NBER WORKING PAPER SERIES

INCENTIVE STRENGTH AND TEACHER PRODUCTIVITY: EVIDENCE FROM A GROUP-BASED TEACHER INCENTIVE PAY SYSTEM

Scott A. Imberman
Michael F. Lovenheim

Working Paper 18439
http://www.nber.org/papers/w18439

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2012

We wish to thank seminar participants at Aarhus University, CESifo, Cornell University, the Institute for Research on Poverty, NBER Summer Institute, Purdue University, Tilburg University, the University of Copenhagen, and the University of Houston for helpful comments and suggestions. We gratefully acknowledge the help and input of Kiel Albrecht, Jack Barron and Steve Coate in the development of our theoretical model. We further thank Aimee Chin, Steven Craig, Steve Rivkin, Gary Solon and Lesley Turner for helpful comments and suggestions. Finally, we would like to thank the employees at Houston Independent School District for their help and assistance. All errors, omissions and conclusions are our own. Copyright 2012 by Scott Imberman and Michael Lovenheim. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System
Scott A. Imberman and Michael F. Lovenheim
NBER Working Paper No. 18439
October 2012
JEL No. H41,I21,J33,J38

ABSTRACT

Using data from a group incentive program that provides cash bonuses to teachers whose students perform well on standardized tests, we estimate the impact of incentive strength on student achievement. These awards are based on the performances of students within a grade, school and subject, providing substantial variation in group size. We use the share of students in a grade-subject enrolled in a teacher's classes as a proxy for incentive strength since, as the teacher share increases, a teacher's impact on the probability of award receipt rises. We find that student achievement improves when a teacher becomes responsible for more students post program implementation: mean effects are between 0.01 and 0.02 standard deviations for a 10 percentage point increase in share for math, English and social studies, although mean science estimates are small and are not statistically significant. As predicted in our theoretical model, we also find larger effects at smaller shares that fall towards zero as share increases. For all four subjects studied, effect sizes start at 0.05 to 0.09 standard deviations for a 10 percentage point increase in share when share is initially close to zero and fade out as share increases. These findings suggest that small groups provide productivity gains over large groups. Further, they suggest that the lack of effects found in US teacher incentive pay experiments probably are in some part due to specific aspects of program design rather than failure of teachers to respond to incentives more generally.

Scott A. Imberman
Michigan State University
486 W. Circle Drive
110 Marshall-Adams Hall
East Lansing, MI 48824-1038
and NBER
imberman@msu.edu

Michael F. Lovenheim
Department of Policy Analysis and Management
Cornell University
135 Martha Van Rensselaer Hall
Ithaca, NY 14853
and NBER
mfl55@cornell.edu

1 Introduction

Teacher incentive pay has become an increasingly popular policy in the United States. A key component of many of these incentive schemes, and of performance pay systems in general, is group incentives. In education, this typically amounts to paying teachers based on grade- or school-specific performance on standardized exams in a given subject. The literature to date has found mixed results with respect to the effectiveness of these programs. Lavy (2002) finds evidence that school-wide incentives in Israeli schools boost student achievement as does Muralidharan and Sundararaman (2011) in India. Nonetheless, studies focusing on the United States generally have found small or negligible effects of group incentive pay on performance (Fryer, 2011; Springer, et. al., 2010; Sojourner, West and Mykerezi, 2011; Goodman and Turner, 2011), although Ladd (1999) finds positive effects of a school-based incentive pay system implemented in the early 1990s in Dallas and Fryer et al. (2012) find effects when there is loss aversion.

One drawback of providing rewards based on group performance is the "$\frac{1}{n}$" problem (Holmstrom, 1982; Kandel and Lazear, 1992), whereby teachers in larger groups may be less responsive to financial incentives. The "$\frac{1}{n}$" problem comes about because of "free riding," where workers have an incentive to reduce their effort when others in the group exert more effort, and because of "award salience" effects, where the incentive to work hard is reduced when group size increases, which causes each worker to have less of an impact on the likelihood of winning the award.[1] Further, larger groups make it more difficult for workers to monitor each other's effort, which can encourage shirking.[2] On the other hand, in the case of group incentives for teachers, the existence of larger groups can have positive effects through spillovers, such as the development of team teaching practices, mentoring, technology adoption, peer effects, and mutual professional support. How individual performance responds to group-based incentives thus is an empirical question about which little currently is known in general, and especially

[1] In the context of this paper we define "effort" broadly to include not only an increase in quantity (e.g. time working) but also quality (more effective use of time) and actions that increase the use of productivity enhancing technologies.

