“How Effective are Financial Incentives for Teachers?”

Scott A. Imberman
Michigan State University and NBER

Keywords

teacher incentives
student achievement
pay-for-performance
merit pay

Teaser

Linking teacher pay to student performance has become popular but the evidence on its effectiveness is mixed.

Elevator Pitch

Concerns about poor student performance have led schools to diverge from traditional teacher compensation and base a portion of pay on student outcomes. The number of locations adopting such incentive pay schemes has increased substantially. Since 2004 the number of school districts in the US providing teachers financial incentives to improve student performance has increased by over 40%. Figure 1 shows that across US states there is widespread variation, with some states having no incentives and in others close to 50% of districts have incentives. Overall, 11% of districts in the US provide teachers with performance incentives.

Pros

• Incentives can effectively improve student performance if they are designed well.
• Paying teachers for student performance has been shown to be highly effective in developing countries, specifically Kenya and India.
• Small group incentives strike a balance between loss of effectiveness from free-riding and gains in effectiveness from teachers cooperating with each other.

Cons

• Overall, evidence on individual incentives in developed countries is mixed with some positive and some negligible impacts.
• Large group incentives have been shown to have little impact on achievement and in some cases even generate negative impacts.
• There is no evidence that incentives tied to specific exams generate improvements in other measures of academic performance, suggesting a lack of general improvements in knowledge.

Author’s Main Message
Financial incentives for teachers can be effective with appropriate design, but poorly designed incentives will have little benefit. Policy makers ought to avoid threshold based incentives when possible in favor of piece-rate systems. Incentives should be aligned with multiple outcomes that include both objective (e.g. teacher value-added) and subjective (e.g. class observations) components. If group incentives are used instead of individual, groups should be kept small (e.g. school-grade-subject instead of whole school).

Motivation

Traditionally teachers in many parts of the world are compensated based on credentials (degrees and certifications) and experience. However, research has shown that the returns to experience are limited while there is little impact of credentials on student performance. However, the quality of the teacher is very important. Given this disconnect between teacher compensation and teacher performance, the idea of financial incentives for teachers (often called “performance,” “merit” or “incentive” pay) aligned with measures of student performance has become increasingly popular. In fact, teacher incentives were one of the key pillars of the US President Barack Obama’s “Race to the Top” initiative.

Financial incentives for teachers are becoming increasingly popular. According to the Schools and Staffing Survey conducted by the US Department of Education, the share of public and charter school districts in the United States with financial incentives for teaching excellence has increased by over 40% from 2004 through 2012. As can be seen in Figure 1, there is also wide variation where in some US states there are no incentives and in others nearly half of districts offer incentives. Incentives have also been implemented in many countries worldwide such as in Denmark, Norway, Hungary, Israel, India and Kenya.

This essay will focus specifically on financial incentives for direct improvements in student performance. A number of other types of incentives including recruitment for hard-to-staff schools, incentives to acquire certain credentials such as advanced certifications and incentives to recruit teachers in fields with shortages, while interesting, will not be considered.

Discussion of Pros and Cons

Key Goals of Incentive Pay

The primary concept underlying teacher incentives is that teachers are paid based on their productivity. The goal of such a scheme is to have two key impacts. The first is to encourage teachers to exert more “effort.” Effort here is defined broadly to include both the quantity and quality of effort. For example, for quantity improvements teachers with incentives may be inclined to increase time spent giving students instruction or provide after-school tutoring. Quality can be enhanced by adopting innovative teaching techniques, analyzing data to improve student performance, or by experimenting with different teaching methods. The second goal is to recruit high quality teachers as some economists have theorized that incentives will attract people to the teaching profession who are better at improving student performance.

Variations in Incentive Pay
In practice there are nuances to the implementation of incentive pay that can have large implications for effectiveness. Thus any policy implementer must think carefully about the most appropriate flavors of incentives to provide.

