Lessons on Teacher Pay

Policy Brief on Performance-Based Pay for Teachers*

Hemanth Bharatha Chakravarthy†‡

September 10, 2020

Executive Summary

TODO

Contents

***Who is this policy brief for?** Policymakers, their support staff, civil servants and bureaucrats, development stakeholders, and general readers with an interest in education policy and economic development. The evidence is global and applicable to any developing world context. **Why was it prepared?** To inform deliberations about education policy by summarizing the best available global research evidence about the topic and, consequently, advancing evidence-based policy recommendations.

†A.B. Candidate in Applied Mathematics and Economics, Harvard University, Cambridge MA 02138

‡Mail: hemanthbharathachakravarthy@college.harvard.edu, web: b-hemanth.com

1 Introduction and Motivation

Since independence in India and, particularly, in the past two decades, across the developing world and in India, education priorities have shifted from increasing primary school enrollment to a focus on student learning and attainment. Human capital development, broadly, and learning outcomes in public schools, specifically, have been declared top priorities for the UN Sustainable Development Goals and the 2030 UN Global Human Development Agenda. (Assembly (2015)) Yet, the 2018 Annual Status of Education Report found that only a quarter of Indian third grade students could read and perform subtraction at a second grade level although primary school enrollment rates are over 96%. (ASER (2018))

The Problem: In this context, a series of randomized evaluations[1] have found that simply increasing school resources has little impact on learning outcomes in developing countries. (Glewwe et al. (2009), Blimpo et al. (2015), Das et al. (2013), Pradhan et al. (2011), Sabarwal et al. (2014), De Ree et al. (2017)) Simultaneously, research shows the central role of teachers in determining educational outcomes: individuals exposed to better teachers perform better in school and are more likely to attend college and earn higher salaries. (See Hanushek and Rivkin (2012), Chetty et al. (2014a), Chetty et al. (2014b)) Teacher pay is also a significant part of the global education budget—in India, a 2014 report found that teacher salaries alone accounted for 80% of all non-capital education expenditure. (Dongre et al. (2014)) Despite this, Kremer et al. (2005) find that 25% of Indian public school teachers are absent on any given day and less than half of them actually teach even if they attend class.

Furthermore, Kremer et al. (2005) show that in Indian government schools, teachers reporting high levels of job satisfaction are more likely to be absent. Based on further discussions with teacher focus groups, they suggest that this was because teachers who put in low effort and still get by were quite satisfied, while hard-working teachers were dissatisfied because there was no difference in the career/professional outcomes between them and those who shirked work. Moreover, Mullainathan (2005)'s psychological analysis describes how the high initial intrinsic motivation of teachers can diminish over time if they feel that the government does not appreciate or reciprocate their efforts. This creates great potential to modify teaching by tinkering with pay.

There are many commonly explored supply-side solutions to improve learning outcomes in economic development research. Some examples include hiring supplementary or extra contract teachers, hiring part-time staff to focus on math and language, increasing funding or providing books, computer-assisted learning and other ed-tech innovations, coaching teachers, and mixing poor and rich students (as the Indian *Right to Education Act* does). However, teacher pay is still an important consideration: it affects the incentives of teachers to attend class, plan lessons, assign more homework, teach more advanced materials, help slow-learners or otherwise put in more effort into teaching (which is demonstrably crucial in affecting student performance) and, potentially, changes who is attracted to become teachers in the first place.

If teacher pay does indeed incentivize teacher effort and performance, the current system of teacher pay distorts incentives that would otherwise allow the best, most hard-working teachers to earn the most and motivate all teachers to contribute more effort. Instead, right now, teacher pay rewards experience and additional qualifications like masters degrees. However, research shows that **teacher experience and additional qualifications are a poor predictor of their students' performance**. (Rockoff (2004), Rivkin et al. (2005)) Thus, there have been attempts globally—by governments in several US states, Australia, Brazil, Chile, Peru, Israel, and the United Kingdom—to tie teacher pay to student outcomes. The overarching idea in these attempts—i.e., in performance-based pay for teachers—is simple: find some measure of student performance and financially reward better performing students' teachers.

