastructure upgrading and hiring contract teachers. In other words, winning the grant does not provide individual teachers with money for private use.

School performance was measured using two indicators, each with equal weight: first, students' average examination performance over the previous two years (2013 and 2014); and second, the percentage point improvement in students' average examination performance over the same two years. The scheme awarded the grant using a competition. The performance metrics were used for ranking and schools in the top 25% were awarded the grant in the following year.

The study presents difference-in-differences estimates of the program impact, using non-government schools (non-eligible) as controls and show evidence of parallel pre-treatment trends.

Results

The central findings of the study are that the program had on average a small negative effect for primary schools, which is significant in one year; and the program had a positive mean effect in junior secondary schools in all years studied. In primary schools, the program did not work as expected and resulted in negative or not significant results on average, driven largely by negative effects for schools in the third and fourth quartile. The authors hypothesize that this was the result of unproductive investments by principals in these schools, for example sending teachers away for training and so reducing contact hours.

The paper provides subgroup effects by estimating the exam score regression separately by quartile of the pre-treatment exam scores (Table 7 of the paper, results copied in Appendix Table 2 below). For the quartile estimates, if there is consistent pattern in the point estimates, we coded it either positive or negative. If there is an inconsistent pattern, for example basic has higher impacts compared to both pre-basic and higher level, we coded it as zero. We calculated the significance levels by conducting a

two-sample t-test based on the group estimates and reported standard errors. For the exam score quartiles, we based it on a test of quartile 4 versus quartile 1.

The grants competition increased inequality among junior secondary schools as high performers were much more likely to receive the grant than low-performing schools. Particularly in the second a third year of the program, the pattern of subgroup estimates is remarkably consistent: the program impact estimate increases across quartiles of the pre-treatment exam scores.

In contrast, the program narrowed the gap between low- and high-performing primary schools, because of a positive effect on the worst performing schools and negative effects on the best performing schools. The pattern of the impact estimates, moving from positive to negative across quartiles, is again consistent with reduced inequality.

Appendix Table 3: Estimates by quartile (of pre-treatment mean exam scores 2013-14) from Al-Samarrai et al. (2018)

Tables 5 and 7	(Table 5)	(Table 7)			
	ATE	Q1	Q2	Q3	Q4
Primary					
	-0.26	1.8***	0.08	-1.1***	-2.1***
2015	(0.25)	(0.34)	(0.33)	(0.32)	(0.31)
	-1.25***	1.57***	-1.16***	-2.66***	-2.82***
2016	(0.30)	(0.40)	(0.40)	(0.39)	(0.38)
	0.07	3.71***	0.58	-1.41***	-2.85***
2017	(0.29)	(0.41)	(0.40)	(0.41)	(0.38)
Secondary					
	2.61***	1.8***	3.0***	3.3***	2.1***
2015	(0.18)	(0.32)	(0.30)	(0.26)	(0.23)
	4.55***	2.29***	3.96***	5.36***	7.17***
2016	(0.38)	(0.47)	(0.41)	(0.52)	(0.48)
	4.34***		2.24***	5.05***	9.88***
2017	(0.46)	0.59 (0.50)	(0.51)	(0.61)	(0.63)

Notes: dependent variable is the mean exam score (percent)

Study 2: Lautharte, I., de Oliveira, V. H., Loureiro, A. (2021). Incentives for mayors to improve learning: evidence from state reforms in Ceará, Brazil. (Policy Research Working Paper No. 6694). World Bank. https://openknowledge.worldbank.org/handle/10986/35024

Program and evaluation

Lautharte et al. (2021) studies a performance-based reward aimed at (municipality) mayors in Ceará State, Brazil, the only design of this kind in our review. The performance reward was introduced with a 2007 tax reform that introduced an RBF instrument in the allocation of a sizable share (25 percent) of the "Quota Parte" consumption tax revenue to municipalities. This allocation was henceforth based on

a formula that included performance indicators for education, health, and environmental performance. Of these, the education index receives the highest weight (18/25). The calculation of the performance rewards is done by an independent think-tank using state government data.

The tax revenue obtained can be spent at will by mayors. The authors do not describe the importance of the tax reform incentives in municipal budgets. However, they do show (Table 6) that mean municipality per capita spending is 787 USD per capita or about 11 percent of mean per capita GDP. The Quota Parte tax distribution is described as an important revenue source for municipalities.

In the basic design, the education index is a function of two indices, Literacy (weight ⅔) and Learning; both are measured using state-managed assessments in primary schools. Moreover, the literacy index corrects for literacy level variation (punishing unequal literacy performance). Both the literacy and the learning index reflect the level and the change in learning.

The study analyses three RBF policy reforms: the first is the introduction of the basic RBF design in 2008. The second is the addition of a technical assistance package on top of the RBF scheme, (2009 for 5th grade, 2009-2011 for 9th grade). The technical assistance consisted of three main elements: (1) provision of structured literacy textbooks and teacher training; (2) "knowledge exchange practices", with training on school management, M&E of education policy, support to teacher career reform and meritocratic selection for principals; and (3) pedagogical action guided by learning assessment data.

The third is a reformulation of the RBF formula to "penalize municipalities with higher percentages of students performing below pre-defined minimal thresholds on test scores". In effect, the last reform specifically increases incentives in the lower tail part of the distribution by giving more weight to municipalities who manage to increase performance among students who are lagging behind.

The paper provides difference-in-differences impact estimates of these three reforms on test scores for Portuguese and Mathematics using a sample of schools located in the three immediate municipalities on both sides of the Ceará state border, with the control schools coming from the non-treated adjacent state municipalities. The test score data are from an assessment of students at the end of grades 5 and 9 (end of primary and lower secondary, respectively). This assessment is separate from the one used to calculate the municipality incentives. Equal trends in the pre-treatment years cannot be rejected.

Results

The mean results can be summarized as follows. First, the basic RBF reform had no impact on grade 5, but a significant positive impact of about 0.15 SD for both math and Portuguese in grade 9 (junior high). Second, the addition of the teaching assistance package nearly doubled the base reform impact in grade 9 and resulted in 0.16 SD (Portuguese) and 0.22 SD (math) test score improvements in grade 5; a large part of this effect is accounted for by the provision of textbooks. The reformulation of the RBF instrument (that penalized higher fractions performing below minimum levels) further improved test scores, with sizeable coefficient estimates of between 0.10 and 0.17 SD.

To address the distributional aspects of the reforms, the authors present results of quantile regressions. Overall, these results show that the base RBF design increased inequality, because the impact was concentrated in the upper part of the distribution.

Across grades and subjects, there was little to no impact on test scores in the lower parts of the distribution. The introduction of the teaching assistance package does benefit the lower performing students, but does not alter the regressive distributional impact: the impact estimates are a factor 2-3 higher for the 90th percentile students, compared to 10th percentile students.

Finally, the reform of the RBF formula is progressive and tends to reduce the impact differential across quantiles. In one instance (Portuguese in grade 9), the impact on the lower half of the distribution is in fact larger than for the upper part of the distribution. In the words of the authors: "Altogether, Figure 3 indicates that when the RBF formula conditions the redistribution of resources to the performance of students at the lowest bottom (minimum thresholds), it tends to reduce, but not eliminate, the learning enlarged by the RBF mechanism."

Appendix Table 4: Subgroup estimates from Table 2 and Figure 3 in Lautharte et al., 2021

Program	Test subject	ATE	SE	Subgroup estimates (Interpretation from graphs in Figure 3)
Tax incentive	Math G5 (graph)	0.036	0.4	Impact estimate increasing over quantiles, close to zero for Q1-Q3
Tax incentive	Portuguese G5 (graph)	0.06*	0.037	Impact estimate increasing over quantiles
Tax incentive	Math G9 (graph)	0.155***	0.026	Impact estimate increasing over quantiles
Tax incentive	Portuguese G9 (graph)	0.153***	0.023	Impact estimate flat, slight increase over quantiles
Tax incentive + TA	Math G5 (graph)	0.22***	0.029	Impact estimate increasing over quantiles
Tax incentive + TA	Portuguese G5 (graph)	0.162***	0.021	Impact estimate increasing over quantiles
Tax incentive + TA	Math G9 (graph)	0.148***	0.032	Impact estimate increasing over quantiles
Tax incentive + TA	Portuguese G9 (graph)	0.118***	0.028	Impact estimate increasing over quantiles
Tax incentive, revised	Math G5 (graph)	0.17***	0.038	Impact estimate increasing over quantiles
Tax incentive, revised	Portuguese G5 (graph)	0.166***	0.032	Impact estimate increasing then decreasing over quantiles
Tax incentive, revised	Math G9 (graph)	0.092**	0.035	Impact estimate decreasing then increasing over quantiles

Tax incentive, revised	Portuguese G9 (graph)	0.096***	0.028	Impact estimate decreasing over quantiles

Study 3: Contreras, Dante and Tomás Rau (2012) "Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile", Economic Development and Cultural Change, Vol. 61, No. 1 (October 2012), pp. 219-246

Program and evaluation

The Chilean National System of School Performance Assessment (SNED) is notable because it is at national scale and has operated since 1996. It includes all public and subsidized schools in Chile, both primary and junior secondary (basic education), and covers 6500 schools and over 90 percent of students.

The SNED is a teacher group incentive program in which schools compete against their peers on the basis of their average performance and in which monetary rewards are distributed equally among all teachers in the winning schools (10% used to pay outstanding teachers). Mean bonus in 1997 is 40% of monthly wage; increased to 80% in later years. The bonus is paid every two years.

The SNED is a threshold competition. Schools are ranked according to formula and top 25% of schools win (literally "Winning schools account for 25% of enrollment in each group"). The schools compete within homogeneous groups (within district, and within SES).

This program is in the "mixed" category because the bonus formula consists of 6 factors: 37 and 28 percent for learning level and improvement, respectively (based on student assessments in 4th and 8th grades); that is, 65% is based on objective learning measures. The other factors include: initiative/innovation, parental involvement, working conditions improvement, equality of opportunities.

The study uses private fee-paying schools as a control group and three different procedures to estimate treatment effects: a) using a nearest-neighbour matching estimator; b) using a difference estimator; c) using a panel estimator.

Results

The impact analysis shows large mean effects of between 0.16-0.25 SD (Math) and 0.14-0.25 (Language). We do not show all estimation results in the paper, but focus on the panel data results in Table 7. The estimated coefficient sizes in this table are similar to the (double robust) difference estimator and in the center of the estimate range according to the matching estimator.

The paper does not present estimates by baseline quantiles. Instead, it uses pre-treatment information to predict the post-treatment probability of winning following the competition formula. It then plots the ex-ante probability of winning against the difference between the actual and predicted exam scores (a measure of the program effect).

This graph (Figure 1 in the paper) shows a positive correlation between ex-ante chance of winning and learning impact. In the words of the authors: "It can be seen that the tournament seems to affect schools with a probability of winning greater than the 60th percentile. This suggests the existence of a large fraction of schools that do not respond to the incentive program." We conclude that there is a regressive effect on learning. We only provide one (averaged) mean treatment effect estimate because we only have one piece of evidence Figure 1) for the subgroup effect.

Appendix Table 5: Subgroup estimates from Contreras and Rau (2021)

	ATE (Mean of grade-subject results)	SE (Median of grade-subject results)	Subgroup estimates (Interpretation from Figure 1)
SNED	0.236	0.036	Test score prediction errors are strictly positively correlated with the probability of winning the competition.