The first issue is whether to provide incentives based on the performance of teaching groups or individual teachers. Individual awards are provided to teachers based on how well they improve their own students’ performance. Group incentives provide rewards based on average performance of a group of teachers. Most often the group will comprise of all teachers in a school, school-grade, school-subject, or school-subject-grade. In some cases teachers are put into smaller teams of teachers. There are two key economic distinctions between these award types that drive their effectiveness. An important worry is that group incentives promote “free-riding” where teachers do not increase their effort as much as with individual incentives because they can take advantage of effort improvements in other group members. On the other hand, group incentives encourage cooperation amongst teachers whereas individual performance promotes competition. Given that teachers tend to benefit from the help of their colleagues and a collegial school environment is better for productivity, there is worry that individual incentives could damage these relationships.

The second key design characteristic is the metrics used to identify award winners. Typically at least a portion of the incentives are aligned to some test score measure. Since basing incentives on unadjusted test scores tends to reward teachers for having high ability students, rather than how they improve student performance, districts have relied on test score growth to assess teacher performance. More complex statistical models called teacher value-added models are also widely used and purport to identify the direct contribution of teachers to student achievement growth. Nonetheless, rewarding teachers based on specific exam performance encourages teachers to target that exam alone potentially with little impact on broader learning. Thus it is common for districts to base pay on multiple outcomes in addition to test scores such as classroom observations and principal evaluations.

The final design characteristic is the structure of the incentive system. Incentives can be implemented through three methods – absolute targets, rank-order tournaments, and piece-rate. Absolute incentives provide teachers with bonus pay if their students achieve certain outcomes regardless of how other teachers perform. For example, the Advanced Placement Incentive Program (APIP) in Texas provided teachers awards for students passing Advanced Placement exams. Rank-order tournaments provide teachers awards if they perform better than a certain percentage of other teachers on the metric. An example of this is Houston Independent School District’s ASPIRE program which pays teachers bonuses if they receive value-added scores above the 50th percentile which then double above the 75th percentile. Finally piece-rate systems pay teachers based on incremental improvements in student performance. For example, a piece rate system may pay teachers $100 times his or her value-added score. While economic theory suggests piece-rate systems may be more effective and have less perverse incentives than threshold based systems, districts tend to prefer the latter as it ensures budget security – any relative system will place a cap on total payouts while piece-rate systems could generate far larger than expected payouts. An intriguing compromise between these two methods called “pay for percentile” has recently been proposed. [1] The idea is that teachers are paid based on how students perform relative to a set of observably similar comparison students. The proposers of this method show theoretically that it aligns incentives so that teachers provide optimal levels of effort but this has yet to be shown empirically.

Evidence on Incentive Pay
A fundamental difficulty with evaluating teacher incentives is that districts that choose to have incentives may differ from other districts in unobservable ways. For example, a typical concern is that districts may be more inclined to implement incentive pay programs if they are having problems recruiting effective teachers. Thus any impacts of the plan would pick up the existing, low performing state of the district. Because of this much of the academic research on incentives has turned to using randomized controlled trials (RCTs) to evaluate teacher incentives. The benefit of this is that RCTs eliminate the unobserved differences between teachers affected by incentives and those not affected. However, RCTs are often limited in scope and teachers may respond differently to these experiments since they know they are temporary. Thus I will also consider evidence based on non-experimental data but that use methods that provide causal impacts of incentive pay. Table 1 provides a list of the studies discussed here and provides some key takeaways from those studies.

**Evidence on US Incentive Pay Programs**

Most research on teacher incentive pay has been conducted in the United States. In particular there have been a number of experimental evaluations of incentive pay schemes using various designs. The plans were implemented in many locations across the country though all were in urban or suburban areas. Thus the programs tended to be in districts with large minority and low-income populations. Of the studies described below, it is important also to note that, with the exception of the Chicago Heights experiment, all of the payments were based on teachers meeting thresholds rather than piece-rate incentives.