This issue is particularly important in developing nations like India. Evidence suggests that changing teacher incentives tends to be more effective in contexts where there is low accountability in the education system. (Imberman (2015), Ganimian and Murnane (2016), Glewwe and Muralidharan (2016)) Particularly in contexts like India where curriculum and teaching tend to target above median students and leave behind students at the lower ends of the distribution (see Glewwe et al. (2009)), researchers learned that success [in getting teachers to "teach at the right level"] required getting instructors to embrace the mission of targeting the learning needs of students who have fallen behind. (Banerjee et al. (2017))

The economic theory insights on performance-based teacher pay are mixed: while theory expects responses to pay incentives, these responses could very much be perverse ones like teaching from a heavily test-focused perspective and encouraging rote learning at the cost of long-term learning and human capital development. This makes the question of teacher pay an empirical one. That is, the best way then, to understand whether performance-based teacher pay works and, in particular, what model of it works best is to look at experimental studies that attempted performance-based pay and natural experiment[2] settings that study instances where a governmental or non-governmental body attempted performance-based pay. That is exactly what this policy brief does, reviewing the evidence on performance-based pay and, consequently, advocating an "ideal" model of performance-based pay for teachers.

2 Favorable Evidence

2.1 Summary of Favorable Evidence

I summarize the favorable evidence on performance-based pay showing improvements in student test scores in incentivized subjects (i.e., those subjects whose outcomes are measured to determine pay rewards), spillover effects, if any, on other non-measured subject performance, and any effects on teacher and school effort in Table 1. The different kinds of models explored in these pieces of evidence like Pay-for-Percentile Tournaments (P4P), Gains, and Thresholds/Levels based incentives are all explained and compared in Section 4.

Other, less reliable or less relevant studies also produce favorable results. For instance, Barrera-Osorio and Raju (2017) conducts a RCT providing cash bonuses to teachers with most student enrollment, final exam participation, and scores in 600 of the worst performing primary schools in Punjab, Pakistan, and observes higher exam participation, especially amidst the worst performing schools within the sample[3]. There have also been studies in developed nations like the US, in different educational contexts, that offer favorable evidence for performance-based pay.

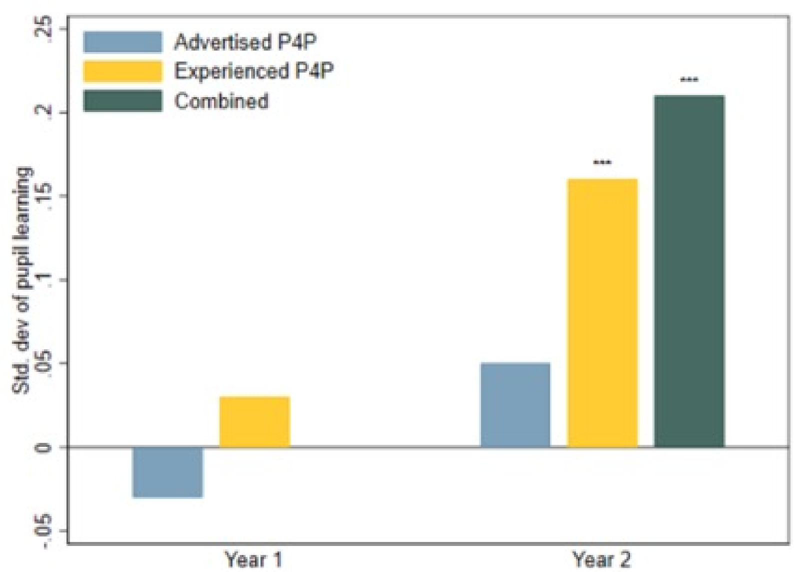

Figure 1: Student Improvements under Pay-for-Percentile in Rwanda (Leaver et al. (2019))

Setting	Learning Outcomes	Spillover Effects	Teacher Effects
Muralidharan and Sundararaman (2011), Muralidharan (2012)			
RCT in 300 government schools in rural Andhra Pradesh, India, with bonus of Rs. 500 for every % over 5% of average class gains in Math and English test scores	After 2 years and 5 years, performance improved by 0.27 and 0.17 SDs and **0.54 SDs** and **0.37 SDs** in math and language tests respectively	After 5 years, spillover improvements of 0.52 SDs and 0.3 SDs in science and social studies	Higher observed and self-reported teacher effort
Leaver et al. (2019)			
Pay-for-percentile (P4P) RCT in Rwandan schools, randomly assigning recruited teachers to different contacts, studying paying mean bonuses of 15% of annual pay to top 20% of P4P teachers based on student performance and teacher presence, preparation, and pedagogy	After 2 years, **0.21 SDs** increase in composite pupil learning	(Incentives across subjects)	Changed the nature of recruitment, attracting more money minded but also more effective teachers; teachers placed in P4P schools were more intrinsically motivated
Mbiti et al. (2019), Mbiti et al. (2019)			
RCT in 350 schools (120K students) in Tanzania, offering different combinations of test performance thresholds incentives and funds	Students in schools with higher funds and PBP were **0.36 SDs** better on high stakes test and **0.23 SDs** on the low stakes test, 49% and 116% more likely to pass Math and English tests, and 25% less likely to repeat a grade	Combination schools had gains in Science scores after incentivizing just Math and English	Teacher presence and effort increases