Study 4: Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. American Economic Journal: Applied Economics, 2 (3), 205-27. https://economics.ucr.edu/pacdev/pacdev-papers/incentives_for_effort.pdf

Program and evaluation

The Glewwe et al. (2010) study provides evidence from a performance pay field experiment in the Busia and Teso districts in Western Kenya. The program was implemented by an NGO in a sample of low performing primary schools, in grades 4-8. Teachers in these grades are rewarded as a group, based on the school average performance on a district exam, with a penalty for students not sitting the exam. The incentive has a competition (with threshold design), as only the highest scoring half (24 out of 50) of treatment schools could win a prize. The prizes were in-kind, with the value of the prize increasing in the performance. One half of prizes was for performance based on absolute learning levels, and one half based on improvements, potentially creating more opportunities to win for low-performing schools. The prizes range in value from 21 to 43 percent of the typical teacher's monthly salary.

Results

The impact of the intervention on mean performance is limited to the formula metric, which is a function of the number of exam takers and the (high-stakes) test scores. There is no mean impact on non-incentivized test scores, nor on measures of teacher behaviour.

No subgroup (distributional) learning effects were reported. To assess whether there are any subgroup effects behind the mean effects, we used the published data and re-estimated two of the paper's tables (Table 2, panel A, reward formula; and Table 3, panel B, NGO exam), with interactions for quartiles of the initial test score distribution.

We present the additional estimates in the four tables below. We do not find any significant subgroup effects in either table. The interaction coefficients are small compared to the positive and highly

significant quartile dummies for Q3 and Q4. The signs of the interaction coefficients are consistent across the two tables: positive interaction estimates for quartiles 2 and 3, and negative for quartile 4. These results are relative to quartile 1, so do not suggest a consistent interaction across quartiles. We interpret these findings as representing a zero interaction.

Appendix Table 6: Program Impacts on Score on Teacher Reward Formula with dependent variable being score on formula used to reward teachers from Glewwe et al. (2010)

		Year 0 (1)	Year 1 (2)	Year 2 (3)	Year 3 (post-program) (4)
Model from the paper (Table 2, Panel A)	Incentive school	0.036	0.131*	0.215***	0.026
		(0.083)	(0.079)	(0.075)	(0.06)
	Observations	63812	76509	73789	57674
Model with interactions	Incentive school		0.149*	0.159*	0.075
			(0.086)	(0.095)	(0.092)
	Quartile 2		0.05	0.024	-0.015
			(0.055)	(0.063)	(0.065)
	Quartile 3		0.181***	0.187***	0.131*
			(0.059)	(0.07)	(0.073)
	Quartile 4		0.482***	0.562***	0.416***
			(0.078)	(0.081)	(0.097)
	Interaction Incentive*Q2		0.027	0.035	0.071
			(0.075)	(0.094)	(0.102)
	Interaction Incentive*Q3		0.031	0.009	0.072
			(0.08)	(0.097)	(0.112)
	Interaction Incentive*Q4		-0.083	-0.088	-0.115
			(0.103)	(0.113)	(0.145)
	Observations		30819	22341	9621

Notes: additional calculations for review paper

Appendix Table 7: Program Impact on Test Scores with dependent variable being score on NGO exam, from Glewwe et al. (2010)

		Year 0	Year 1	Year 2
Model from the paper (Table 3, Panel B)	Incentive school	0.048	0.092	0.026
		(0.093)	(0.086)	(0.095)
	Observations	33487	39966	26082
Model with interactions	Incentive school		0.048	-0.005
			(0.055)	(0.095)
	Quartile 2		0.282***	0.303***
			(0.028)	(0.05)
	Quartile 3		0.603***	0.597***
			(0.034)	(0.062)
	Quartile 4		1.272***	1.03***
			(0.076)	(0.093)
	Interaction Incentive*Q2		0.024	0.023
			(0.036)	(0.067)
	Interaction Incentive*Q3		0.026	0.044
			(0.043)	(0.074)
	Interaction Incentive*Q4		-0.042	-0.032
			(0.085)	(0.106)
	Observations		33397	22470

Notes: additional calculations for review paper

We also calculated subgroup interaction effects for girls. We present these additional estimates in the two tables below. We do not find any significant subgroup effects in either table. However, the sign is negative for each regression. We conclude that the subgroup effect is negative bur not significant.

Appendix Table 8: Program Impacts on Score on Teacher Reward Formula with the dependent variable being the score on formula used to reward teachers, from Glewwe et al. (2010)

		Year 0	Year 1	Year 2	Year 3 (post-program)
Model from the paper (Table 2, Panel A)	Incentive school	0.036	0.131*	0.215***	0.026
		-0.083	(0.079	(0.075	(0.06)
	Observations	63812	76509	73789	57674
Model with interactions	Incentive school		0.156*	0.232***	0.046
			(0.085	(0.083	(0.069
	Girl		-0.045**	-0.127***	-0.102***
			(0.021	(0.029	(0.03
	Interaction incentive*girl		-0.054	-0.034	-0.043
			(0.035	(0.051	(0.049
	Observations		76509	73789	57674

Appendix Table 9: Program Impact on Test Scores with dependent variable being score on NGO exam, from Glewwe et al. (2010)

		Year 0	Year 1	Year 2
Model from the paper (Table 3, Panel B)	Incentive school	0.048	0.092	0.026
		-0.093	(0.086	(0.095)
	Observations	33487	39966	26082
Model with interactions	Incentive school		0.053	0.008
			(0.054	(0.096)
	Girl		0.107***	0.105***
			(0.016	(0.030)
	Interaction incentive*girl		-0.01	-0.023
			(0.026	(0.038)
	Observations		33397	22470

Study 5: Barrera-Osorio, F., & Raju, D., (2017). Teacher performance pay: Experimental evidence from Pakistan. Journal of Public Economics, 148, 75–91. https://doi.org/10.1016/j.jpubeco.2017.02.001

Program and evaluation

This experimental program started in 2010 in three districts of Punjab, Pakistan, in a sample of 600 public primary schools. The Government implemented the program using administrative data. The program has three treatment arms (150 schools each) and a pure control arm (150 schools), with random assignment. The three treatment arms differ in the bonus beneficiary definition: in the first arm only head teachers are incentivized; in the second arm, both head teachers and other teachers are incentivized using the same bonus levels; in the third, both are incentivized but the head teacher unit incentive is twice the teacher unit incentive.

The mean bonus payment for subject teachers was between (21,000 and 23,000 rupees or) 7 and 9 percent of the annual baseline salary for the average teacher.

Serious thought went into the design, taking note of many issues discussed in the literature. The authors also note that this is a government-implemented program, which means that some design choices are constrained that may explain the lack of impact on the test scores (see below). We discuss a few of the design choices.

The performance pay formula is based on Grade 5 exams – so Grade 1-4 teachers behaviour affects these grades only indirectly and with a time lag. The formula pays for school level mean exam scores to include all students, but this can induce free-rider behavior: teachers are not paid at the grade-subject level for which they are directly responsible.

The bonus is based on a mix of metrics: test scores, enrolment and exam participation. The formula (Table 2) has a piece rate structure, with the total bonus an increasing function of: increases in test scores (weight 0.6); increases in enrolment (weight 0.25); and the level of exam participation (weight 0.15). There is no threshold, so the formula in principle incentivizes across the distribution of schools. The incentive includes exam participation rate (level) to avoid the exclusion of weaker students.

The analysis (and our summary) focuses on a comparison that pools the three treatment arms, i.e., comparing enrolment, exam participation and test scores across schools with and without any randomly assigned incentive treatment, for three years; see Table 4 in the original paper.

Results

Table 4 (original paper) shows that there are no significant mean effects on school enrolment, nor on test scores. For exam participation, there are positive significant mean effects in years 2 and 3, and a negative effect in year 1.

The authors analyse subgroup effects (for year 3 data, where the effects are largest) by modelling and predicting control and treatment outcomes and comparing these with observed outcomes. For learning, the graph is noisy with large fluctuations for both the control and treatment group; and there is no clear pattern suggesting that effects are concentrated in a certain part of the distribution of predicted scores. For enrolment, the graph suggests positive effects for larger (urban) schools.

Appendix Table 10: Subgroup estimates from Table 4 and Figure 1 in Barrera-Osorio and Raju (2017)

Mean effect	Subgroup effects
Pooled treatment	Based on Figure 1.C and description
Year 3	Year 3
0.0151 (0.07)	No clear pattern of impact across range of predicted exam scores.

Study 6: Ferraz, C., & Pereira, V. (2016) Can Students Benefit it if Teachers Lose their Bonus? Behavioral Biases Inside the Classroom. (Working paper as of August 5, 2016).

Program and evaluation

This paper provides an impact evaluation of a teacher bonus program in Pernambuco state, Brazil, that started in 2008. Program is part of an ambitious school accountability program. It is targeted at the school level: all school employees are paid for school performance, in both primary and secondary schools. The program is implemented state-wide and targets 950 schools, 50,000 teachers and 1.3 million students.

The bonus targets the teachers as a group: all teachers paid are paid, irrespective of their assignments or responsibilities. Payments are based on a school target that is set in terms of improvement in IDEPE index (= test score x pass rates). The program is implemented by the government, with testing organized by a specialized firm that hires external evaluators. The targets are set yearly by the department of education.

The incentivized tests are administered in grades 5, 9 and 12. The bonus is linear above threshold of 50% of target attained, up to 100% of target. The bonus value varied from 1.5 to three monthly teacher wages; the mean bonus is 221% of monthly wage in 2009 (or around 18 percent of an annual wage; see Bruns and Luque, 2015).

Results

The impact estimates are based on a regression-discontinuity (RD) design. The estimates reflect mean differences in outcomes between schools that barely missed the 50% threshold and those that just crossed the threshold. The mean impacts are sizeable. By pooling all tested grades for 2009, 2010 and 2011, the paper finds an RD impact of 0.134 SD (significant at 10%) for language, and 0.119 SD (significant at 10%) for math.

There is no explicit subgroup estimate, but there is suggestive evidence that the program may have had a regressive effect. First, school targets are set such that schools with lower past indexes are required to make bigger improvements (page 5). Figures 1 and 2 confirm this: on the left side of the bonus threshold, the distance from the threshold is generally negatively correlated with test scores. So weaker schools get higher improvement targets, and we can assume that larger improvements are

harder than smaller improvements; and, therefore, that weaker schools are more likely to miss their improvement target.

Second, Table 10 (p. 39) shows that there is a discouragement effect: a school that barely misses the 50% threshold for the first time has a strong positive effect on test scores; but this is not the case for schools that have previously missed the target. And the further a school is away from the target, the more negative is the math test score coefficient for schools that had already missed the target (but these coefficients are insignificant).