A particularly widely publicized randomized controlled trial was conducted in the city of Nashville, Tennessee.[2] The program provided teachers with large bonuses of up to $15,000 for math performance. It was a fixed-threshold model that set relatively high thresholds – teachers needed to reach at least the 85th percentile of pre-program value-added scores to receive any award. The study found no statistically significant impact on math scores from the awards. However, the design of the incentives has some key problems that may have reduced effectiveness. First, the high thresholds may have left many teachers unwilling to respond. Second, the focus on mathematics leaves open the question of impacts on other exams. Third, the incentives were entirely based on test score performance. While this is an advantage in some sense as it allows the study to isolate this particularly focused incentive, it nonetheless limits what we learn about incentives more generally.

Nonetheless an experiment in Chicago Heights, Illinois – a suburb of Chicago – found a similar lack of impacts from individual incentives, though I caution that this study is a working paper and thus has yet to undergo peer review.[3] Their results show no significant impact of both individual and small team group incentives based on giving teachers up to $8,000 for student test performance. Despite these null results, there are two particularly unique aspects of this study that are worth mentioning. First, this is the only study that uses the “pay for percentile” incentive system that bases awards on how well a student performs relative to a similar comparison group of students. Second, despite finding no impact for incentives that provided teachers with monetary gains, the authors are also able to test the impacts of monetary losses. This is important as there is substantial economic evidence that people care more about losing money more than gaining money, even if the amounts are the same. This concept is referred to as “loss aversion.” The study replicated this by giving all teachers a bonus prior to the start of the school year and had them sign contracts that required them to pay it back after the year ended if their students did not perform sufficiently well on an end-of-year exam. This simple change in the structure of the incentive program generated very large positive impacts on test scores. Nonetheless, while intriguing, such a method may be hard to implement in practice.
Another randomized experiment in individual incentives was Chicago’s implementation of the Teacher Advancement Program which paid teachers up to $6,400 based on a mix of value-added, class observations and teachers’ involvement in the school. Unlike the previous two experiments, entire schools were randomized into earlier or later adoption of the program. An advantage of this methodology is that it better reflects how adoption would occur in practice, as typically whole schools or districts would adopt en masse. Even so, as with the other studies mentioned, they find no significant impacts of incentives after one year.

Two studies, nonetheless find positive impacts of incentives on student performance. The first evaluates a unique characteristic of the IMPACT incentive program implemented in Washington, DC. This program provided teachers with an opportunity to earn one-time bonuses and permanent salary increases of up to $27,000 per year, which makes it considerably more expensive than typical programs that provide temporary bonuses but also provides an especially large incentive. As in the Chicago TAP program, incentive receipt was determined by a metric that incorporated teacher value-added scores, classroom observations and the teacher’s involvement in the school. While the program was not implemented experimentally, the authors use the fact that to qualify for the increases in permanent salary teachers needed to be rated “highly effective” two years in a row. This means that those who were just barely rated highly effective in the first year have a much stronger incentive in the second year than those who just barely miss even though they are otherwise virtually identical. Hence, the study is able to compare these two types of teachers. Strikingly, with the study does show significant positive impacts on teacher performance. However, as with the study in Chicago Heights, one must be cautious in using this evidence as it has not yet undergone peer review. The second study looks at Minnesota’s Q-Comp program. Unlike the programs discussed above, Q-Comp gave school districts substantial flexibility in the design of the incentives including the factors used and whether group or individual incentives are used, thus the study does not differentiate between incentive types. Nonetheless, by comparing districts based on their timing of adoption and whether or not they adopt at all, the paper finds small positive effects for reading but none for math.

The research described above indicates that the impacts for individual incentive awards in the US are mixed and, at best, one needs very large awards for them to be effective. Nonetheless, it remains possible that group-based awards are able to capitalize on encouraging cooperation amongst teachers despite the potential for free-riding. Once again, though, the evidence here is mixed. The best evidence on the impacts of group awards in the US comes from two studies that looked at schools in New York City that were randomly assigned to an incentive program. While the program was designed to give schools flexibility in defining how incentives were distributed (though they had to be awarded based on test scores), in practice nearly all schools adopted incentives based on average school-wide achievement of around $3,000 per-teacher. In this case there is no significant positive impact on math or reading scores and in middle schools there was even a significant negative impact. However, one of the studies points out that smaller schools had better responses to incentives suggesting that free-riding, which would be a bigger problem in large schools, plays an important role in group incentives.