Table 1: Summary of Favorable Evidence on Performance Based Pay

Table 1: Summary of Favorable Evidence on Performance Based Pay (Continued)

Setting	Learning Outcomes	Spillover Effects	Teacher Effects
Contreras and Rau (2012)			
Natural experiment studying a government P4P program run in 90% of Chilean public schools for Math and English scores	Between **0.14 and 0.25 SDs** for language and math test scores		
Gilligan et al. (2018)			
RCT in 302 Ugandan rural primary schools attempting combinations of giving books and P4P pay incentives	7% increase in two-year student attendance as long as the program was accompanied with providing more books	Dropout rates fall (that are currently high because teachers encourage weaker students to dropout before they reach the PLE exam and hurt school reputation (if they fail))	Increased teacher effort independent of whether the school received books or not; more time spent grading and more overall effort exerted
Loyalka et al. (2019)			
RCT in 216 middle schools in Western China to compare different models of performance-based pay: levels, gains, and P4P	Scores improve by **0.148 SDs** in the P4P program; adding any incentive has a positive increase by 0.074 SDs	(Incentives across subjects)	Curriculum coverage and intensity of teaching improved; teachers taught more advanced concepts

The outcomes discussed in Table 1 represent serious gains: for instance, the program studied by Muralidharan and Sundararaman (2010) had an average treatment effect of 0.22 standard deviations improvement in Math and English. This means, if student grades followed a normal distribution[4], the median student improved scores by 9 percentile points. These scores tangibly matter in the long term: Chetty et al. (2014a) find that the extent of 'gross' value-addition of teachers in grades 4-8 is correlated with the long-term wages of the affected students. Even more directly, Muralidharan (2012) finds that students who completed their full five years of primary school under the program performed significantly better than their peers in control schools by 0.54 and 0.35 standard deviations in math and language tests respectively. These students also scored 0.52 and 0.3 standard deviations higher in science and social studies tests even though there were no incentives on these subjects.

Why did these scores improve? Muralidharan and Sundararaman (2011) and Muralidharan (2012) suggest that the main mechanism for the impact of the incentive program was not increased teacher attendance, but greater and more effective teaching effort conditional on being present. Even comparing between incentive and non-incentive schools within the same Mandals (subdistricts), teacher value addition was comparatively higher in program schools. The interviews they conducted with teachers indicate that teachers in incentive schools are significantly more likely to have assigned more homework and class work, conducted extra classes beyond regular school hours, given practice tests, and paid special attention to weaker children. They also find that teachers supporting the program before it began was correlated to how hard they worked, supporting the idea that incentives work and that in the longer term, there is a possibility that better teachers will be attracted into the system.

Meanwhile, Leaver et al. (2019), who actually study the effects on the composition of teachers, posit that one quarter of the modest but significant impact they observed can be attributed to selection of better teachers at the recruitment stage with the remaining three-quarters arising from increased effort of existing teachers. Hence, they highlight the importance of advertizing the performance-based pay program widely, asserting that that is important in ensuring teachers respond to the program and changing recruitment trends.

2.2 Cost Effectiveness

Dhaliwal et al. (2013) stress the importance of cost effectiveness analysis, particularly in choosing within different education policy alternatives. Their basic model pushes policy makers to identify the key objective for which they are considering policy options and then compare the impact on outcomes per dollar spent. However, for broader context, figure 2 offers background on the costs of various South Asian studies aimed at achieving various goals in education and their outcomes.