Appendix Table 11: impact on test scores by past experience in almost winning the bonus from Table 10 in Ferraz and Pereira (2016)

	Window: 30% to 49%		Window: 20% to 49%		Window: 10% to 49%	
Ever missed the threshold with achievement in window						
	No	Yes	No	Yes	No	Yes
A: Language test scores						
RD estmate	0.241***	-0.019	0.267***	-0.01	0.300***	0.005
Std error	(0.092)	(0.097)	(0.104)	(0.078)	(0.106)	(0.070)
Observations	235	84	204	145	186	175
B: Math test scores						
RD estmate	0.216**	0.039	0.232**	-0.038	0.299***	-0.06
Std error	(0.092)	(0.095)	(0.099)	(0.082)	(0.100)	(0.07)
Observations	235	84	216	154	191	187

Teacher Incentive summaries

Study 7: Loyalka, P., Sylvia, S., Liu, C., Chu, J., Shi, Y., (2019). Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement. Journal of Labor Economics 37, 621–662. https://doi.org/10.1086/702625

Study 8: Chang, F., Wang, H., Qu, Y., Zheng, Q., Loyalka, P., Sylvia, S., Shi, Y., Dill, S. E., & Rozelle, S. (2020). The Impact of Pay-for-Percentile Incentive on Low-Achieving Students in Rural China. Economics of Education Review. 75 : 101954. https://doi.org/10.1016/j.econedurev.2020.101954

Programs and evaluation

Loyalka et al. (2019) and Chang et al. (2020) report on randomized experiments conducted rural schools in designated poverty counties in Western China over the period from Sept 2013 - May 2014 (Loyalka et al. 2019) and Sept. 2014-Sept. 2015 (Chang et al. 2020). As the two studies have been conducted in the same set of schools, and test similar interventions, we summarize their results jointly. Both studies focus math in grade 6 of primary school.

The study design is copied in Appendix Table 12 below. The cells indicate the number of schools. In the Loyalka study there are three treatment groups: Pay for levels, pay for gains, and pay for percentile. In each, teacher pay is based on a ranking of their performance relative to their peers, but the measure on which they are ranked differs across intervention arms. In the levels it is average endline test score, in the gains it is the average gain in test scores, and in the pay for percentile, it is the average percentile of the endline test score of students when compared to the distribution of similar students in the control group. The Chang study is a follow up study in the Pay for Percentile and control schools. In the pay for percentile schools, Chang implements a modified Pay for Percentile scheme which provides a higher reward for learning gains of students who started at a lower learning levels. Note that the cohort of students across the studies is different, as both studies focused on grade 6, but were implemented in different years.

Appendix Table 13: Experimental design of Loyalka and Chang studies, with the number of schools in each treatment cell

	Control	Levels		Gains		Pay for Percentile		Total
		Low	High	Low	High	Low	High	
Loyalka	52	28	26	30	26	28	26	216
Chang	Control					Modified Pay for Percentile		
	51					52		103

In Loyalka payment to teachers were set equal to 3500-(99-P)*35 for teachers in low incentive schools, and 7000-(99-P)*70 in high incentive schools. In Chang, the payment was set equal to 5000-(99-P)50+3000-(99-PB)*30, where P is the average percentile rank of the teacher in the treatment group. For the levels treatment, rank is based on the average endline test score. For the Gains treatment, the average endline-baseline test score is used, and for the Pay for Percentile, percentile rank of the students endline test score is used, where the percentiles are based on students with similar baseline test scores. Note that the distribution of payments to teachers is the same across treatments, as each incentive scheme used the rank of the teachers within an incentive group as the basis to make payments. In Chang, the teachers receive an additional award based on average percentile of the students in the lowest tercile at baseline (PB).

The median payment is 1750 in Loyalka for the low incentive, 3500 for the high incentive, and 4000 in the Chang experiment. It is a linear scale, with the lowest ranking teacher receiving almost nothing. The incentive was around one month salary. In Loyalka, the average monthly baseline salary is 2852 Yuan in the control group (table A1). So the incentive in Chang is the most generous.

Results

The main effects of both papers are reported in the appendix table 12 below. Because the Loyalka paper is not sufficiently powered to analyze the interaction of the choice of performance measure and

the size of the reward, these results are presented separately from one another. The Levels and Gain intervention have no significant impact ton learning. The pay for percentile treatment raises test scores by 0.15 std deviations. There is no significant difference between the low and high reward. The modified Pay for percentile continues to deliver perform in the second year, but with a lower point estimate of 0.10 SD.

Appendix Table 14: main impacts from Loyalka et al. and Chang et al.

	Levels	Gains	Pay for percentile	Low reward	High reward
Loyalka et al.	0.084 (0.052)	0.001 (0.050)	0.148** (0.064)	0.081 (0.055)	0.067 (0.046)
Chang et al.			0.10* (0.06)		

Note: Copied from table 2 of the Loyalka and table 3 of Change study. We use the main results with controls

We focus the discussion of heterogenous treatment effects on the type of incentive only. The table below shows the estimated differences in impacts by gender, rank of baseline test score in classroom, and teachers' perceived value added for the student for the Loyalka study. Girls seem to benefit less from the interventions across all treatment, but the difference is never significantly different from zero. Columns 3 and 4 show how the impact differs by baseline ability of the student. Students are ranked within their classrooms, in three groups based on their baseline test score. All the effects are negative, suggesting that students in the lowest ability rang benefit most from the interventions. The effects are however never significantly different from zero, and there is also no sign of a greater focus on the lowest ability students in the level incentive treatment (as one would expect, theoretically). The last two columns show how the impacts differed by "teacher benefit". To this end, for a random subset of 12 students, teachers were asked how much they though the student would benefit from an extra hour of private lessons. For the levels and Gains incentive, we find that students who the teacher believes would benefit, are also significantly benefiting more from the intervention. This is not the case for the pay for percentile treatment. The last column shows whether the impacts differ by student wealth based on an asset index. This is never the case.

Appendix Table 15: Heterogenous treatment effects in Loyalka et al.

	ATE (1)	Girls (interaction term) (2)	Medium ability (interaction term) /a (3)	Higher ability (interaction term) /a (4)	Medium teacher benefit (5)	High teacher benefit (6)	Richest half (7)
Levels	0.084 (0.052)	-0.037 (0.050)	-0.026 (0.059)	-0.071 (0.060)	0.053 (0.111)	0.213** (0.122)	-0.033 (0.048)

Gains	(0.001) (0.050)	-0.050 (0.050)	-0.031 (0.059)	-0.041 (0.064)	0.163 (0.0146)	0.333** (0.152)	-0.002 (0.052)
Pay for Percentile	0.148** (0.064)	-0.062 (0.057)	-0.055 (0.065)	-0.063 (0.082)	0.056 (0.139)	0.056 (0.151)	-0.010 (0.059)

Notes: Columns 2-6 are copied from table 6 in the Loyalka paper, Column 7 are based on authors calculations using the replication files. /a Note that ability level is based on within in class rank

The pay for percentile does not provide an explicit incentive for teacher to focus on weaker students. A social planner may find this an undesirable feature. The experiment reported in Chang provides an example of a Pay for percentile level where the rewards are set higher for gains of students in in the bottom tercile of the baseline ability level. Theoretically, one would expect this to result in higher impacts for weaker students.

The Chang paper reports separate impacts for three ability levels, defined on the basis of the ability observed in the control group (so across the entire distribution). No replication files are available. For the purpose of the comparison, we report similar results for the Loyalka paper, based on our own calculations using the replication files.

Appendix Table 16: Comparison of heterogenous treatment effects by baseline ability in Loyalka et al. and Chang et al. papers

	Bottom tercile	Middle tercile	Upper tercile
Loyalka et al.	0.112 (0.100)	0.217** (0.680)	0.117 (0.73)
Chang et al.	0.15** (0.07)	0.03 (0.07)	0.10** 0.06)

Notes: copied from table 4 from Chang paper, and authors calculations using replication files for Loyalka study.

The comparison indicates that the incentive introduced in Chang indeed resulted in teachers paying more attention to students in the lower tercile of the ability level. In Loyalka, the middle tercile is the group that benefit most, in Chang it is the lower tercile. Note that these patterns were not apparent when we focused on the within class rank of the students in the Loyalka paper. This suggests that the Chang paper motivated teachers with many students in the lower ability range to put in more effort.

Study 9: Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. American Economic Review, 102 (4), 1241–1278. https://doi.org/10.1257/aer.102.4.1241

Program and evaluation

Duflo et al (2012) evaluated a pilot where teachers were paid based on presence as evidenced by pictures they were required to take at the start and end of the school day. The pictures, which were transmitted online, substantially reduced the cost of monitoring teachers' presence. The experiment

was implemented in schools which were run by a private NGO in India, making it relatively easy to adjust salaries based on recorded absence. Teacher absence at baseline was 35 percent. The study started with a baseline in August 2003, a mid-test in April 2004 and a post-test in September 2004. There were 57 schools in the treatment group, 60 in the control. The incentive has a linear scale. Treatment teachers receive a 500+50(days worked-10) per month, teachers in control schools receive a flat wage of 1000. Therefore, with 20 days worked, both teachers earn the same. With about 20 workdays in a month, this is mostly an incentive which punishes teachers for not showing up.

This is a serious pay cut. With a presence rate at endline of 78 percent, 21 work days, teachers in treatment schools worked about 16,4 days worked, and received 12(500+50(16-10))=9600 per year, whereas teachers in the control schools received 12000. The average pay cut is thus about 2.4 months' salary.

Results

The intervention resulted in an increase in teacher presence, and in test scores. The main results are in Appendix Table 17 and Appendix Table 18 below. All columns report the subgroup effects for the group indicated in the label. The study resulted in a 0.17 SD increase in test scores. The impact was substantially higher for those who took the baseline test in written form, that is, the students who were literate at baseline. Boys and girls benefited equally from the intervention. The study did not report subgroup effects by wealth. Teacher presence increase from baseline to midline by 21 percent point as a result of the intervention. This effect was stronger for teachers who taught classes with below median test scores. This was a catching up effect, by the endline the teacher attendance rates were similar in both groups.

Appendix Table 19: Impact on test scores in Duflo et al. (2012) paper

	All	Took pretest oral	Took pretest written	Girls	Boys
Pay for presence (camera)	0.17* (0.09)	0.16 (0.10)	0.25* (0.13)	0.17* (0.09)	0.16 (0.10)

Notes: standard errors in parenthesis, source table 9

Appendix Table 20: Impact on teacher presence Sept 2003-Feb 2006 in Duflo et al. (2012)

	All	Teachers with above median test scores	Teachers with below median test scores
Pay for presence (camera)	0.21*** (0.03)	0.15*** 0.04	0.24*** (0.04)

Note: standard errors in parenthesis, source table 2

Study 10: Gaduh, A., Pradhan, M., Priebe, J., & Susanti, D. (2021). Scores, Camera, Action: Social Accountability and Teacher Incentives in Remote Areas. (Policy Research Working Paper No. 9748). World Bank. https://openknowledge.worldbank.org/handle/10986/36112

Program and evaluation

Gaduh et al. (Gaduh et al. 2021) test the impact of three interventions: (1) community monitoring (SAM), (2) Community monitoring + a teacher pay incentive based on teacher presence as recorded by camera (SAM+CAM) and (3) Community monitoring + teacher pay incentive based on outcome of community monitoring (SAM+Score), in remote public schools in Indonesia. The community monitoring, or social accountability mechanism (SAM), was against standards agreed by the community and teachers through a facilitated process. Monitoring results were discussed in monthly public meetings. Teacher presence was always included in the list of indicators. In the second and third intervention, a teachers' remote area allowance could be cut if performance fell short of standards. In Sam+Cam, only teacher presence was considered for the pay cut, in the Sam+Score, the composite score resulting from the community monitoring was used. Note that not all teachers received the remote area allowance. Those who did not were not directly affected by salary incentives in treatments 2 and 3.

The actual pay cuts were relatively small: incentivized teachers in SAM+Score received an average pay cut of around 6.9 percent, whereas those in SAM+Cam received an average cut of 10.1 percent. As the remote area allowance is about half of total teacher pay, these translate into a (12*0.5*6.9)=41 percent of a monthly wage on a yearly basis for the SAM+score and a *12*0.5*10.1=61 percent of a monthly wage pay cut on a yearly basis for the SAM+Score.