A study that looks at Houston’s ASPIRE program is able to test this free-riding issue more directly. The study focuses on high school teachers where awards are provided at the subject-grade level (e.g. science teachers in 9th grade) providing substantial differences in the sizes of groups. The awards are entirely based on test score value-added and were as large as $7,700 per teacher. The study finds substantial evidence of free-riding in large groups of teachers indicating that the most effective group incentives are those that put teachers in teams of 5 or fewer. Though the paper does not provide direct impacts of the
incentives, the authors are able to calculate impacts indirectly from the free-riding estimates and find large positive effects of incentives on incentivized exams. However there was no impact on exams in the same subjects but were not linked to the incentives. This highlights the important problem that teachers may target their efforts on the exams themselves rather than overall learning and provides support for adopting a range of metrics for awards.

International Evidence on Teacher Financial Incentives

Outside the United States there has been much less research on teacher incentives. Even so, there are some key studies that show that the international evidence on incentives is much more positive than the US evidence. To start two studies in Israel find substantial positive impacts of teacher incentives on student performance.[10][11] Given that Israel is a developed country with an education system similar to those in the US and Europe, one could consider this additional evidence, combined with the US evidence, on how incentive pay works in developed countries. The papers focus on two different incentive systems that reward teachers for how well their student performed on the Bagrut a combined high-school exit and college entrance exam. Though they are not randomized experiments, they nonetheless are done in ways that their results should be considered to be causal. The first study estimates the impact of a relatively low-stakes school-wide (group) incentive program that gave teachers up to $1,000 each. The second looks at an individual incentive program that had much higher payouts – as much as $7,500. In both cases the incentives had positive and statistically significant impacts on student performance.

As shown above there is wide ranging evidence on the impacts of teacher performance pay in developed countries, mostly the United States, but our knowledge of impacts in developing countries is far more limited. Nonetheless, two important experiments provide us with some insight that suggests that in developing contexts incentives can be highly effective and are far cheaper to implement. The first study was conducted in the state of Andra Pradesh in India.[12] This randomized controlled trial assigned teachers either to having no incentive, a school-wide incentive or an individual incentive. As a percent of a teacher’s salary the incentives were substantial, but in absolute terms they were very inexpensive – typically less than $100 per teacher. The incentives were also based on a piece-rate system rather than a threshold system. The study finds large and statistically significant impacts on math and language achievement from individual incentives with a still sizable and significant, but 50% smaller, impact for group incentives. It should be noted also that, unlike schools in the US, the schools here are small, averaging only three teachers per school. The New York City incentive schools, for example, typically averaged 16 incentivized teachers. Hence, the existence of group incentive effects here are consistent with the Houston study that showed small groups respond more. It is also worth noting that the authors attribute much of this impact to the fact that incentivized teachers put more effort into student preparation for the exams including through additional test prep, more homework/classwork, extra tutoring outside of school hours, and paying extra attention to weaker students.

The second developing country experiment was conducted in Kenya.[12] Schools were randomly assigned to have incentives that provided prizes worth up to $51 based on average test score performance in the school. The incentives were successful at improving student performance on the incentivized exams. However, as in the Houston study, there was little evidence of impacts on non-incentivized exams in the same subjects.

Background Information
1) **Teacher value-added.** Using test scores alone for teacher evaluation is problematic as it does not distinguish between the teacher’s effectiveness and the existing ability of the students. Value-added uses statistical adjustments to test-score growth to isolate the teacher’s contribution to student achievement.

2) **Piece-rate compensation.** Paying individuals for production on a per-unit basis.

3) **Threshold compensation.** Paying individuals for reaching production thresholds.

4) **Pay for percentile.** A method of paying teachers for student performance proposed by Barlevy and Neal (2012) that bases payments on how students rank in achievement gains relative to an observably similar comparison group of students.