[2] In the context of supermarket workers, Mas and Moretti (2009) show that peer-monitoring can substantially reduce free-riding behavior.

with regard to teachers.

In this paper, we test directly for whether the strength of a group-based incentive a teacher faces affects her productivity using the implementation of the ASPIRE[3] teacher incentive pay program in the Houston Independent School District (HISD). In the 2006-2007 school year, HISD began a rank-order tournament incentive pay program that pays teachers based on the relative value-added of their students' performance on math, English, science and social studies state exams. In 2007-08 and later for high school teachers, the incentives are group-based, rewarding teachers for the performance of all students in each year-school-grade-subject group. The awards are allocated using a rank-order tournament for each subject, with sharp cutoffs in award amounts at the 50^{th} and 75^{th} percentiles of the district-wide, subject-grade value-added distribution. The award amounts are substantial: the maximum award in the 2009-2010 school year was $7,700.[4]

Our analysis begins with a simple model of worker effort in which we develop predictions for how teachers will respond to the implementation of an output-based rank-order tournament incentive pay system as a function of the proportion of the output for which they are responsible. We model a group incentive pay system with two teachers per group, where teachers differ only in the share of students they teach. We focus on the effect of a teacher's share on her own effort and find that, in general, the model yields ambiguous predictions for how teacher share should affect teacher effort. When we parameterize the model and solve it numerically, our results point to increasing share having mostly a positive effect on effort but that the effect is non-linear in share. At small shares, increasing share has large, positive effects on effort that declines with the proportion of students in a group teachers instruct, such that it approaches zero or even becomes negative for some parameter values.

We then take these theoretical predictions to the data, using individual student-level data from before and after implementation of the incentive pay program. We test for responses to incentive strength by examining whether teachers who are responsible for a larger share

[3] ASPIRE stands for "Accelerating Student Progress, Increasing Results and Expectations."

[4] The maximum award includes a 10% bonus for perfect teacher attendance. Additional smaller awards based on school-wide performance metrics also are provided. The maximum for these awards combined was $3,410 in 2009-10 when applying the attendance bonus.

of students in each grade and subject generate more achievement gains after implementation of the award system than those who are responsible for teaching fewer students. Under a group incentive scheme, the share of students a teacher instructs is a strong proxy for incentive strength because, as the teacher share increases, a teacher's impact on the probability of award receipt rises. We argue that this share is a more direct measure of cross-teacher incentive differences than the number of workers in the group, which is the measure typically used to examine group-size effects. Thus, our key explanatory variable is the share of a subject-school-grade cell enrolled in each teacher's classes, and we identify how the effect of this share changes when the incentive pay program is implemented using a difference-in-difference methodology. By controlling for pre-ASPIRE share, lagged student test scores, student demographics, school-year and grade-year fixed effects, we argue our empirical models account for the non-random sorting of students into classrooms with teachers of differing quality and who teach a larger or smaller share of students. The key identifying assumption we invoke is that the effect of share taught on student achievement is not shifting systematically when the incentive system is implemented for reasons not having to do with the program. We present extensive evidence that this assumption holds in our data.

Doubts about whether teachers respond to financial incentives, at least in the United States, arise from the fact that recent experimental studies in the US have found little overall impact of incentive pay on achievement (Fryer, 2011; Springer et al., 2010; Goodman and Turner, 2011). Hence, a major contribution of our analysis is to establish that teachers are responding to incentive pay. In particular, we find that teacher productivity increases as financial incentives become stronger. The estimates are largest for math, where we find that a 10 percentage point increase in the share of students taught post-ASPIRE increases test scores by 0.024 standard deviations. A similar increase in teacher share increases achievement in English and social studies by 0.014 and 0.020 standard deviations, respectively. There is no effect on science scores, on average. We show as well that teachers shift their focus across grades in response to the program, such that among students in different grades in the same year with the same teacher, test performance increases more for students in the higher-share classroom. Further, using another math exam that is not incentivized under this system, we test whether the achievement gains in math are specific to the incentivized exam. Despite the large effects we find on the incentivized math exam, there is no impact on the non-incentivized exam.