Table B – Costs of Rigorously Evaluated Education Interventions in South Asia

Study	Interventions	Costs
Andrabi, Das and Khwaja (2013)	report cards with school and child test scores	USD 1 per child / USD 20 per marginal child enrolled
Banerjee, Cole, Duflo and Linden (2007)	hiring of young women to teach students lagging behind in basic literacy and numeracy skills	USD 2.25 per child per year
Banerjee, Cole, Duflo and Linden (2007)	computer-assisted learning program focusing on math	USD 15.18 per child per year
Barrera-Osorio, Blakeslee, Hoover, Linden, Raju and Ryan (2013)	publicly funded private primary schools	USD 18 per child per year
Chaudhury and Parajuli (2010a)	CCT for girls	USD 36 per child per year
He, Linden and MacLeod (2008)	English education curriculum (different implementations)	USD 11.20-20.46 per child per year
Lakshminarayana, Eble, Bhakta, Frost, Boone, Elbourne and Mann (2013)	supplementary remedial teaching by community volunteer, provision of learning material and additional material support for some girls	USD 48-61 per child per 18 months
Linden (2008)	computer assisted learning program (different implementations)	USD 5 per child per year
Muralidharan and Prakash (2013)	bicycles for girls	USD 12 per child per year
Muralidharan and Sundararaman (2010)	diagnostic tests and feedback to teachers and monitoring of classroom processes	USD 6 per child per 24 months
Muralidharan and Sundararaman (2011) and Muralidharan (2012)	group-based bonus payments to teachers based on improvement in students' test scores	USD 2 per child per year
Muralidharan and Sundararaman (2011) and Muralidharan (2012)	individual-based bonus payments to teachers based on improvement in students' test scores	USD 3 per child per year
Muralidharan and Sundararaman (2013a)	school choice program featuring lottery-based allocation of school vouchers	savings of 102 USD per child per year
Muralidharan and Sundararaman (2013b)	extra contract teachers	USD 6 per child per 24 months

Figure 4 – Meta-Analysis of Interventions' Impacts on Composite Test Scores

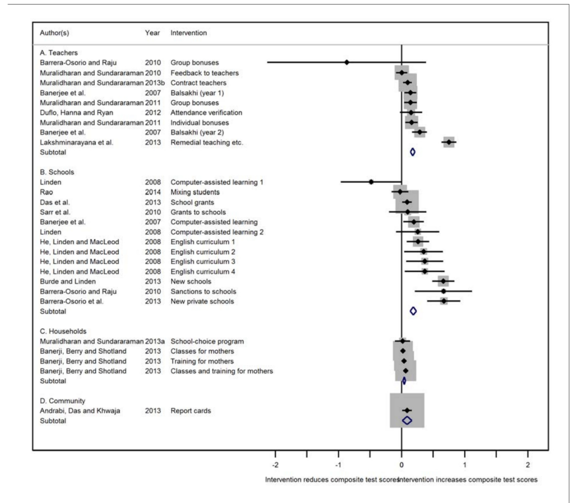

Figure 2: Costs-Benefits of South Asian Education Policy Experiments (Asim et al. (2015))

As per this model of cost effectiveness analysis, Mbiti et al. (2019) report that their Levels treatment (one of two models of performance-based pay they trial) improves performance by 0.22 standard deviations, implying a cost-effectiveness of 3.04 standard deviations per USD 100 spent per child. Such estimates suggest that both the models of performance-based pay that they trial are more cost-effective compared to several other interventions aimed at improving teacher incentives and quality in developing countries analyzed in the overview by Kremer et al. (2013).

Similarly, the average cost of the program studied by Muralidharan and Sundararaman (2011) was USD 3 per student when the program paid, on average, 3% of annual salary as bonuses to teachers after measuring their improvements. They posit that this makes performance-based pay a significantly more cost effective way of increasing student test scores compared to spending a similar amount of money unconditionally on school inputs. Similarly, Gilligan et al. (2018) also support this, finding that performance-based pay costs USD 3 per student.

Further, several studies identify strong complementarities which are relevant for understanding the cost-effectiveness of performance based pay. Muralidharan and Sundararaman (2011) find that impact is maximum when there is a combination of additional funds and performance-based pay. Similarly, Mbiti et al. (2019) find complementarities between increasing school costs and performance-based pay and Gilligan et al. (2018) find that performance-based pay works if schools are given more books. Moreover, Mbiti et al. (2019)

find that, using a composite measure of test scores across subjects, the "interaction" effect between school inputs and teacher incentives was equal to 0.18 standard deviations. They find that the combination treatment had effects over five times greater than the sum of the impact of the Grant (just giving funds) and Incentives treatments after two years. This means that for different forms of education expenditure, for the same amount of money, policy makers can make significantly larger gains if they combine them with incentivizing teachers.