5) **Free riding.** A person free-rides off another if he/she reduces his/her contribution towards achieving a common goal in response to an increase in the contribution from someone else.

**Limitations and Gaps**

While there are a number of excellent studies evaluating teacher incentive pay, there remains much that is not known. First of all, the research is heavily centered on the United States. Only a handful of rigorous studies have been conducted in other countries, none of which are from Europe, East Asia, or South and Central America. Second, most of the incentive schemes studied are based on thresholds where teachers must achieve certain achievement levels to receive bonuses. This is mainly due to the attractive feature that such methods (especially those based on rankings of teachers) provide certainty in budgeting. However, theoretically we would expect threshold incentives to be less effective than piece-rate methods that pay for each unit of additional performance. Third, direct comparisons of group and individual incentives are rare. While both have been studied in different contexts, comparisons in the same location are limited. Fourth, there is a lack of empirical evidence on how basing incentives on multiple outputs compare to basing incentives on single outputs like teacher value-added.

**Summary and Policy Advice**

Table 2 provides a broad overview of the research discussed in this essay. In general the evidence on the impacts of incentives is mixed. While they appear to be quite successful in developing countries, the results are unclear in developed countries like the US and Israel, though tend to weigh more towards positive effects than negative. Even so, in cases where there are positive impacts it appears that the effects are centered on the directly incentivized exams which indicate that the financial incentives may not improve general learning if they are narrowly targeted. Studies that look at incentives based on multiple outcomes tend to show more positive effects. As a result of this research, the following recommendations are made for policy makers considering implementing teacher incentive pay to improve student performance.

1) Teacher incentives appear to have the potential to improve student outcomes but the design of the system matters and poorly designed systems can even make outcomes worse.

2) Incentives should be based on multiple outcomes where teacher value-added is one of a few metrics, at least one of which should be subjective (e.g. principal evaluation or classroom observations).

3) When possible, thresholds and tournaments should be avoided in favor of piece-rate systems. “Pay for percentile” is a promising method but remains to be empirically tested.
(4) Group incentives should keep the size of groups low. Incentives at the school level are typically ineffective, but the evidence suggests that small groups can generate improvements in student performance.
Key References


Further Readings


Additional References


Figure 1: Percent of Public and Charter School Districts with Teaching Incentives by US States in 2011-2012