Consistent with our theoretical model, our estimates show that there are large, positive effects of increasing share on achievement at low shares in all four subjects post-ASPIRE and that this effect declines with share. The effect of increasing share by 10 percentage points is between 0.05 to 0.09 standard deviations at very low shares and falls until reaching zero at shares between 0.2 to 0.3. These results are unlikely to be driven by monitoring: conditional on share, there is no change in the relationship between the size of the department and student test scores. Our results thus suggest that there are large returns in terms of student achievement to incentivizing smaller groups of teachers but that these returns disappear as group sizes decline sufficiently. A back-of-the-envelope calculation based on our estimates, along with some assumptions about the nature of the groups, points to achievement-maximizing group sizes of between 3-5 teachers, which is far smaller than the school-wide groups common in teacher incentive pay programs.

In showing that there are heterogeneous effects of group incentive pay on teacher effectiveness, this paper demonstrates that previous work examining average effects may miss an important part of how group incentives operate. While we cannot determine whether our results are due to award salience or free-riding behavior, they nonetheless indicate that targeting incentives to smaller groups likely yields larger achievement improvements than using large groups. To date, there has been little work on how to optimally structure teacher incentive pay,[5] and the results from this analysis indicate that design features matter a lot in determining how effective an incentive system is in increasing student achievement. Our results underscore the importance of focusing on such design issues in future work.

The rest of this paper is organized as follows: Section 2 describes the previous literature on teacher incentive pay and on free riding in group incentive programs. Section 3 describes our theoretical model. In Section 4 we describe the HISD incentive pay program, and our data are discussed in Section 5. We present our empirical methodology in Section 6. All results are

[5] Barlevy and Neal (2012) provide the only theoretical work on optimal teacher incentive pay of which we are aware.

discussed in Section 7, and Section 8 concludes.

2 Previous Literature

The prior literature on teacher incentive pay mostly focuses on group-level incentives. However, these studies typically examine whether there is an average effect of these incentive pay systems on student achievement, not how individuals respond to their specific incentives for increasing output.[6]

Lavy (2002) studies a school-wide performance incentive program in Israeli Public Schools that was implemented in 1995. Schools received bonuses based on dropout rates, the average number of credit units per student and the proportion of students receiving a matriculation certificate. Eligibility for the program is based on school type and geography, creating treatment and control groups that are of the same type and that are observationally similar. His main finding is that the school-based incentives led to an increase in student test scores, a decrease in dropout rates and an increase in the proportion of students receiving a matriculation certificate. He does not examine whether teachers in schools with more teachers are more or less responsive to the implementation of the award system.

Lavy (2009) analyzes another teacher bonus tournament in Israeli schools that began in 2000. He finds that among teachers in eligible schools, test-taking, passing, and scores on a math high school exit test increase significantly due to exposure to financial incentives. While Israel is a developed country, there are substantial differences between the Israeli and US public education systems, making it unclear how relevant these findings, as well as the findings in Lavy (2002), are to the United States.[7] In addition, experimental studies in developing countries have found positive effects of both group and individual incentive pay on student outcomes (Muralidharan and Sundararaman, 2011; Glewwe, Ilias and Kremer, 2010). However, the educational contexts in these countries differ substantially from the United States, which makes their results difficult to generalize to the US schooling environment.

[6] See Neal (2011) for a review of the teacher incentive pay literature.

[7] For example, Lavy (2009) reports that the predominant way in which teachers reacted to the financial incentive was to increase instruction time, which would be difficult for teachers in the US system to do without district- or school-level policy changes.

In the United States, several studies have use randomized experiments to assess the average impact of school-level group incentive pay in New York (Fryer, 2011; Goodman and Turner, 2011) and individual incentive pay in Nashville (Springer, et al., 2010). They find no significant impact of teacher incentives on student performance on average. Sojourner, West and Mykerezi (2011) examine the effect of Minnesota's Q-Comp pay for performance system. In this system, schools can choose whether to provide incentives at the individual teacher or school level. They find very small but significant positive effects. Ladd (1999) estimates the effect of a school-based, rank-order incentive pay system that was implemented in Dallas from 1991 through 1995. She compares trends in academic achievement in Dallas to schools in other large cities over this period, and she shows evidence that academic performance in Dallas rose relative to these other cities. Her empirical methodology cannot differentiate between incentive effects and differential secular trends or shocks across cities, though.[8] Finally, Fryer, et al. (2012) conduct an experiment that gave some teachers award bonuses and other teachers fixed cash pay-outs prior to the school year who were then required to return some of the money based on performance. Consistent with the rest of the literature, they find no significant impacts from the first group, however they do find improvements from the second group. This paper highlights how the particular design of a program matters. It is also interesting to note that they find some larger impacts for two-teacher group awards relative to individual awards in the second group.