3 Key Concerns and Unfavorable Evidence

3.1 Summary of Negative/Mixed Evidence and Features of Failed Programs

Setting	**Negative Outcomes**	**Positive Outcomes**	**Features**
	Glewwe et al. (2010)		
RCT in 50 Kenyan primary schools (grades 4 to 8) with group in-kind incentives based on test scores	Evidence of **teaching to the test**: after the performance-based pay program ended, students who improved lost their improvements; gains came primarily from teachers trying to improve test scores (rather than focusing on deeper learning)	After 2 years, scores improved by **0.22 SDs** ; exam prep classes and exam participation increased	In-kind gifts given as incentives; group incentives as opposed to incentivizing individual teachers
	Bellés-Obrero and Lombardi (2019)		
Natural experiment studying a nationwide government P4P program in 8th grade in Peru	Program did not cause any significant positive effects		Program not well advertized; group incentives; the researchers used internal grades and not the incentivized standardized test

Table 2: Summary of Negative and Mixed Evidence

There is other less reliable negative evidence as well. Behrman et al. (2015) find in a RCT in 88 Mexican high schools that there weren't any significant positive outcomes to having teacher incentives but there were to having student pay bonuses for performance. But this program is carried out in different types of schools including agro, industrial, and marine ones, was vulnerable to cheating, had small bonuses based on a very complicated system, and applied in a schooling context where every subject has different teachers for different semesters in the same year sharing the bonus if all of them took effort and improved performance. Further, the test they use is a low stakes external one and offers no incentive for students to work hard. However, it is worth noting this to realize the features of failed programs.

3.2 Key Concerns

Neal and Schanzenbach (2010) posit that performance-based pay could create perverse incentives for teachers and encourage them to **cheat** or **"teach to the test"**, i.e., teach from a heavily test focused lens, encouraging rote learning, for instance. Evidence supports this. Glewwe et al. (2010) find that student performance improved more in multiple choice type questions than fill in the blank type ones as "memorization is easier" for the former. High stakes tests also can push schools to push poor performing students to drop out to improve overall performance. ((Gilligan et al. (2018))

Although this concern of perverse incentives has truth to it, it is not sufficient to dismiss performance-based pay. For one, Muralidharan and Sundararaman (2011) explicitly test students on both questions requiring more application and cognitive effort and mechanical questions that favor rote learning. While they do find higher score improvements in the latter category, there is still signficant growth in the former. Furthermore, Glewwe et al. (2010) themselves note, "Certain rents (e.g., high-paying government jobs or rationed places in secondary education) may heavily depend on test scores. Thus, higher test scores can benefit pupils as well as teachers." They still say that if this program only helps students to signal to employers or universities and does not build underlying human capital, it is socially wasteful. However, it is important to note that, particularly in developing world contexts like India, where exam scores are important for students' college applications or jobs, even if deeper human capital does not develop as much as test-taking ability, performance-based pay programs might still be entirely worth it.

There are other concerns around performance-based pay that only selectively apply to some models of implementing it. That is to say, these concerns can be evaded by a better designed performance pay program. Some prominent concerns of this category include that this might encourage teachers to selectively focus on students whose performance is most likely to improve or that performance-based pay punishes disadvantaged teachers who have to teach in schools lacking resources with students facing more disadvantages. Programs like Muralidharan and Sundararaman (2010) only compare a teacher's improvements to the teacher's own students baseline performance (past performance) and include all students in the calculation of score gains. This sidesteps these concerns. Further, Mbiti et al. (2019) find no evidence selective focusing. Ultimately, it is valid to note that teachers' tasks are multidimensional and test scores do not fully reflect their work. However, in developing nations, this is not a fatal concern either as student learning is still the primary objective.

Finally, Gilligan et al. (2018), Mbiti et al. (2019), and Muralidharan and Sundararaman (2011) all note either strong complementarities between funding and performance pay or that performance pay only has significant outcomes if it comes alongside funding. This is attributed to the depravity of schools in the contexts they study. That is, if basic resources and books are not a given, any amount of teacher incentives might still not translate to learning outcomes. This could explain how Loyalka et al. (2019) in China and Gilligan et al. (2018) in Rwanda obtain different outcomes despite applying a similar program of pay-for-percentile on sixth grade students. Thus, it is important to ensure that there is a bare minimum of school infrastructure for performance pay to actually start working.

Independent of the evidence itself, it is also worthwhile to note that evidence and, particularly, randomized controlled trials are not complete proof that performance based pay will work. Particularly, there are concerns around external validity: it is not necessary that the same program when implemented in a different country context by a different government will have consistent outcomes with the evidence. See Banerjee et al. (2017) for suggestions on scaling up government policy based on RCT evidence.

4 An 'Ideal Model' of Performance-Based Pay

Two notes here: first, there are scaling considerations that are not directly related to the intervention, particularly, navigating teacher unions that are often quite strong and budgetary limitations; second, program design is crucial for performance-based pay. A plausible explanation for the mixed evidence is precisely the differences in program design.