The study was implemented in 270 schools, in 2 districts, in remote villages. The schools were equally distributed across the three intervention groups and a control group. A baseline and endline survey were fielded in Nov 2016 and March 2018. After that, the project was handed over to the local government and communities. A follow up survey was conducted in April 2019, in all villages except those who received the Cam+Score treatment.

Results

The main results are in Appendix Table 21 below. While all treatments resulted in learning improvements, the SAM+CAM treatments clearly outperformed the other treatments. The learning impacts were also sustained to the second year. The results indicate that boys benefited more than girls, but the difference is small and not significant. Students who scored above the median benefited more than those who scored below. The point estimates for SAM and SAM+CAM are substantial, but not significant. The study does not report impacts by wealth of students.

Appendix Table 22: Impacts on learning in Gaduh et al. study

	Main effects		Male interaction term		Above median student	
	End Line	Follow up	End Line	Follow up	End Line	Follow up
SAM	0.084**	0.028	0.034	0.011	0.031	0.055
	(0.044)	(0.032)	(0.029)	(0.037)	(0.047)	(0.044)

SAM+CAM	0.198***	0.133***	0.012	0.020	0.064	0.045
	(0.036)	(0.034)	(0.029)	(0.037)	(0.044)	(0.042)
SAM+Score	0.110***		0.018		-0.008	
	(0.033)		(0.030)		(0.043)	

Note: effect on average Indonesian and math score reported. Copied from table 3 and 4.

Study 11: Andrabi, T., & Brown, C. (2021). Subjective Versus Objective Incentives and Teacher Productivity. (Working Paper as of March 11 2021).

Program and evaluation

Andrabi and Brown (2021) compared two incentive mechanisms in the context of urban private schools, which cater to the middle class. In the *subjective* pay for performance scheme, teachers were held accountable to performance standards which their managers had communicated to them beforehand. In the *objective* pay for performance, teachers were evaluated on the basis of learning improvements of students in their class, using a Pay for Percentile mechanism. In all cases, teachers were ranked within schools, so every school received the same budget. Most of the teachers, from the 16th to the 60th percentile received a 5 percent wage increase. Below that, the wage increase was either 0 or 2 percent. Above it, it was either 7 or 10 percent. Control teachers received a 5 percent flat wage increase. There are 234 schools in the experiment. 145 received the subjective treatment, 40 the control and 32 the objective treatment. The incentives applied to math language and science teachers teaching grade 4-13. Timeline: February 2018 to January 2019, so one year effects.

Results

Appendix Table 23 provides the main results. Both treatments have similar effects on test scores. There is weak evidence that the objective treatment resulted in a drop in socio-emotional welfare of students (table 4 in paper, not copied here). The heterogenous treatment effects indicate very little differentiation of impact by baseline test score for the pay for percentile treatment. We see this both if we compare by student baseline score, or by school baseline score. The point estimates indicate that for the subjective treatment, student who scores low at baseline benefited more, which could indicate that managers encouraged teachers to focus more on weaker students. The interaction effects are however always insignificant.

Appendix Table 24: Impact on test scores in Andrabi and Brown

	All	Student baseline score	School baseline score
Objective treatment	0.0918**	0.00183	-0.0199
	(0.0575)	(0.0243)	(0.0165)
Subjective treatment	0.0859**	-0.0345	-0.0681
	(0.0220)	(0.0290)	(0.120)

Notes: Table 3, and calculations by Christina Brown at request of the authors

Study 12: Gilligan, D.O., Karachiwalla, N., Kasirye, I., Lucas, A., & Neal, D.A. (2018). Educator Incentives and Educational Triage in Rural Primary Schools. Journal of Human Resources. 10.3368/jhr.57.1.1118-9871R2

Program and evaluation

Gilligan et al. (2018) is fairly large experiment in Uganda using pay for percentile. There are 2 arms. The experiment is implemented in primary schools in Uganda. There are 7 grades and very high dropout rates. One of the reasons for the high dropout is that teachers are held accountable for high passing rates at the primary school leaving national exam (PLE). This creates an incentive to let only best students enter grade 7. Administrative data support the high dropout that increases with grade level (promotion rate is about 60 percent at higher grades), but only anecdotal data indicates that there is a discontinuity from grade 6 to 7.

There are 150 treatment and 150 control schools. In the treatment schools, a pay for percentile treatment was implemented which was based on the math assessment that the researchers implemented in March and October 2016 for grade 6 students. Test covered questions from all grade levels. So the time span is only 6 months. A third round of data collection was done in oct 2017, to see whether children were promoted to grade 7, but no test was done. In addition, they have data from the primary school leaving exams. The intervention rewarded a teacher based on a linear function of the percentile performance of the student if compared with students with the same baseline test level. If the child was not present for the second test, it was awarded as if all questions were wrong (for example, if 20 perc of the grade bucket did not show up, the award was set equal to the 20^{th} percentile for a missing student). The median performance-based pay was 10,000 per student and the maximum 20,000. The median payment for a teacher with a class of 33 students was 329,000 shillings, which is about 1 month pay for an average teacher.

The hope is that this experiment will induce teachers to also give attention to children of lower end of the ability distribution. In the control group, the incentives are skewed towards focusing on the higher end. Note however that there is no explicit incentive in the experiment to reduce drop out from grade 6 to grade 7. So one could argue that possible reductions in the dropout rate are resulting from students performing better in grade 6, not from teachers lowering the standards. The authors also explicitly designed the experiment to study whether the impact depended on whether math books were available at school. Through stratification, the authors ensured that half of the sample did, and the other half of the sample did not have math books available at baseline in grade 6. The hypothesis is that students with textbooks can better take advantage of the additional effort of teachers.

Results

Appendix Table 25: Impact results in Gilligan et al.

	ATE	Boys	Girls	Without books	With books	Lower ability	Higher ability
Present round 2	0.018	0.017	0.018	0.022	0.014		
	(0.017)	(0.022)	(0.020)	(0.025)	(0.023)		
Test score round 2	0.018	0.005	0.029	-0.031	0.072	-0.001	0.038
	(0.030)	(0.036)	(0.034)	(0.036)	(0.047)	(0.031)	(0.040)
Test low grade items	0.003	-0.012	0.014	-0.023	0.032	-0.002	0.011
	(0.028)	(0.035)	(0.033)	(0.036)	(0.044)	(0.034)	(0.038)
Test high grade items	0.042	0.036	0.049	-0.029	0.118**	-0.002	0.067
	(0.035)	(0.042)	(0.040)	(0.042)	(0.056)	(0.034)	(0.048)
Present round 3	0.042**	0.041*	0.042**	0.013	0.072***		
	(0.018)	(0.022)	(0.021)	(0.026)	(0.025)		
Took PLE	0.023	0.036	0.011	0.019	0.026		
	(0.019)	(0.023)	(0.022)	(0.026)	(0.029)		
Passed PLE	0.010	0.007	0.010	0.009	0.008		
	(0.018)	(0.021)	(0.021)	(0.024)	(0.026)		

Notes: the source is Table 3 and 4

We see no effect on attendance in round 2, which also was not expected. We observe no overall impact on test scores, but some patterns emerge from the data. Generally, the small positive effects are higher for the items at the appropriate grade level, suggesting that students which were at that level benefited from the intervention, and not the students which had fallen behind. Schools with books clearly benefited more than those without, suggesting a complementarity between teacher effort and books. Higher ability students also benefit more. The strongest effects on learning, not included in this table, are found in schools with books, for students above the median achievement level at baseline. For them the impact on learning is 0.18*** std dev. There are hardly gender differences. If at all, girls benefit slightly more than boys.

The results also clear positive effect on teacher effort (Table 5, not copied here)

We do see an effect on attendance in grade 7, with no gender difference, but again a stronger effect in schools with books. There are no significant impacts on taking or passing the PLE. Breaking this further down, it seems like the program had a stronger effect on boys, but only in schools with books. For them the prob of taking the PLE increased by 0.067** (table 3b). This is interesting, because the test score results suggested that girls benefited more (all weakly) in schools with books (Table 4b). Boys have higher test scores in the control group. So probably the effect on PLE taking for boys is higher because they were closer to the threshold needed to pass to the next grade.

Overall, a very carefully done and nice paper. It is too bad that the performance incentives were based on such a small time interval, which may explain the low point estimates. Otherwise, it is a great study. The intervention had some effects, but still nowhere close to solving the problem. Only 56 percent of the students attended the round 3 test, so an impact of 0.04 is only a small step. The results clearly suggest that performance pay works better in better resourced schools, and for better students. So even though the incentive is set up to direct attention to all students, these are the students that benefit most from the intervention.

Study 13: Muralidharan, K. & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. Journal of Political Economy, 119 (1), 39-77. https://doi.org/10.1086/659655

Study 14: Muralidharan, K. (2012). Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India. (Working paper for Society for Research on Educational Effectiveness). https://eric.ed.gov/?id=ED530172

Program and evaluation

Muralidharan and Sundararaman (2011) and Muralidharan (2012) report on a long-term experiment with performance pay based on test scores in a representative sample of government primary school in Andra Pradesh, India. Teachers were paid bonuses based on percentage point (these are absolute increases, not pay for percentile) increases in the average increase in test scores in the classroom (individual incentive) or in the school (group incentive). There were no sanctions in the case of a decrease. Teachers were paid 500Rp for every percentage point. Average yearly payout was about 3 percent of salary, so this was a small bonus. The original experiment had 100 schools in each group (individual, group, control). After 2 years, a random subset of 50 schools of the 2 intervention groups continued the intervention for another 3 years. For these schools, the bonus formula was also changed. Instead of paying for gains in test scores, the interventions paid for a (student score-target student score), where the target score was a predicted score based on students' characteristics, using a model estimated on the control sample. Again, teacher bonuses remained positive, but individual students could contribute negatively to the bonus if they scoreD below the target score or were absent at the test.

Results

Appendix Table 26 and Appendix Table 27 shows the impact for the individual and group incentive respectively. The individual incentive treatment has sustained impacts, the group incentive were lower and less sustained. So, the free riding overrides the need for collective action. There is not much going

on in terms of heterogenous treatment effects. Estimates reported are based on the full sample, which includes incoming cohorts. Household affluence is the only indicators which consistently has stronger impacts of the intervention.