The most closely related study to our own is Goodman and Turner (2011). They use variation in the number of math and English teachers in each school in a school-level randomized teacher incentive pay experiment in New York City to identify whether free-riding reduces the program's effectiveness. They present evidence that achievement declined slightly in larger schools and may have increased by a small amount in smaller schools. However, given the aggregate nature of their data at the school-year level, they cannot test whether the differences in responsiveness by school size are due to free-riding or whether they are due to school

[8] Jackson (2010, 2012) examines the effects of the Advanced Placement (AP) Incentive Program in Texas and finds that offering student and teachers incentives to pass AP exams increases test taking, test passing, college-going and future earnings. However, given the structure of the award program, he is unable to disentangle the impact of teacher-specific awards, *per se*, on student outcomes from the effect of offering both students and teachers financial incentives to pass AP exams.

attributes that are correlated with school size. Furthermore, without individual teacher data, their analysis cannot examine whether there are non-linear effects of group size. Nonetheless, the results from this analysis point to the potentially important role that group size plays in determining how teachers respond to group-based incentives.[9]

Outside of education, there has been more work examining how group incentive schemes influence productivity. Prendergast (1999) provides an overview of this literature. Several of these studies suggest that workers respond less to group incentives when they are part of a larger group (Newhouse, 1973; Liebowitz and Tollison, 1980; Gaynor and Pauly, 1990). While suggestive of the existence of free-rider behavior, none of these analyses are able to control fully for the endogeneity of group size nor do they have exogenous variation in award amounts (i.e., in the returns to effort). Hamilton, Nickerson and Owan (2003) do have exogenous variation driven by a garment plant switching from individual to group-based piece rate pay. They find positive productivity effects of this switch, which are due to increased worker cooperation.

Thus, despite the sizable previous literature on group-based merit pay, little is known about how group incentives impact worker behavior. In education, very little attention has been paid to the effects of group-based teacher incentive pay on teacher behavior when there are many teachers that dilute each worker's impact on the likelihood of receiving an award. Given the pervasive nature of group-based incentive pay in education and in the private sector, understanding how group size interacts with teacher behavior is critical to developing optimal merit pay systems. The unique structure of the HISD teacher incentive pay system for high school teachers provides an unusually clean test of the impact of the strength of group incentives on individual behavior that will allow us help fill this gap in the literature.

[9] Ahn (2011) also presents evidence that free-riding may exist in group incentive pay systems in education. He estimates a structural model of teacher effort and student achievement in which he proxies for teacher effort with teacher absences to analyze a school-level incentive pay system in elementary schools in North Carolina. He estimates free-rider effects by simulating optimal effort responses by teachers in response to a change from school to classroom level incentives. While his parameter estimates point to an increase in average bonus receipt, average teacher effort declines due to a change in which teachers find themselves marginal to an award threshold. These estimates are consistent with a role for group size, but they are only suggestive because he is unable to disentangle group-size effects from changes in the marginal incentives of teachers.

3 A Model of Teacher Responses to Group Incentives

In order to illustrate the role of teacher share in a rank-order incentive pay tournament and to motivate our empirical analysis, we start with a theoretical model based on Kandel and Lazear (1992) that we modify to account for the tournament structure of the award system. We assume workers are homogenous, except for the share of the output for which they are responsible. We impose this simplification in order to isolate the role of the output share in driving effort differences across workers in a tournament model. Throughout, we assume that share is exogenously assigned.[10] Teachers choose effort, e, which determines value-added through the function $S(e)$, where $S'(e) > 0$ and $S''(e) < 0$. This function does not differ across teachers by assumption. The cost of effort, $C(e)$, also is the same for each teacher and is convex in effort: $C'(e) > 0$ and $C''(e) > 0$.