4.1 Pay-for-Percentile vs. Gains vs. Levels

The two attempts to compare the different models show mixed results. In Tanzania, levels schools were 0.096 standard deviations higher relative to the Pay for Percentile schools

	P4P	Gains	Levels
Meaning	Tournament that compensates teachers based on the percentile ranks of every student at the year end in the homogeneous comparison set (based on location and socioeconomic status)	Compensation based on the average gains of student scores compared to baseline level	Compensation and rewards based on average achievement: reaching certain fixed student score goals or thresholds
Student Distributional Effects and Triage	Strongest incentives to focus on all students across the distribution (Loyalka et al. (2019))	Incentive to focus on students most capable of improvement (this could be weaker or stronger students)	Incentive to focus on students close to the threshold (Neal and Schanzenbach (2010), Lavy (2009))
Teacher Distributional Effects	Comparison only within comparison set of similar schools	Only improvements with respect to own students' baseline performance measured	Comparison independent of starting point: punishes teachers serving disadvantaged students
School-level Triage	Top pre-program percentile ranked and bottomed ranked schools lose incentive; Contreras and Rau (2012) find impact only on schools already above the 65th %ile		
Clarity or parsimony	More complicated; higher risk of misunderstanding a la Fryer (2013), Goodman and Turner (2013)	Relatively simple; lower risk of being misunderstood	

Table 3: Comparison of Pay Models

(Mbiti et al. (2019)). However, in China, P4P outperformed the levels treatment which in turn outperformed the gains treatment. Particularly, adjusted for effects on students across the distribution of baseline student abilities as opposed to just average or total effects, P4P further outperformed the other two models. Economic theory also favors the P4P model. (Barlevy and Neal (2012))

Regardless of the model chosen, it is important to advertise performance-based pay well. For one, Leaver et al. (2019) find that the positive impact of such programs is higher when well advertised. Moreover, if teachers have to respond to the incentives created by performance-based pay, as a prerequisite, they need to know of their existence, understand the system well, and have an idea of how they can work to improve student human capital.

Similarly, regardless of the model chosen, it can be important to accompany the program with appropriate funding increases or books. As described in Section 3.2, performance-based pay often only works if basic school infrastructure and books are a given, allowing teacher effort to actually have a meaningful impact. This is further strengthened by the discussion of complementarities between funding and performance incentives, as described in Section 2.2.

4.2 Test Design and Grade Choice

To ensure that the test does actually measure the performance of students, the following must be ensured:

1. The test metrics should be consistent with minimal noise: if a student got 80% in year 1 and 90% in year 2, that should mean that the student improved by around 10% and not that the test was easier or scored more liberally in year 2. This was a crucial lesson from Barrera-Osorio and Raju (2017)'s study of a program in Punjab, Pakistan.

2. The content of the test to should require as much cognitive application as possible: a test that is harder to rote learn for or to use score-boosting tactics like those around multiple choice questions for does a better job of measuring actual human capital development. This would help prevent the result Glewwe et al. (2010) find where teachers become overtly test-focused instead of developing student human capital.

3. The test should have low risk of cheating. Avoid allowing self-reporting of scores by teachers and schools and increase external supervision, exchange correction of test papers, surprise checks on test centers, and so forth. Teacher-enabled cheating should be heavily deterred. In nations like India with pre-existing high stakes tests such as the grade 10 and 12 school-leaving exams (the "board exams"), these tests could serve

as the incentive. So, even in the event where teachers become overly test-focused, this is likely to be less different from current approaches to crucial tests and, in the worst case, cost human capital in return for better scores in career-determining exams.

4. Insulate against externalities of high stakes tests such as student stress or pressure from school authorities for worst-performing students to leave the school and dropout.

5. Teachers should understand the test and understand how to teach accordingly. More feedback for teachers on their pedagogy also helps. (Muralidharan and Sundararaman (2010))

6. Finally, it is also potentially useful to track other indicators in considering bonuses and rewards; for instance, dropout rates could also be incentivized. Based on ability and state resources, tracking other performance indicators like teacher attendance and incentivizing these could also help.

4.3 Bonus Design

Certainly, the pay increment must be modeled as a bonus. This ensures that all teachers receive a standard basic minimum pay to prevent resentment or loss of current and potential teachers.

4.3.1 Group vs. Individual Bonuses

The evidence heavily favors individual bonuses targeted at those teachers who performed well over group bonuses to schools that performed well. This is also clearly intuitive: group bonuses require some amount of co-operation and risk the creation of free-riders. Whereas, individual performance is directly under the individual teacher's control and, when incentivized, more likely to inspire more effort.