Appendix Table 28: Impact and heterogenous treatment effect of individual incentive bonus in Muralidharan and Sundararaman (2011) and Muralidharan (2012) papers

			Individual incentive		
Year	1	2	3	4	5
Effect on test scores	0.154 (0.045)***	0.204 (0.050)***	0.191 (0.056)***	0.331 (0.072)***	0.444 (0.101)***
Interaction effects					
Household Affluence	0.023 (0.020)	0.004 (0.021)	0.032 (0.021)	0.013 (0.031)	-0.017 (0.044)
Male	0.006 (0.028)	-0.042 (0.032)	0.044 (0.045)	0.06 (0.057)	0.066 (0.102)
Baseline Test Score	0.002 (0.032)	0.048 (0.037)	0.031 (0.045)	0.006 (0.076)	

Notes: the source is Table 5 and 8a in Muralidharan (2012)

Appendix Table 29: Impact and heterogenous treatment effect of group incentive bonus in Muralidharan and Sundararaman (2011) and Muralidharan (2012) papers

			Individual incentive		
Year	1	2	3	4	5
Effect on test scores	0.106 (0.044)**	0.061 -0.049	0.089 (0.051)*	0.123 (0.067)*	0.129 -0.085
Interaction effects					
Household Affluence	0.032 (0.018)*	0.031 (0.020)	0.040 (0.021)*	0.008 (0.037)	0.032 (0.049)
Male	0.018 (0.028)	0.022 (0.038)	0.004 (0.045)	-0.021 (0.083)	-0.020 (0.120)
Baseline Test Score	0.015	-0.002	0.044	-0.063	

	(0.030)	(0.038)	(0.046)	(0.078)

Source: Table 5 and 8a in Muralidharan (2012)

Study 15: Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., Rajani, R. (2019a). Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania. The Quarterly Journal of Economics, 134 (3), 1627-1673. https://doi.org/10.1093/qje/qjz010

Program and evaluation

Mbiti (2019a) report on an experiment which provided unconditional grant, teacher incentives, and a combination of the two to primary schools in Tanzania. The experiment included 350 schools, of which 70 where in each of the three treatment groups, and 140 were in the control. The experiment continued for two years, and focused on teachers in grade 1-3. For the teacher incentive, teachers were paid a fixed amount for the number of students that participated in the baseline, and passed a test at the end of the school year. As this study focuses on pay for performance incentives, we will not discuss the grant only arm of this study.

The direct costs for the incentive program are 2.52 USD per student, for the combined program it is 8.71 per student (table A14). A teacher has on average around 58 students (table 1). So the cost per teacher per year are 2.52*58=146 for the teacher incentive and 8.71*58=505 US dollar for the teacher incentive and school grant combined. Teachers' monthly salary is about 312 US dollar, so the programs cost about 3.9 percent and 13.5 percent of annual teacher pay, respectively.

The results are summarized in Appendix Table 30. The teacher incentives with a school grant component clearly outperform the teacher incentive alone. Note that the paper also reports that the school grant alone does not have an impact on learning, so the complementarity between additional resources and teacher effort is likely to explain the higher impacts. For this combined incentive plus grant program, we find that girls benefited more from the intervention, and so did those with lower baseline scores.

Appendix Table 31: Table Impact on learning as reported in Mbiti(2019a)

	ATE	Male	Baseline ability (lagged test score)	Wealth
Teacher incentive				
Y1	0.06*			NA
	(0.04)	-0.07*	-0.01	
Y2	0.03	(0.04)	(0.02)	NA
	(0.04)			
Teacher incentive and unconditional grant				
Y1	0.12***	-0.10**	-0.06**	NA
	(0.04)	(0.04)	(0.03)	

Y2	0.23*** (0.04)	NA

Notes: the source is Table IV for main effects. Heterogenous treatment effects are from Table VII and are based on data from both years.

This paper is one of the few that uses a threshold design to pay teachers. Teacher were paid a fixed amount depending on whether the student passed the test, independent of baseline ability. Theory would predict that to maximize their bonus, teacher would focus on those who are likely to pass close to the threshold. The authors test this hypothesis by seeing whether the effect differs depending on how far students are away from the threshold that would make them pass. If anything, the results indicate the opposite (Table A12). The interaction of distance with the treatment is never significant, for both treatments. Combining all treatments, subjects and years, and just looking at the sign of the point estimates, there are 9 estimates that indicate that the impacts become larger as the distance to the threshold increases (and 2 estimates that indicates they become smaller). This result is consistent with the negative interaction effect with baseline test scores, which is significant for the combination arm.

Study 16: Romero, M., & Schipper, Y. (2019b). Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania. (NBER Working Paper No. 25903). National Bureau of Economic Research. https://www.nber.org/papers/w25903

Program and evaluation

Mbiti, Romero, and Schipper (2019b) report on a large experiment on the effect of different teacher incentives in rural Tanzania, public schools. They compare threshold-based regimes against pay for percentile. The intervention is 2 years. There are three groups, each of 60 schools: Control, levels and pay for percentile. The levels treatment has 3 different levels, where a higher amount is provided if the student reaches a more advanced level. There are 3 to 5 different thresholds defined for each grade. All treatments are competitions, that is, the average award amount for a teacher is kept constant across treatments. The average bonus equal to 3.5 percent of teachers annual salary, or half a month's pay.

For the threshold intervention, theory predicts that teachers will focus on those students were the relationship between reward and effort is the steepest. These are mostly likely the students around the threshold, for whom with little effort it can be ensured that they pass the threshold. With multiple thresholds, the same prediction holds, but effort will around these thresholds. A uniformly distributed set of thresholds across the ability distribution, will result in similar effort for all students regardless of their ability. As higher thresholds have a higher payout, one would expect better scoring students, who are closer to the higher thresholds, to benefit more from this intervention

For pay for percentile we expect teachers to devote equal attention to all students on average, as they are all in an equal competition. Teacher may vary their attention depending on their ability and the composition of the classroom. But it will always be relative to other teachers, so teachers have an incentive to focus on the group of students for which they have a comparative advantage (because, for

example, they are better able to teach these students than other teachers, or they have a lot of a certain type of students in their class), but this can never lead to a systematic bias across all schools for a certain type of student.

Results

Overall, the levels incentive outperforms the pay for percentile incentive. We also observe stronger effects for the incentivized test score, which is used to calculate the bonus, when compared with a low stakes test score. Both incentives do somewhat better in the second year. The heterogenous treatment effects are not clear. They are present for math in pay for percentile by baseline test score (better scoring students benefit more). This must be the first-year effects, as in the graphs this is seen only for the first year. In year 2, the positive gradient with respect to baseline test score is visible anymore in the graphs. All the interactions with baseline test scores are positive for the Levels intervention, which is in line with the hypothesis. This however is also the case for the pay for percentile, for which we did not expect this. The point estimate is even lightly higher. This suggests is not driven by the payout scheme of the level threshold, but more a factor which is common across the two incentive payment mechanisms, like complementary with teacher effort.

Appendix Table 32: Impacts on low stakes test scores in Mbiti, et al. (2019b)

	Language + math combined		Year 1-2 heterogenous treatment effects			
	ATE Y1 Lang+math combined	ATE Y2	Male math (Interaction term)	Male language	Baseline test score Math	Baseline test score language
Levels	0.057	0.096**	-0.022	0.017	0.034	0.015
	(0.052)	(0.046)	(0.037)	(0.051)	(0.032)	(0.029)
P4Perc	-0.029	0.044	0.020	-0.024	0.068**	0.030
	(0.039)	(0.044)	(0.041)	(0.051)	(0.026)	(0.030)
Level (incentive)	0.17**	0.22**				
	(0.064)	(0.059)				
P4P (incentive)	0.059	0.13**				
	(0.54)	(0.56)				

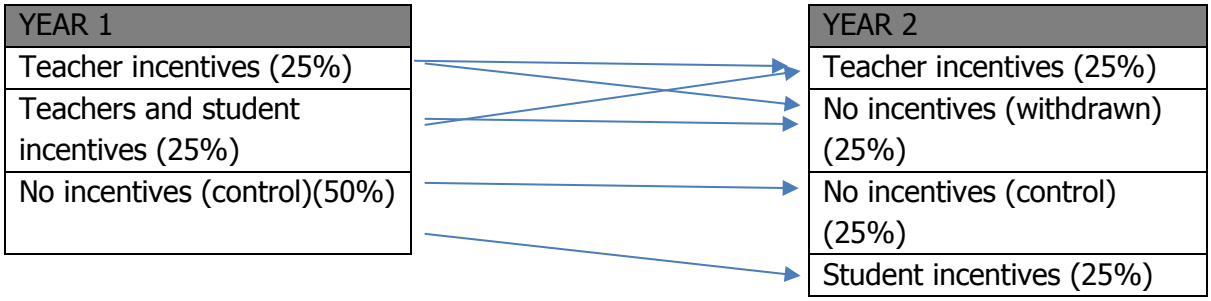
Notes: based on Table 3 and appendix Table D.12. Heterogeneity effect estimates are based on data from both years.

There are no clear effects on teacher effort. The only thing that comes out clearly is that teachers who of themselves believe they can make a difference in the students learning outcomes react stronger to both treatments.

Study 17: Filmer, D. P., Habyarimana, J. P., & Sabarwal, S. (2020). Teacher Performance-Based Incentives and Learning Inequality. (Policy Research Working Paper No. 9382). World Bank. https://openknowledge.worldbank.org/handle/10986/34468

Program and evaluation

This study evaluates the impact of a program which provided in kind prizes for the best performing teachers and/or students in lower secondary schools in Tanzania. It is a large experiment, with a sample of 400 schools, of which 83 percent are public. The intervention focused on teachers and students in grade 10. The study started in 2013 and lasted 2 years. Note that the cohort of students is different for each year, but teachers often be the same. The graph below indicates the experimental design and the allocation of schools to treatment arms:



The authors analyze the experiments in year 1 and year 2 separately. The "No incentives (withdrawn)" is analyzed as a separate treatment group to investigate whether teachers reduce their effort once the incentives for them are withdrawn.

The teacher incentives let teachers compete on the average test score improvement of their students. They compete in a group of around 10 schools/teachers. There are three prizes for the top performers, with a value of $190 for the first place, and 120 and 110 dollars for the 2nd and third. According to the paper, the yearly cost of teacher program is $442. In terms of award money, it is (190+130+110)/10=$43 per teacher. Monthly income is 376. So the program costs 442/(12*376)=9.7 percent of annual income and 43/(12*376)=0.9 percent of annual income in terms of reward money. It is not clear why the cost of the program reported in the paper exceeds the reward money by tenfold.

Students compete with other students in the same class. The top 3 students in the class (greatest gain in test scores) received a price. The cost reported in the paper is a quarter of the teacher reward cost, consequently, the price students receive imust be much smaller than that of teachers.

The paper reports heterogenous treatment effects by whether or not the schools is in the top 30 percent in baseline test scores. No heterogenous treatment effects are reported by baseline test score of students, but the paper does report conditional quantile regressions, which tells us something about how the distribution of the test score changes over time. The paper does not control for baseline student test scores, which is very standard in this literature. This unusual way of presenting result indicates the authors were not able to match student test scores over time at the student level.

Results

The paper reports heterogenous treatment effects by whether or not the schools is in the top 30 percent in baseline test scores. No heterogenous treatment effects are reported by baseline test score of students, but the paper does report conditional quantile regressions, which tells us something about

how the distribution of the test score changes over time. The paper does not control for baseline student test scores, which is very standard in this literature. This unusual way of presenting result indicates the authors were not able to match student test scores over time at the student level.

The impacts on learning are copied in the Appendix Table 33 below. The first table indicates the 1st year effects, the second table the second-year effects.

In the first year, the program only seems to work if student and teacher incentives are combined (0.16 std dev ATE). There are no effects of the teacher incentive only. The interventions shifts the entire test distribution to the right, but more at the upper tail. Under the assumption of no rank reversal, this indicates that the better off students benefited more.

The first-year effects suggest that students-level incentives make the difference. It is therefore puzzling why the student level effects only treatment, introduced in the second year, does not have a significant effect on learning. The student incentives do seem to motivate the higher scoring students. Better schools show strong effects (0.22 std dv ATE), though not significant. Similar effects, but significant, are shown in the quantile treatment

The teacher incentives are more puzzling. Schools where teachers who continue to receive the teacher incentives now show impacts, similar in magnitude as the ones for the teacher and students incentives observed in the first year. Even more surprising is that withdrawing the teacher incentive (which was not effective in the first year) has a positive impact (but not significant) in the second year. It almost seems as if teachers respond in a delayed way to the treatment. Possibly they respond not as much to the announcement, but more on the news to the prizes being rewarded.