Teachers compete in a rank-order tournament for a monetary award, A. Teachers only win the award if the teacher's group achieves value-added at or above a critical value S^*, which is randomly drawn from an i.i.d., single peaked continuously differentiable distribution with mean μ, variance σ^2 and highest density at μ. The CDF for this distribution is defined by $F(x)$, with pdf $f(x) = F'(x)$. While S^* is potentially endogenously determined by strategic interaction of teachers, in practice since there are many teachers and many groups competing in the tournament we assume that teachers take S^* to be exogenous.[11] S* is unknown to teachers, but they do know the distribution that generates S*. The teacher's utility is

$$U_i = A \times 1\Big(\sum_k^N \theta_k S(e_k) > S^*\Big) - C(e_i) \tag{1}$$

where $1(\cdot)$ is the indicator function, θ is the teacher's share of students in a school-subject-grade and $\sum_k^N \theta_k = 1$. Hence, if the group exceeds the critical value, the teacher receives a flat award of A.

[10] This is an innocuous assumption in this model, because with equal ability teachers, principals will have no incentives to assign higher shares to the "best" teachers.

[11] Note that even with this simplifying assumption, this model yields quite complex predictions about the relationship between teacher share and effort. Endogenizing S^* will add considerable complexity without, we believe, adding much more insight into this relationship because each teacher would exert at most a small effect on S^* due to the large number of teachers and groups.

The effect of the award incentive on each teacher's behavior will be a function both of the teacher's own share of output and the effort other teachers in the group exert. We therefore must solve for the Nash equilibrium in this static game. In order to make this problem easier to solve, consider a group with two teachers, 1 and 2. Each has the same expected utility function given by equation (1). Define $\bar{S} = \theta S(e_1) + (1-\theta)S(e_2) - \mu$ as the group-average value-added less the mean of the cutoff distribution (e.g., $\bar{S} > 0$ if the mean value-added is greater than the expected cutoff). The expected utility functions for teachers 1 and 2 are given by:

$$E(U_1) = AF(\bar{S}) - C(e_1) \tag{2}$$

$$E(U_2) = AF(\bar{S}) - C(e_2) \tag{3}$$

These expected utility functions lead to two first order conditions, which are the best response functions for teacher 1 and teacher 2, given the award structure and exogenously assigned teacher share:

$$A\theta F'(\bar{S})S'(e_1) - C'(e_1) = 0 \tag{4}$$

$$A(1-\theta)F'(\bar{S})S'(e_2) - C'(e_2) = 0 \tag{5}$$

The intersection of these best response functions implicitly define the optimal amount of effort for each teacher, e_1^* and e_2^*. The empirical focus of this paper is on identifying $\frac{\partial e_1^*}{\partial \theta}$. However, solving for optimal effort through applying the implicit function theorem to equations (4) and (5) and then calculating $\frac{\partial e_1^*}{\partial \theta}$ yields a complicated expression that cannot be signed in general.[12] Thus, in order to give us insight into how optimal effort in this tournament setting varies with output share, we parameterize the model and solve it numerically.

Let $S = \sqrt{e}$, $C = \frac{1}{2}e^2$ and A=1. These functions are chosen to be concave and convex, respectively, and we note that the resulting comparative statics are similar using alternative concave and convex functions for S and C. We parameterize $F(\cdot)$ with a normal CDF, $\Phi\big(\frac{\bar{S}}{\sigma}\big)$, and we will explore below how the model predictions vary with the mean and variance of this

[12] The full closed-form solution is provided in Appendix B. We note that there is clearly no general sign here as below we show examples where $\frac{\partial e_1^*}{\partial \theta}$ is positive and examples where it is negative.

distribution. The first order conditions using these parameters are:

$$\frac{\theta\phi\left(\frac{\bar{S}}{\sigma}\right)}{2\sqrt{e_1}} - e_1 = 0 \tag{6}$$

$$\frac{(1-\theta)\phi\left(\frac{\bar{S}}{\sigma}\right)}{2\sqrt{e_2}} - e_2 = 0 \tag{7}$$

where $\phi(\cdot)$ is the standard normal PDF. The set of equilibrium achievement values (and by extension teacher effort) of teacher 1 as a function of θ, which is teacher 1's share, is shown in Figure 1 for different values of μ and σ.[13] The relationship between effort and own share is concave for most values of θ in each μ-σ combination, although these parameters affect the specific shapes of these curves.[14] Note that since the share for teacher 2 is $1-\theta$, the equilibrium achievement values of that teacher's students would be defined by the mirror image figure. Also note that, since achievement is a concave function of effort, the shape of this relationship will show the same patterns as the direct relationship between θ and effort.