Source: Schools and Staffing Survey 2011-2012. Available at http://nces.ed.gov/surveys/sass/tables/sass1112_2013311_d1s_005.asp. Hawaii, Iowa, Montana, New Jersey, Oklahoma, South Dakota and West Virginia are missing due to insufficient data.
<table>
<thead>
<tr>
<th>Study</th>
<th>Location (Program Name)</th>
<th>Incentive Type</th>
<th>Max Award</th>
<th>Student Achievement Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dee &amp; Wyckoff (2014)</td>
<td>Washington, DC (IMPACT)</td>
<td>Individual based on observations, school involvement, and test score value-added.</td>
<td>$27,000 per year permanently</td>
<td>Positive on IMPACT scores that include, as part, value-added from test scores.</td>
</tr>
<tr>
<td>Fryer (2013) and Goodman &amp; Turner (2013)</td>
<td>New York City</td>
<td>Schools were permitted to decide on metric but in practice almost all schools used school-wide group awards based on test scores.</td>
<td>$3,000 one-time bonus that varied by position.</td>
<td>Negative and significant for middle school, negative and insignificant for elementary on math and reading.</td>
</tr>
<tr>
<td>Fryer, et al. (2013)†</td>
<td>Chicago Heights, IL</td>
<td>Individual or small group award based on &quot;pay for percentile&quot; method of Barlevy and Neal (2012).</td>
<td>$8,000 one-time bonus</td>
<td>No significant effects for standard awards but large positive effects when money is taken away from low performing teachers.</td>
</tr>
<tr>
<td>Glazerman &amp; Seifullah (2012)</td>
<td>Chicago (Teacher Advancement Program)</td>
<td>Individual based on observations, school involvement, and test score value-added.</td>
<td>$6,400 one-time bonus</td>
<td>No statistically significant effect after one year.</td>
</tr>
<tr>
<td>Glewwe, Ilias and Kremer (2010)</td>
<td>Busia &amp; Teso, Kenya</td>
<td>Group awards at the school level based on test-score levels.</td>
<td>Prizes worth up to $51.</td>
<td>Positive for incentivized exam but no significant effect on non-incentivized exams.</td>
</tr>
<tr>
<td>Imberman &amp; Lovenheim (2015)</td>
<td>Houston, TX.</td>
<td>Group awards at the subject-grade level based on teacher value-added.</td>
<td>$7,700 one-time bonus</td>
<td>Positive for incentivized exam but zero impact on non-incentivized exams.*</td>
</tr>
<tr>
<td>Lavy (2002)</td>
<td>Israel</td>
<td>Group awards at the school level based on improvements in school-wide average scores.</td>
<td>$1,000 one-time bonus</td>
<td>Positive effects on average test scores.</td>
</tr>
<tr>
<td>Lavy (2009)</td>
<td>Israel</td>
<td>Individual awards based on test scores.</td>
<td>$7,500 one-time bonus</td>
<td>Positive effects for math and English.</td>
</tr>
<tr>
<td>Muralidharan &amp; Sundararaman (2011)</td>
<td>Andhra Pradesh, India</td>
<td>Individual or group awards at the school level based on test score gains.</td>
<td>$11 per percentage point increase in average gain.</td>
<td>Significant positive effects on language and math for both award types. Group effects about 1/2 size of individual effects.</td>
</tr>
<tr>
<td>Sojourner, Mykerezi &amp; West (2014)</td>
<td>Minnesota (QComp)</td>
<td>Mix of individual and group based on multiple factors including observations and test scores.</td>
<td>Varied by district. Average maximum of $2,200 one-time bonus.</td>
<td>Positive for reading no significant effect for math.</td>
</tr>
<tr>
<td>Springer, et al. (2012)</td>
<td>Nashville, TN (POINT Experiment)</td>
<td>Individual awards based on teacher value-added.</td>
<td>$15,000 one-time bonus</td>
<td>No significant effects on math.</td>
</tr>
</tbody>
</table>

† Denotes study that had not completed peer review at the time of writing and thus results are not final.
* Note that Imberman & Lovenheim (2015) does not identify the impact directly and rather estimates impacts indirectly through the effect of changing the teacher’s impact on likelihood of award receipt.
<table>
<thead>
<tr>
<th>Incentive Type</th>
<th>United States &amp; Israel</th>
<th>India &amp; Kenya</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Results</td>
<td>Number of Locations</td>
<td>Results</td>
</tr>
<tr>
<td>Individual Incentives</td>
<td>Mixed zero and positive.</td>
<td>6</td>
<td>Positive</td>
</tr>
<tr>
<td>Group Incentives</td>
<td>Mixed negative and positive.</td>
<td>5</td>
<td>Positive</td>
</tr>
<tr>
<td>Threshold Incentives</td>
<td>Mixed negative and positive.</td>
<td>7</td>
<td>Positive</td>
</tr>
<tr>
<td>Piece-Rate Incentives</td>
<td>Mixed zero and positive.</td>
<td>2</td>
<td>Positive</td>
</tr>
<tr>
<td>Directly Incentized Exams</td>
<td>Mixed negative and positive.</td>
<td>9</td>
<td>Positive</td>
</tr>
<tr>
<td>Non-Incentivized Exams</td>
<td>Zero</td>
<td>1</td>
<td>Zero</td>
</tr>
<tr>
<td>Multiple Outcomes</td>
<td>Mixed zero and positive.</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>Test Scores Only</td>
<td>Mixed negative and positive.</td>
<td>6</td>
<td>Positive</td>
</tr>
<tr>
<td>Total Number of Locations</td>
<td></td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>