This study is interesting as it works with prizes for the a few winners, and nothing for the rest. This threshold type of incentive should incentive those who have a good chance to be among the winners. With rankings on test score gains(teachers), and test scores within the class (students), there should be no systematic advantage for schools with higher baseline scores. We clearly find there is, both for teacher and student incentives in the second year. Maybe well performing students and teachers feel generally more confident to compete, and respond more to the treatment, even though they do not have a higher chance to win. Overall, this is a nice study, which has some data problems, and puzzling results.

Appendix Table 34: Impact results in Filmer et al. (2020)

Year 1	ATE	Quant 10	Quant 0.25	Quant 0.5	Quant 0.75	Quant 0.9
Teacher incentives						
Main effects (ATE)	0.002	-0.045	-0.038	0.000	0.018	0.051
	(0.068	(0.029)	(0.027)	(0.027)	(0.034)	(0.041)

Top 30 percentile school (interaction term)	0.059 (0.143)	-0.020 (0.060)	-0.044 (0.055	0.080 (0.057)	0.143** 0.066	0.124 (0.087)
Teacher and student incentives Main effects (ATE)	0.155** (0.068)	0.100*** (0.029)	0.114*** (0.027)	0.140*** (0.027)	0.171*** (0.033)	0.208*** (0.041)
Top 30 percentile school (interaction term)	0.203 (0.144)	-0.041 (0.059)	0.062 (0.054)	0.225*** (0.056)	0.403*** (0.065)	0.321*** 0.086)

Notes: the source is Table 3,5 using Average test scores of Kiswahili language, English language and math.

Appendix Table 35: Year 2 Impact results in Filmer et al. (2020)

Year 2	Year 2 ATE	Quant 10	Quant 0.25	Quant 0.5	Quant 0.75	Quant 0.9
Teacher incentives Main effects (ATE)	0.127** 0.057)	0.020 (0.019)	0.066*** (0.015)	0.108*** (0.018)	0.121*** (0.021)	0.168*** (0.031)
Top 30 percentile school (interaction term)	0.345*** (0.119)	0.21*** (0.039)	0.27*** (0.032)	0.36*** (0.034)	0.36*** (0.044)	0.42*** (0.063)
No Incentives (withdrawn) Main effects (ATE)	0.105 (0.067)	0.032 (0.022)	0.087*** (0.019)	0.112*** (0.021)	0.096*** (0.026)	0.056 (0.037)
Top 30 percentile school (interaction term)	0.294* (0.157)	0.23*** (0.047)	0.24*** (0.039)	0.31*** (0.042)	0.33*** (0.054)	0.31*** (0.078)
Student incentives Main effects (ATE)	0.057 (0.620)	0.009 (0.023)	0.037* (0.019)	0.057*** (0.022)	0.036 (0.026)	0.009 (0.038)
Top 30 percentile school (interaction term)	0.221 (0.151)	0.22*** (0.049)	0.22*** (0.041)	0.25*** (0.044)	0.21*** (0.056)	0.22*** (0.081)

Study 18: Barrera-Osorio, F., Cilliers, J., Cloutier, M. H., & Filmer, D. (2021). Heterogenous Teacher Effects of Two Incentive Schemes: Evidence from a Low-Income Country. (Policy Research Working Paper 9652). https://openknowledge.worldbank.org/handle/10986/35565

Program and evaluation

This paper reports on a large experiment implemented in public primary schools of Guinea, West Africa. The study included 2 treatment groups, and a control group, each of 140 schools, and spanned 2 school years. It focused on grade teachers in grade 3 and 4. In the first treatment arm, well performing teachers could receive an in-kind reward at the end of the year. In the second treatment arm, well performing teachers could receive a certificate and recognition at public meetings.

It was an absolute award scheme, where gains in student test scores made up 70 percent. of the performance measure, and an assessment of teaching quality during an inspection visit the other 30 percent. In the second year, the weight for test score gains was reduced to 60 percent, and a new indicator for student attendance during test taking was added with a weight of 10 percent.

For both treatments, there are 4 award levels, with 4 different performance and award levels. For the in-kind award, the value of the award increases from 4 percent to 49 percent of yearly teacher pay. On average, the in-kind program provided 17 percent of teachers pay in rewards (calculated form table A2). 36 percent of teachers received no award, 25 percent received the highest reward level.

The table below shows the main results. The in-kind rewards outperform the recognition rewards, with the effects of both programs fading in the second year.

Appendix Table 36: Impact results in Barrera-Osorio et al (2021)

	ATE	Interaction term Boys
In kind – Year 1	0.239*** (0.084)	-0.072 (0.060)
In kind – Year 2	0.156 (0.099)	Na
Recognition Year 1	0.125 (0.087)	-0.124** (0.062)
Recognition Year 2	0.088 (0.103)	

Notes: the source is Table 1 (combining grades and subjects)

No heterogenous treatment effects are reported by wealth, or baseline test score.

The paper shows some interesting gender effects, indicating that female teachers are motivated by recognition whereas male teachers are only rewarded by in kind rewards. We also see that the recognition had a substantially larger effect on learning for girls as compared to boys. So it seems like this incentive appeals more to female teachers, and also has a greater impact on female students.

Study 19: Behrman, J.R., Parker, S.W., Todd, P.E., & Wolpin, K.I. (2015). Aligning learning incentives of students and teachers: results from a social experiment in Mexican high schools. Journal of Political Economy, 123 (2), 325-364. https://doi.org/10.1086/675910

Program and evaluation

This paper reports on a social experiment in Mexican public highschools. Three treatment arms were included. In treatment 1, students received payments for test scores, in treatment 2 teachers received payment for test scores of students in their class, and in the treatment 3 both students, teachers and school administrators received payment for their joint performance. The intervention focused on mathematics in grade 10 to 12. 88 schools participated in the experiment, with 20 in each treatment

group and 28 in the control group. The experiment started in school year 2008-2009 and lasted for 3 years.

The incentive payments for students were based on the level at which they started their school year, and at which level they ended. Four different levels are distinguished: pre-basic, basic, proficient and advanced. Students who dropped a level, or stayed in the pre-basis or basic levels received nothing. Students which improved, or stayed at the higher levels received money. The amount of money increased with a higher level. For example, staying proficient yielded 6000 pesos, whereas staying advanced yielded 10,500 pesos (copied from table 3).

In treatment 2, teachers received 5 percent of what their students would have received in the student level treatment. If a student fell one of more levels, a negative incentive of 125 pesos was given. The total incentive could never be negative however.

Treatment 3 is complicated. It included the treatment 1 and 2, but with a bit lower payments. Payments are not only based on individual achievement, but also on achievement of others, that is, students are rewarded for performance of classmates, and teachers are rewarded based on fellow mathematic teachers. There are also payments to administrators and non-mathematics teachers included in this experiment.

The costs of the experiment were just over 2080 pesos per student in treatment 1 (text with table 10). As the average class is 36, this is 36*2080=74880 for a class. An average teacher receives 6332 pesos for the students in their class in treatment 2. For treatment 3, students received 2991+1108=4099 (10^{th} grade) and teacher received 15330+3779=19109 pesos, which in total comes down to 36*4099+19109=166673 per class. Teachers often teach multiple classes, so can earn more than the amounts stated. Annual teacher salary is about 200,000 persos (calculated by 25000/0.125 p 338), so the treatments costs (74,880/200,000)=37%, 6332/200,000=3% and 166,673/200,000=83% of annual teacher salary per class per year, a huge difference in costs!

Results

The authors noted considerable cheating. There was a concern that the treatment results would be affected by cheating, as also the stakes, and thus the incentive to cheat, varies across treatment. For this reason, the authors also present an adjusted set of results, which correct for cheating. The correction is based on suspicious answer patterns observed of students in the same class. We present the adjusted results.

Different cohorts were exposed to the treatment for different times. We present the results for the cohort that was in grade 10 in the first year of the experiment. This cohort experienced the treatment for the full three years. This is also the cohort that the authors use to present heterogenous treatment effects in table 8.

The authors standardize their test scores to have a std dev of 100, rather than 1, as is standard in the literature. We have adjusted the estimates to comply with the standard. The heterogenous treatment results were presented in the paper by gender and baseline achievement level.

Appendix Table 37: Impacts of Student, teacher, and student and teacher incentives combined in Behrman et al. (2015)

Treatment	ATE	Girls	Boys	Pre-basic	Basic	Proficient
Student incentives (T1)						
Year 1	0.169***	0.187***	0.15**	0.15***	0.182***	0.28**
	(0.049)	(0.0565)	(0.0591)	(0.0407)	(0.0592)	(0.125)
Year 2	0.297***	0.338***	0.253***	0.244***	0.353***	0.473***
	(0.0489)	(0.0562)	(0.0579)	(0.0359)	(0.0595)	(0.135)
Year 3	0.227**	0.288***	0.147*	0.236***	0.225**	0.456**
	(0.0749)	(0.0785)	(0.0784)	(0.0628)	(0.0887)	(0.161)
Teacher incentives (T2)						
Year 1	0.0127	0.0151	0.0132	0.0195	-0.017	0.0119
	(0.0574)	(0.0639	(0.0642)	(0.0449)	(0.0743)	(0.161)
Year 2	0.0211	0.0471	0.0232	0.0211	-0.0015	-0.0212
	(0.0605)	(0.064)	(0.0664)	(0.0479)	(0.0732)	(0.161)
Year 3	0.0399	0.0672	0.211	0.0475	0.0272	0.179
	(0.0754)	(0.0757)	(0.0903)	(0.0632)	(0.0876)	(0.177)
Student and teacher incentives (T1+T2)						
Year 1	0.314***	0.358***	0.33***	0.268***	0.308***	0.453***
	(0.0579)	(0.053)	(0.0748)	(0.0484)	(0.0771)	(0.175)
Year 2	0.437***	0.51***	0.455***	0.334***	0.489***	0.581***
	(0.0833)	(0.0743	(0.0998)	(0.0598)	(0.0954)	(0.198)
Year 3	0.567***	0.639***	0.637***	0.507***	0.574***	0.702***
	(0.151)	(0.158)	(0.149)	(0.127)	(0.166)	(0.237)

Notes: the source is table 7 and 8 from the original paper. Estimates have been converted to express standard deviation increase in test scores.

It is clear that the treatments which included student incentives perform best, even when correcting for cheating (T1 and T1+T2). The teacher incentive (T2) alone was not successful. The high value of the student incentives may explain this.

For the teacher incentive treatment (T2), none of the subgroup estimates is significantly different from zero, and neither are they significantly different from each other.

For the treatments that included a student incentive, girls generally benefit more from the intervention than boys, but none of the differences are significantly different from zero (based on a two-sample t-test). By baseline proficiency level, the point estimates show a clear pattern indicating that better off students benefitted considerably more. Often the point estimates for students at the proficient level

were double that of students who were at the pre-basic level at baseline. However, none of the differences are significantly different from each other. The fact that higher payments were provided for higher proficiency levels may have motivated students who were close to that level more.

Student Incentive summaries

Study 20: Blimpo, M. P. (2014). Team incentives for education in developing countries: A randomized field experiment in Benin. American Economic Journal: Applied Economics, 6 (4), 90-109. DOI: 10.1257/app.6.4.90

Program and evaluation
Blimpo (2014) studies a student level incentives experiment in 100 secondary schools in Benin. The experiment has four arms: (1) Individual Target, a standard incentive design where students receive monetary rewards individually for reaching a performance target, 22 schools; (2) Team Target, where teams of four students received rewards based on the performance of the team as a whole, 22 schools; and (3) Team Tournament, where teams of four students competed across schools for three substantial monetary prizes (in a competition among 84 teams), 28 schools; (4) a control arm, 28 schools.