In each case, increasing teacher share for teachers with low share increases achievement by much more than increasing share for those with a higher share. This non-linearity is more apparent in Figure 2, which shows $\frac{\partial S(e_1^*)}{\partial \theta}$, calculated empirically for each 0.001 change in θ. The effect of increasing teacher share on achievement is much larger at lower share levels, and it declines with share. This finding is robust to different assumptions about the mean and standard deviation of the cutoffs for award eligibility, although for $\sigma = 0.5$ and $\mu = 1.5$ the effect of share on achievement approaches zero at lower share levels. This non-linear effect is driven by several factors. First, free-rider effects are larger at lower levels of θ. Increasing θ for those with low share reduces the incentive to free ride more than increasing θ for a high share teacher, which translates into a larger effort increase. Second, because the marginal cost of effort is increasing in effort and the marginal benefit of effort is decreasing in effort, there is a limit to how much a teacher will increase her effort, regardless of the strength of the incentives. As share increases, teachers increase effort, but because effort is costly and due to diminishing marginal returns to effort, this increase occurs at a decreasing rate.

[13] The values of μ are different for each σ in order to make the range of means examined be roughly proportional to the standard deviation.

[14] The exceptions are for $\sigma = 0.5$ and $\mu = 1$ or $\mu = 1.5$, where the curve becomes convex when $\theta > 0.7$.

As discussed above, in general, $\frac{\partial e_1^*}{\partial \theta}$ cannot be signed. While the derivatives shown in Figure 2 are mostly positive, the effect of increasing share can become negative when share is high and σ is small and/or μ is small. This result stems from the fact that increasing θ reduces the share for the other teacher (1 - θ). Since at high θ the first teacher is exerting a lot of effort while the second teacher is exerting little effort, the convexity of the cost function implies that teacher 1 can get large utility gains from reducing effort. This effect is strongest when the cutoff is known with more certainty, as then it becomes easier for teachers to target effort. Furthermore, the model predicts that teacher effort is particularly sensitive to share variation at low shares. Thus, reducing teacher 2's share when she only teaches a small proportion of the students could lead to a sizable reduction in her own effort, which would reduce the group value-added. If the group is far from the expected cutoff, this reduction in effort could place the group too far to reasonably expect to be in contention for the award, which also could induce teacher 1 to reduce effort.

With the exception of these more extreme cases, Figures 1 and 2 show that in a theoretical model in which there are two teachers in each group and heterogeneity is only a function of share, increasing share tends to have a positive effect on student achievement that declines with share (e.g. $\frac{\partial S(e_1^*)}{\partial \theta} > 0$ and $\frac{(\partial S(e_1^*))^2}{\partial^2 \theta} < 0$). As a result, this model provides two testable predictions. First, an increase in the share of students for whom a teacher is responsible should generate an increase in average achievement on the incentivized exam. Second, this effect becomes smaller as the teacher share increases. We will test both of these predictions in our empirical analysis below. The model also predicts that under certain conditions increasing teacher shares when they are already high could have a negative impact on achievement. While this is an intriguing result, unfortunately we will not be able to test this prediction directly as our data do not have sufficient variation at high teacher shares (see Figure 3).

4 The ASPIRE Teacher Incentive Pay Program

The Houston Independent School District is one of the largest school districts in the United States, with more than 200,000 students enrolled. The district began providing teachers bonus compensation for the performance of their students on standardized exams in 2005-06. The initial program contained a mix of school-level and individual teacher rewards based on student achievement growth on the Stanford Achievement Test and the state accountability test. In total, teachers who taught "core" courses - math, reading, science, social studies and English\language arts - could receive up to $6,000 in payments above their base pay. There were no rewards provided at the department level that year; all awards were either individual or school-wide. In the 2006-2007 academic year, all merit-based bonuses were awarded at the school-wide or school-subject level.