4.3.2 Bonus Amount

In terms of the bonus amount itself, the key learning seems to be that it doesn't matter too much. The only explicit trial of this, in China, by Loyalka et al. (2019) finds roughly no differences in impact between different bonus sizes (in fact, doubling the bonus had no significant impact). The average bonus in Muralidharan and Sundararaman (2011) was around 35% of a monthly salary, whereas prizes were in-kind in Glewwe et al. (2010), and had an average value ranging between 12% and 21% of a teachers' monthly salary. The Chilean program studied by Contreras and Rau (2012) awarded an average bonus of 10%

of a monthly salary whereas an Israeli program studied by Lavy (2002) awards prizes of 10%-40% of an average teacher's monthly salary to approximately one third of participating teachers. The average bonus studied by Mbiti et al. (2019) was approximately half a month's salary.

4.3.3 Bonus Structure

The bonus structure relies on the model of measuring performance as discussed in Section 4.1. For the sake of an example, however, the bonus model in Muralidharan and Sundararaman (2011) was structured as a bonus of Rs. 500 for every percentage point increase in average scores over 5%, that is:

$$\text{Bonus} = 500 * (\text{\% gains in class average test scores} - 5\%)$$

5 Conclusion

TODO

Notes

[1]A *randomized controlled trial* (or RCT) is a study where subjects are randomly assigned to *treatment groups* where they experience interventions and a *control group* where nothing changes and observed. The idea is that any changes in outcomes in the *treatment group* are *causally tied* to the only difference between the two groups—receiving the treatment or not.

[2]A *natural experiment* is an observational study where a subpopulation is chosen by some natural process that resembles random selection and then selectively exposed to a treatment (policy or intervention). By observing the differences between similar individuals after they either received the treatment or did not, researchers make causal inferences about the effects of the treatment. In this way, it is a natural setting resembling a RCT. Simultaneously, these are generally different from RCTs in that those implementing policies are often actual governments (whose impact is then studied by researchers) as opposed to researchers themselves.

[3]However, as they themselves add, the test they use isn't reliable or consistent and, therefore, test scores are hard to compare due to noise. Their incentive program was also weak.

[4]A *normal distribution*, also known as a bell curve, is a symmetric distribution where most of the observations cluster around the central peak and the probabilities/frequencies for values further away from the mean taper off equally in both directions.

References

ASER (2018). Annual status of education report india 2018. *ASER Centre, New Delhi*.

Asim, S., R. S. Chase, A. Dar, and A. Schmillen (2015). *Improving education outcomes in South Asia: Findings from a decade of impact evaluations*. The World Bank.

Assembly, U. G. (2015). Transforming our world: the 2030 agenda for sustainable development: Draft resolution referred to the united nations summit for the adoption of the post-2015 development agenda by the general assembly at its sixty-ninth session. Technical report, A/70/L. 1, 18 September 2015. http://undocs. org/A/70/L. 1 (accessed 19 . . . .

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives 31*(4), 73–102.

Barlevy, G. and D. Neal (2012). Pay for percentile. *American Economic Review 102*(5), 1805–31.

Barrera-Osorio, F. and D. Raju (2017). Teacher performance pay: Experimental evidence from pakistan. *Journal of Public Economics 148*, 75–91.

Behrman, J. R., S. W. Parker, P. E. Todd, and K. I. Wolpin (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.

Bellés-Obrero, C. and M. Lombardi (2019). Teacher performance pay and student learning: Evidence from a nationwide program in peru.

Blimpo, M. P., D. K. Evans, and N. Lahire (2015). *Parental human capital and effective school management: evidence from The Gambia*. The World Bank.

Brodaty, T. and M. Gurgand (2008). Teacher and peer effects in higher education: Evidence from a french university. Technical report, Working Paper. Paris School of Economics Paper No. Otober 2008-10.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review 104*(9), 2633–79.

Contreras, D. and T. Rau (2012). Tournament incentives for teachers: evidence from a scaled-up intervention in chile. *Economic development and cultural change 61*(1), 219–246.

Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan, and V. Sundararaman (2013). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics 5*(2), 29–57.

De Ree, J., K. Muralidharan, M. Pradhan, and H. Rogers (2017). *Double for nothing? experimental evidence on an unconditional teacher salary increase in indonesia*. The World Bank.

Dhaliwal, I., E. Duflo, R. Glennerster, and C. Tulloch (2013). Comparative cost-effectiveness analysis to inform policy in developing countries: a general framework with applications for education. *Education policy in developing countries*, 285–338.