The expected individual pay-offs are equalized across treatment arms and are about USD 6.

The incentive design has multiple thresholds, with clear absolute pay-offs that are sharply increasing in exam performance: for example, in the Individual Target group, each participant was promised 5,000 Francs CFA (≈ $10) if her average score was between 10/20 and 12/20 (exclusive); and 20,000 francs CFA (≈ $40) if they scored at least 12/20. Scores on the national examination, administered at the end of the school year, were used as the basis for the incentive payments.

Results
The results of the experiment show substantial mean treatment effects on students' learning outcomes, across the three designs. Blimpo reports an average treatment effect of 0.29 SD on the overall test score for the Individual Target group; 0.27 SD in Team Target; and 0.34 SD in Team Tournament. The impact is statistically the same across interventions, indicating the first-order mean effect is present irrespective of the specific incentive design.

There is no difference in treatment effects between boys and girls. There is some evidence of a positive interaction effect with the baseline score (continuous variable) in the team tournament treatment. There are no baseline score interaction effects in the other (target) treatment arms. (Also, there are no IA effects in a specification with quartile dummies). Overall, this does not provide (strong) evidence for regressive learning effects.

Appendix Table 38: Mean impact and heterogeneity estimates (Blimpo 2014, Tables 6 and 7)

	ATE	Interaction baseline score
Individual target	0.26** (0.11)	-0.03 (0.10)

Team target	0.24* (0.14)	0.21 (0.13)
Team tournament	0.28** (0.13)	0.24** (0.11)

The paper provides some evidence that, controlling for average team score at baseline, teams with a larger score variance at baseline had higher scores on the exam. This suggests that teams were capable of supporting the lower performing team members to do well.

Study 21: Kremer, M., Miguel, E., Thornton, R. (2009). Incentives to Learn. The Review of Economics and Statistics. 91 (3), 437-456. https://doi.org/10.1162/rest.91.3.437

Program and evaluation

The paper evaluates the impact of the Girls Scholarship Program in Kenya, started in 2001. The goal of the program is to incentivize girls in Grades 6-8 to complete school (to not drop out), using a scholarship competition, based on test scores. The scholarship pays for fees and school inputs in Grades 6-8. This intervention is clearly targeted to one of our subgroups of interest. The sample includes 127 schools across two districts.

The design of the incentive program does not pay special consideration to weak performers (and could thus be expected to result in regressive distributional effects). First, the winners were chosen based on their total test score on government exams across five subjects. Second, the incentive has a threshold design: the scholarship is assigned to the highest scoring 15 percent across *all* grade 6 girls in treatment schools.

Not surprisingly, the baseline score is a strong predictor of the (high stakes) exam score that determines the competition ranking. The study also finds that parent education (but not parent wealth) predicts the exam score ranking. As a result, "Schools varied considerably in the number of winners: 56% of program schools (36 of 64 schools) had at least one 2001 winner, and among schools with at least one winner, there was an average of 5.5 winners per school." So, some schools are more likely to have (multiple) winners than others.

Results

The paper (Table 4) finds evidence for positive mean program impacts on academic performance: girls who were eligible for scholarships in program schools had significantly higher test scores than comparison schoolgirls. (However, there is important district heterogeneity: the significant treatment effects are from Busia only, no effect at all in Teso).

The paper also finds some subgroup effects. Table 6 shows a positive and significant subgroup effect for girls in the second quartile. This subgroup effect is consistent with the relatively short "distance" of these girls to the top 15 percent, and so is possibly explained by the threshold incentive design. The authors remark that ".. the largest gains [are] in the second quartile: those students striving for the top 15% winning threshold". However, the authors cannot reject the hypothesis that treatment effects are equal in all quartiles. In fact, the paper also finds positive program effects among girls with

low pretest scores, who were unlikely to win. (And Table 5 shows spill-over effects for boys, who were not eligible for the scholarship, Busia only).

The paper also finds (unexpected) significant test score gains among boys, even though they were ineligible for the scholarship.

The study describes not finding (positive) significant treatment interactions with baseline household wealth variables; but the regression estimates are not tabled. "Interactions of the program indicator with these characteristics are not statistically significant at traditional confidence levels for any characteristic (regressions not shown), implying that test scores did not increase significantly more on average for students from higher-socioeconomic-status households." We code the wealth result estimate as not available.

Appendix Table 39: Mean impacts and heterogeneity estimates (Kremer et al., 2009)

		Test score quartile interaction			
	ATE	Q1	Q2	Q3	Q4
Girls	0.19* (0.11)	0.00 (0.13)	0.23** (0.10)	0.13 (0.09)	0.12 (0.20)
Boys	0.08 (0.13)	-0.11 (0.12)	0.18** (0.09)	0.11 (0.09)	0.18** (0.10)

Notes: Based on Kremer et al. 2009. Mean effects are from full sample estimates, panel A in tables 4 and 5. Interaction effects are from Table 6.

Study 22: Hirshleifer, S.R. (2016). Incentives for effort of outputs? A field experiment to improve student performance. (Working Paper No. 201701), University of California at Riverside, Department of Economics.

Program and evaluation

Hirshleifer (2016) evaluates a small-scale experiment in 4th-6th grade in 18 schools (45 classrooms) in Mumbai and Pune, India. The program takes place in a setting where a technology-based platform (KA Lite) is integrated in Math instruction. There are two treatments: an input and an output incentive. In the input incentive treatment, students are incentivised for correct answers during the process of completing modules throughout a unit, while the output incentive rewards correct answers during a test at the end of a unit.

Students earn credits that are linearly increasing in the number of correct answers. The credits can be used to "buy" tangible rewards (choosing from 47 items including an eraser and a chess set) from the online platform.

The idea of the input incentive intervention is to overcome present-bias: for learning, continuous effort is necessary, but self-control problems may further limit effort. In addition, students may not know their learning production function well. Treatment was assigned randomly and announced at the start of a unit, making sure the students understood their assignment. The main outcome variable is the score on a (non-incentivized) math test on the KA Lite platform.

Results

The findings of the study are summarized in the table below. The input treatment shows a strongly significant and large mean effect, the output incentive does not. The author explains that the results are driven by the more frequent and salient payments in the input treatment. The study does not find evidence of heterogeneity with respect student characteristics (baseline test score, grade and gender). We copy the main effects and heterogeneity effects for baseline test score and gender from Tables 6 and 8 in the paper.

Appendix Table 40: Mean impacts and heterogeneity estimates (Hirshleifer, 2016)

Treatment	ATE	Test score	Female
Input incentive	0.577*** (0.141)	0.063 (0.073)	-0.030 (0.139)
Output incentive	0.242 (0.170)	-0.033 (0.088)	0.055 (0.319)

Notes: based on Tables 6 and 8 from the original paper

Study 23: Berry, J. (2015). Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India. Journal of Human Resources, 50 (4), 1051-1080. http://jhr.uwpress.org/content/50/4/1051.abstract

Program and evaluation

Berry (2015) provides evidence from a small-scale experiment among students in eight primary schools in urban slums in Gurgaon, India. The experiment provides monetary and in-kind incentives linked to individualized competency targets based on a pre-test. Rewards were based on the scores on a post-test, taken two months after the announcement, that measured progress relative to the target.

Children were assigned individual targets. The reward for reaching the target was 100 rupees (USD 2.50) or about one local daily wage for an unskilled worker. Children were also invited to attend free after school classes.

The experimental design intended to estimate the effects of treatments that differed in recipient type and incentive format. The results presented in the paper focus on the following treatments: a) Parent money; b) Child money; c) Child toy; d) Child voucher for toy.

The paper uses a quasi-experimental control group: "The group of children included in the randomization but not reached for the program announcement serves as a quasi-experimental control group. These children had remarkably similar pre-test scores compared with those reached for the announcement". Based on this comparison group, the paper finds a large effect of being reached by program: 27 percent more likely to achieve target and 0.61 SD higher test score.

The paper further focuses on a comparison of effects between these randomized treatments. Most results, including the subgroup results, are relative to "parent money treatment". In particular, it compares baseline test score subgroup effects between an arm that provided in-kind rewards (toys or vouchers for toys) and cash rewards.

Results

The main treatment effects are presented in Table 4 focus on a comparison of treatments b, c and d with the omitted category a). We use the results in panel B "Money versus Toy"; these are consistent with the individual treatment effects.

The author finds no evidence that the type of incentive or the identity of the recipient affected mean outcomes. However, there is some evidence that these impacts varied by the child's initial test score (relative to the cash treatment). In particular, the estimates show that the in-kind rewards (toys) had a larger effect for children with lower pre-test scores.

Appendix Table 41: Mean impact and heterogeneity estimate (Berry, 2015)

Treatment	ATE	Test score (col 3, cat score)
Toy / Voucher	0.008 (0.053)	-0.084* (0.042)

Notes: Based on Tables 4 and 5 from the original paper

Appendix C: Estimates

Appendix Table 42: Heterogeneity estimates

Reference	Level	Intervention	Year	Outcome	Main Effect, in SD units, (S.E.)	Category (-2, -1, 0, 1, 2)			Weight
						Test score	Girls	Wealth	
Lautharte et al. (2021)	Groups	tax incentive	2009	Math G5 (graph)	0.036 (0.40)	1	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive	2009	Portuguese G5 (graph)	0.061* (0.037	1	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive	2009-11	Math G9 (graph)	0.155*** (0.026)	1	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive	2009-11	Portuguese G9 (graph)	0.153*** (0.023)	0	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive + TA	2011	Math G5 (graph)	0.220*** (0.029)	1	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive + TA	2011	Portuguese G5 (graph)	0.162*** (0.021)	1	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive + TA	2015-17	Math G9 (graph)	0.148*** (0.032)	1	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive + TA	2015-17	Portuguese G9 (graph)	0.118*** (0.028)	1	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive, revised	2013-17	Math G5 (graph)	0.170*** (0.038)	1	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive, revised	2013-17	Portuguese G5 (graph)	0.166*** (0.032)	0	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive, revised	2013	Math G9 (graph)	0.092** (0.035)	0	NA	NA	0.5
Lautharte et al. (2021)	Groups	tax incentive, revised	2013	Portuguese G9 (graph)	0.096*** (0.028)	-1	NA	NA	0.5
Al-Samarrai et al. (2018)	Groups	2015, primary	2015	Exam scores	-0.26 (0.25)	-2	NA	NA	1
Al-Samarrai et al. (2018)	Groups	2016, primary	2016	Exam scores	-1.25*** (0.30)	-2	NA	NA	1
Al-Samarrai et al. (2018)	Groups	2017, primary	2017	Exam scores	0.07 (0.29)	-2	NA	NA	1
Al-Samarrai et al. (2018)	Groups	2015, secondary	2015	Exam scores	2.61*** (0.18)	-1	NA	NA	1
Al-Samarrai et al. (2018)	Groups	2016, secondary	2016	Exam scores	4.55*** (0.38)	2	NA	NA	1