The current incarnation of ASPIRE started in the 2007-08 academic year, when HISD modified the teacher award for high school teachers so that they are determined within grade and subject rather than by school. The district contracted with the SAS Corporation and moved to a more complex method of calculating teacher value-added using the Education Value-Added Assessment System (EVAAS). The system is based on a model developed by William Sanders and co-authors originally under the moniker "Tennessee Value-Added Assessment System" (Sanders, Saxton and Horn, 1997; Wright, Sanders and Rivers, 2006). For department-based awards, where a department is defined by school-grade-subject, the model estimates a department-grade-year fixed effect that accounts for prior teachers' or departments' contributions to achievement.[15] The current department fixed effect is captured and then adjusted via a Bayesian shrinkage estimator so that estimates for departments with fewer observations are biased towards the mean. This adjusted department fixed effect is the department value-added score. The value-added measures are then ranked within grade, subject and year.[16] Departments that receive value-added scores greater than zero (indicating value-added greater than the mean) and that are above the median value-added in their group receive an award. The award doubles if the department is within the top quartile of value-added.

Table 1 provides details on the awards available to teachers each year and the requirements for receiving them for high-school teachers who teach core courses - the focus of this study. As the table indicates, although a teacher would receive the awards for all grades in his subject

[15] This procedure amounts to including fixed effects for the teachers\departments each student had in each of the last three academic years.

[16] See Wright, White, Sanders and Horn (2010) for a detailed technical treatment of both methods.

regardless of whether he teaches each grade, each award is based on grade and subject specific performance. For example, a teacher may only teach 9^{th} grade students in science and hence her actions only can contribute to the 9^{th} grade portion of the science award. However, if her department meets the requirements for the 10^{th} grade science award, she will receive that award money as well. Our identification strategy relies on the fact that how much each teachers' actions contribute to the probability of award receipt differs by the share of students she teaches in each grade. Despite the fact that teachers may receive bonus money due the actions of teachers in other grades, the incentive system is designed such that each core teacher's own students enter into some award tournament. This setup means that every core high school teacher is incentivized, to some degree, to get over some award threshold.[17] And, the most salient measure of the specific incentives a teacher faces is the share of students in the group she teaches, as her impact on the likelihood of award receipt is directly proportional to this share.

In addition to the teacher award, there are a series of awards for school-wide performance.[18] Each of these awards is relatively small, ranging from \$150 to \$750 a piece, hence we do not consider them in our analyses.[19]

In 2006-07 and 2007-08, teachers could earn up to \$5,500 from the departmental awards.[20] The maximum total award a teacher could receive was \$8,030. In 2008-09, HISD increased award amounts substantially. The maximum award on the department portion jumped to \$7,700, with a total maximum award of \$11,330. Given that the base salary for a new teacher with a

[17] This design also could lead teachers within a department to act strategically across grades by reducing performance in earlier grades in order to increase growth in later grades. Due to accountability pressures, it is unlikely principals would allow such behavior to persist for very long. However, we have estimated models by grade to examine whether effects are indeed smaller in earlier grades. We find no statistically significant differences across grades, which suggests teachers are not engaging in this cross-grade gaming behavior. These results are provided in Appendix Table A-4.

[18] Each year there are four types of campus-wide awards for which teachers are eligible. Initially, these awards included a bonus for school-wide performance, an award for being in the top half of a state-wide comparison group of schools determined by the state education agency, an award for the school being given one of the two highest accountability ratings, and a writing performance award. In 2009-10, the second campus-wide award was disbanded and replaced with bonuses for participation in and performance on Advanced Placement and International Baccalaureate exams.

[19] Principals and assistant principals also were given awards for school-wide performance in each subject. The incentives under these awards were only partially aligned with those of teachers, as they were based on performance by the entire school in each subject, not by each grade and subject.

[20] This amount includes a 10% attendance bonus that is given to teachers who take no sick days during the year.

bachelor's degree in that year was \$44,027, up to 20% of a teacher's total wage compensation was determined by performance bonuses, with up to 14% from the department award portion. Even teachers at the highest step in the pay scale, \$71,960, could have received up to 14% of their salary from incentive pay. The average award across all core teachers in HISD (not only high school) was \$3,614 in 2009-10. Thus, the large bonus amounts relative to base pay suggests there is substantial scope in this system for teachers to respond to the incentives.

One potential problem with the ASPIRE program is that the use of the EVAAS value-added methodology for determining award receipt might make the award formula complex and difficult for teachers to understand. However, there is some evidence that teachers in HISD were well informed and had a good understanding of the system. In surveys conducted by the district, teachers were asked about their level of understanding of the program parameters.[21] Although the surveys had relatively low response rates (30% - 50%), those who responded generally indicated that they understood the program. For example, in May 2009, 90% of teachers indicated they had very high, high, or sufficient understanding of the program. Nonetheless, we note that teachers do not need to fully understand the value-added system in order to respond to the incentives we study in this paper. A sufficient condition for us to detect responses to student share incentives is that teachers understand that increasing their students' achievement on specific tests leads to an increase in value-added and that their students' contribution to the value-added score is proportional to the share of students they teach in the given subject. Since detailed documents that explain the value-added system are easily accessible to teachers online, we believe this condition is likely met and if anything, a lack of understanding would bias us towards not finding effects.