Dongre, A. A., A. Kapur, and V. Tewary (2014). How much does india spend per student on elementary education? *PAISA District Studies-2013 (Forthcoming)*.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review 101*(5), 1739–74.

Duflo, E. and R. Hanna (2005, December). Monitoring works: Getting teachers to come to school. Working Paper 11880, National Bureau of Economic Research.

Duflo, E., R. Hanna, and S. P. Ryan (2012). Incentives work: Getting teachers to come to school. *American Economic Review 102*(4), 1241–78.

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from new york city public schools. *Journal of Labor Economics 31*(2), 373–407.

Ganimian, A. J. and R. J. Murnane (2016). Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research 86*(3), 719–755.

Gilligan, D. O., N. Karachiwalla, I. Kasirye, A. M. Lucas, and D. Neal (2018). Educator incentives and educational triage in rural primary schools. Technical report, National Bureau of Economic Research.

Glewwe, P., N. Ilias, and M. Kremer (2010). Teacher incentives. *American Economic Journal: Applied Economics 2*(3), 205–27.

Glewwe, P., M. Kremer, and S. Moulin (2009). Many children left behind? textbooks and test scores in kenya. *American Economic Journal: Applied Economics 1*(1), 112–35.

Glewwe, P. and K. Muralidharan (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the Economics of Education*, Volume 5, pp. 653–743. Elsevier.

Goodman, S. F. and L. J. Turner (2013). The design of teacher incentive pay and educational outcomes: Evidence from the new york city bonus program. *Journal of Labor Economics 31*(2), 409–420.

Hanushek, E. A. and S. G. Rivkin (2012). The distribution of teacher quality and implications for policy. *Annu. Rev. Econ. 4*(1), 131–157.

Imberman, S. A. (2015). How effective are financial incentives for teachers? *IZA World of Labor*.

Jacob, B. A. and S. D. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics 118*(3), 843–877.

Kremer, M., C. Brannen, and R. Glennerster (2013). The challenge of education and learning in the developing world. *Science 340*(6130), 297–300.

Kremer, M., N. Chaudhury, F. H. Rogers, K. Muralidharan, and J. Hammer (2005). Teacher absence in india: A snapshot. *Journal of the European Economic Association 3*(2-3), 658–667.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of political Economy 110*(6), 1286–1317.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review 99*(5), 1979–2011.

Leaver, C., O. Ozier, P. Serneels, and A. Zeitlin (2019). Recruitment, effort, and retention effects of performance contracts for civil servants. *World*.

Loyalka, P., S. Sylvia, C. Liu, J. Chu, and Y. Shi (2019). Pay by design: Teacher performance pay design and the distribution of student achievement. *Journal of Labor Economics 37*(3), 621–662.

Majerowicz, S. and R. Montero (2018). Can teaching be taught? experimental evidence from a teacher coaching program in peru. Technical report, Working Paper.

Mbiti, I., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani (2019). Inputs, incentives, and complementarities in education: Experimental evidence from tanzania. *The Quarterly Journal of Economics 134*(3), 1627–1673.

Mbiti, I., M. Romero, and Y. Schipper (2019). Designing effective teacher performance pay programs: Experimental evidence from tanzania. Technical report, National Bureau of Economic Research.

Mullainathan, S. (2005). Development economics through the lens of psychology. In *Annual*

Bank Conference on Development Economics 2005: Lessons of Experience, pp. 45–70.

Muralidharan, K. (2012). *Long-term effects of teacher performance pay: Experimental evidence from India*. ERIC Clearinghouse.

Muralidharan, K. and V. Sundararaman (2010). The impact of diagnostic feedback to teachers on student learning: experimental evidence from india. *The Economic Journal 120*(546), F187–F203.

Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: Experimental evidence from india. *Journal of political Economy 119*(1), 39–77.

Neal, D. and D. W. Schanzenbach (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics 92*(2), 263–283.

Pham, L. D., T. D. Nguyen, and M. G. Springer (2017). Teacher merit pay and student test scores: A meta-analysis. In *Association for Education Finance and Policy Annual Meeting, Washington, DC*.

Pradhan, M., D. Suryadarma, A. Beatty, M. Wong, A. Alishjabana, A. Gaduh, and R. P. Artha (2011). *Improving educational quality through enhancing community participation: Results from a randomized field experiment in Indonesia*. The World Bank.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American economic review 94*(2), 247–252.

Sabarwal, S., D. K. Evans, and A. Marshak (2014). *The permanent input hypothesis: the case of textbooks and (no) student learning in Sierra Leone*. The World Bank.