Al-Samarrai et al. (2018)	Groups	2017, secondary	2017	Exam scores	4.34*** (0.46)	2	NA	NA	1
Contreras and Rau (2012)	Groups	National teacher performance incentive program	Panel estimate, table 7	Average score	0.236*** (0.036)	1	NA	NA	1
Ferraz and Pereira (2016)	Groups	Pernambuco teacher incentive	2008-11	Math score	0.134* (0.069)	1	NA	NA	0.5
Ferraz and Pereira (2016)	Groups	Pernambuco teacher incentive	2008-11	Language score	0.119* (0.064)	1	NA	NA	0.5
Glewwe et al. (2010)	Groups	Teacher incentives	Year 1	Reward formula	0.131* (0.079)	0	-1	NA	1
Glewwe et al. (2010)	Groups	Teacher incentives	Year 2	Reward formula	0.215*** (0.075)	0	-1	NA	1
Glewwe et al. (2010)	Groups	Teacher incentives	Year 3	Reward formula	0.026 (0.060)	0	-1	NA	1
Glewwe et al. (2010)	Groups	Teacher incentives	Year 1	NGO exam scores	0.092 (0.086)	0	-1	NA	1
Glewwe et al. (2010)	Groups	Teacher incentives	Year 2	NGO exam scores	0.026 (0.095)	0	-1	NA	1
Barrera-Osorio and Raju (2017)	Groups	teacher group incentive (pooled)	Year 1	Student exam scores	0.0151 (0.07)	0	NA	NA	1
Berhman et al. (2015)	Groups	Student and teacher incentives combined	Year 1	Test score	0.314*** (0.0579)	1	1	NA	1
Berhman et al. (2015)	Groups	Student and teacher incentives combined	Year 2	Test score	0.437*** (0.0833)	1	1	NA	1
Berhman et al. (2015)	Groups	Student and teacher incentives combined	Year	Test score	0.567*** (0.151)	1	1	NA	1
Filmer et al. (2020)	Students	Teacher and for students: in kind prize for best performers in terms of gain in test score,	Year 1	Test score	0.155** (0.068)	1	NA	NA	1

		competition within class							
Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Groups	pay for average increase in test scores in school (or difference between avg test score and expected avg test score from year 3)	Year 1	Test score	0.106** (0.044)	1	-1	-2	1
Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Groups	pay for average increase in test scores in school (or difference between avg test score and expected avg test score from year 3)	Year 2	Test score	0.061 (0.049)	-1	-1	-1	1
Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Groups	pay for average increase in test scores in school (or difference between avg test score and expected avg test score from year 3)	Year 3	Test score	0.089* (0.051)	1	-1	-2	1
Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Groups	pay for average increase in test scores in school (or difference between avg test score and expected avg test score from year 3)	Year 4	Test score	0.123* (0.067)	-1	1	-1	1
Muralidharan (2012) and Muralidharan and	Groups	pay for average increase in test scores in school (or difference	Year 5	Test score	0.129 (0.085)	NA	1	-1	1

Sundararaman (2011)		between avg test score and expected avg test score from year 3)							
Berry (2015)	Students	Toy rewards	2007	Reached literacy goal	0.008 (0.053)	-2	NA	NA	1
Blimpo (2014)	Students	Individual target	2009	Standardized BEPC score	0.26** (0.11)	-1	0	NA	1
Blimpo (2014)	Students	Team target	2009	Standardized BEPC score	0.24* (0.14)	1	0	NA	1
Blimpo (2014)	Students	Team tournament	2009	Standardized BEPC score	0.28** (0.13)	2	0	NA	1
Kremer et al. (2009)	Students	Girls Scholarship competition	2001-02	Test score program test	0.19* (0.11)	0	NA	NA	1
Kremer et al. (2009)	Students	Girls Scholarship competition	2001-02	Test score program test	0.08 (0.13)	0	NA	NA	1
Hirshleifer. (2016)	Students	Input incentive	2014-15	Stand outcome test score	0.577*** (0.141)	1	-1	NA	1
Hirshleifer. (2016)	Students	Output incentive	2014-15	Stand outcome test score	0.242 (0.170)	-1	1	NA	1
Berhman et al. (2015)	Students	Student incentives (pay for increase in levels)	Year 1	Test score	0.169*** (0.049)	1	1	NA	1
Berhman et al. (2015)	Students	Student incentives (pay for increase in levels)	Year 2	Test score	0.297*** (0.0489)	1	1	NA	1
Berhman et al. (2015)	Students	Student incentives (pay for increase in levels)	Year 3	Test score	0.227** (0.0749)	1	1	NA	1
Filmer et al. (2020)	Students	in kind prize for best performers in terms of gain in test score, competition within class	Year 2	Test score	0.057 (0.620)	1	NA	NA	1

Andrabi and Brown (2020)	Teachers	pay for percentile	Year 1	test (high stakes)	0.0918**(0.0575)	1	NA	NA	1
Andrabi and Brown (2020)	Teachers	pay for rating by manager	Year 1	test (low stakes)	0.0859*(0.0220)	-1	NA	NA	1
Chang (2020)	Teachers	2P4Pc with high reward for low ability students	Year 2	test (low stakes)	0.10* (0.06)	-1	NA	NA	1
Duflo et al. (2012)	Teachers	pay for presence	Year 1	Test (low stakes)	0.17* (0.09)	1	0		1
Gaduh et al. (2020)	Teachers	Social accountabiliyt mechanism(SAM)	Year 1	Test (low stakes)	0.084** (0.044)	1	-1	NA	1
Gaduh et al. (2020)	Teachers	Social accountabiliyt mechanism(SAM)	Year 2 follow up	Test (low stakes)	0.028 (0.032)	1	-1	NA	1
Gaduh et al. (2020)	Teachers	SAM+pay for presence	Year 1	Test (low stakes)	0.198*** (0.036)	1	-1	NA	1
Gaduh et al. (2020)	Teachers	SAM+pay for presence	Year 2 follow up	Test (low stakes)	0.133*** (0.034)	1	-1	NA	1
Gaduh et al. (2020)	Teachers	SAM+pay for performance rating	Year 1	Test (low stakes)	0.110***(0.033)	-1	-1	NA	1
Gilligan et al. (2018)	Teachers	pay for percentile	Year 1	Present round 2	0.018 (0.017)	NA	0	-1	0
Gilligan et al. (2018)	Teachers	pay for percentile	Year 1	Test score round 2	0.018 (0.030)	1	1	1	1
Gilligan et al. (2018)	Teachers	pay for percentile	Year 1	Test score round 2 easy items	0.003 (0.028)	1	1	1	0
Gilligan et al. (2018)	Teachers	pay for percentile	Year 1	Test score round 2 diff items	0.042 (0.035)	1	1	2	0
Gilligan et al. (2018)	Teachers	pay for percentile	Year 2	Present round 3	0.042** (0.018)	NA	0	2	0
Gilligan et al. (2018)	Teachers	pay for percentile	Year 2	Took PLE (national exam)	0.023 (0.019)	NA	-1	1	0
Gilligan et al. (2018)	Teachers	pay for percentile	Year 2	Passed PLE (national exam)	0.010 (0.018)	NA	1	0	0

Loyalka et al. (2019)	Teachers	Payment proportional to endline test score	Year 1	Test scores (high stakes)	0.084 (0.052)	-1	-1	-1	1
Loyalka et al. (2019)	Teachers	Proportional to gain in test score	Year 1	Test scores (high stakes)	0.001 (0.050)	-1	-1	-1	1
Loyalka et al. (2019)	Teachers	Pay for Percentile	Year 1	Test scores (high stakes)	0.148** (0.064)	-1	-1	-1	1
Loyalka et al. (2019)	Teachers	Low Reward	Year 1	Test scores (high stakes)	0.081 (0.055)	NA	NA	NA	1
Loyalka et al. (2019)	Teachers	High reward	Year 1	Test scores (high stakes)	0.067 (0.046)	NA	NA	NA	1
Mbiti et al. (2019a)	Teachers	Incentives	Year 1	Low-stakes test (all subjects)	0.06* (0.04)	-1	2	NA	1
Mbiti et al. (2019a)	Teachers	Grant + incentives	Year 1	Low-stakes test (all subjects)	0.12*** (0.04)	2	2	NA	1
Mbiti et al. (2019a)	Teachers	Incentives	Year 2	Low-stakes test (all subjects)	0.03 (0.04)	-1	2	NA	1
Mbiti et al. (2019a)	Teachers	Grant + incentives	Year 2	Low-stakes test (all subjects)	0.23***(0.04)	2	2	NA	1
Mbiti et al. (2019b)	Teachers	Levels	Year 1	Low-stakes test (all subjects)	0.057 (0.052)	1	0		1
Mbiti et al. (2019b)	Teachers	pay for percentile	Year 1	Low-stakes test (all subjects)	-0.029 (0.039)	2	-1		1
Mbiti et al. (2019b)	Teachers	Levels	Year 2	Low-stakes test (all subjects)	0.096** (0.046)	0	0		1
Mbiti et al. (2019b)	Teachers	pay for percentile	Year 2	Low-stakes test (all subjects)	0.044 (0.044)	0	-1		1

Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Teachers	pay for average increase in test scores in class (or difference between avg test score and expected avg test score from year 3)	Year 1	Test score	0.154 *** (0.045)	1	-1	-1	1
Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Teachers	pay for average increase in test scores (or difference between avg test score and expected avg test score from year 3)	Year 2	Test score	0.204*** (0.050)	1	1	-1	1
Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Teachers	pay for average increase in test scores (or difference between avg test score and expected avg test score from year 3)	Year 3	Test score	0.191*** (0.056)	1	-1	-1	1
Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Teachers	pay for average increase in test scores (or difference between avg test score and expected avg test score from year 3)	Year 4	Test score	0.331 ***(0.072)	1	-1	-1	1
Muralidharan (2012) and Muralidharan and Sundararaman (2011)	Teachers	pay for average increase in test scores (or difference between avg test score and	Year 5	Test score	0.444*** (0.101)	NA	-1	1	1

		expected avg test score from year 3)							
Berhman et al. (2015)	Teachers	Treacher incentives (pay for average change in levels of classroom)	Year 1	Test score	0.0127 (0.0574)	0	1	NA	1
Berhman et al. (2015)	Teachers	Treacher incentives (pay for average change in levels of classroom)	Year 2	Test score	0.0211 (0.0605)	-1	1	NA	1
Berhman et al. (2015)	Teachers	Treacher incentives (pay for average change in levels of classroom)	Year 3	Test score	0.0399 (0.0754)	0	1	NA	1
Filmer et al. (2020)	Teachers	Teachers: in kind prize for best performers in terms perc gain avg test score, competion between similar schools	Year 1	Test score	0.002 (0.068)	1	NA	NA	1
Filmer et al. (2020)	Teachers	Teachers: in kind prize for best performers in terms perc gain avg test score, competion between similar schools	Year 2	Test score	0.127** (0.057)	2	NA	NA	1
Barrera-Osorio et al. (2021)	Teachers	in kind reward, 4 levels	Year 1	Test score	0.239*** (0.084)	NA	1	NA	1
Barrera-Osorio et al. (2021)	Teachers	in kind reward, 4 levels	Year 2	Test score	0.156 (0.099)	NA	NA	NA	1
Barrera-Osorio et al. (2021)	Teachers	recognition reward, 4 levels	Year 1	Test score	0.125 (0.087)	NA	2	NA	1
Barrera-Osorio et al. (2021)	Teachers	recognition reward, 4 levels	Year 2	Test score	0.088 (0.103)	NA	NA	NA	1

Filmer et al. (2020)	Teachers	nothing (teacher incentive removed)	Year 2	Test score	0.105 (0.067)	2	NA	NA	1