The survey responses also provide some insight into whether teachers responded to the incentives in the ASPIRE program. In May of 2009, teachers were asked a series of questions about whether they agree that the award program changed various aspects of their teaching. In each case, at least 47% of teachers responded that they changed a particular aspect. For example, 47% of teachers indicated they devoted more time to professional development, while 60% indicated they used value-added data to make instructional decisions.

[21] The survey results can be found at *http://www.houstonisd.org/portal/site/researchaccountability.*

5 Data

Our data come from matched student and teacher records that cover the 2002-03 through 2009-10 academic years. Since the department-level awards only are provided in high school, we restrict our analysis to grades 9 through 11 (students in grade 12 are not tested unless they fail the grade 11 exams). We further restrict the analysis sample to 2003-04 and after to allow us to control for prior achievement. The data include achievement results from two types of exams. The first is Texas' criterion referenced exam used for accountability called the "Texas Assessment of Knowledge and Skills" (TAKS). Unfortunately, we do not know whether a given seating of this exam is the first or a retake after failing the first exam. Nonetheless, since students often undergo intensive test preparation before retakes, a reasonable assumption is that a student's lowest score in a year is the initial score. Hence, we use each student's lowest score in a year as our achievement outcome for the TAKS exam. The scaled scores are then standardized within grade and year across the district to have a mean of zero and standard deviation of one. The second exam type is the Stanford Achievement Test (SAT), a nationally normed standardized exam. This exam is "low stakes," since it does not contribute to accountability or graduation requirements, although as we explain below it is used in English\language arts and 9^{th} grade science and social studies to determine awards. We also standardize the scaled scores on these exams within grade and year. In addition to the achievement tests, the data have information on student course taking, demographics and grades. Students are linked to teacher id's via course records. Teachers are matched to awards based on a list compiled in 2009 of courses that count for each award.[22]

Each observation in the data is for a student-course unit. As a result, some students who take multiple courses in a subject with either the same or different teachers will be observed multiple times. For example, a student might take a class on US history and a second class on world cultures with two different teachers, both of whom would be eligible for the social studies

[22] Course names were standardized across the district in 2006-07 and remained consistent afterwards. However, prior to 2006-07 some courses had different names. Additionally, some new courses were created and old courses discontinued. Generally, this is not a problem since the awards are only based off of core subjects – math, science, social studies, language arts, and reading – for which course offerings change little over time. Thus, we visually inspected courses that did not match directly to the list to determine whether they should be included as an award eligible course had the ASPIRE program existed at the time.

awards. In this case, the student's achievement only would count towards the value-added metric that determines awards once even though the student would appear in our data twice. In order to ensure that such students are not given excess influence on the estimates, in all of our regressions we assign weights to each observation equal to the inverse of the number of courses the student takes in a subject.[23]

The data are split into four subjects - math, English & language arts (ELA), science, and social studies. Teachers for each of these subjects are eligible for the departmental awards. While reading teachers also are eligible for awards, by high school few students take reading as most have moved on to English literature. Although reading and ELA are combined into a single award, students who take reading enter into the departmental value-added calculation based on reading scores, while students who take ELA courses enter based on language scores. Since very few students take reading in high school, estimates of impacts on reading achievement are very noisy. Hence, we do not provide results for reading. Note that this implies that only students who take an ELA course are included in our analysis of language scores.[24]

We assign teachers to students based on current academic year assignments for both spring and fall regardless of which test is used to determine awards. The TAKS exam is given in late March or early April, making the appropriate teachers for this exam the fall and spring teachers of the current school year. The Stanford exam is given in January, however, making the appropriate teacher assignment more ambiguous. We use the same assignment throughout for purposes of consistency as well as because, for the January exam, the spring semester teachers in academic year t can influence the score through test preparation, extra teaching sessions and review for the exams. On the other hand, the teacher from the spring of academic year $t-1$ may not be inclined to change her effort in response to the year t award, since many of her students in that term may not count towards her year t award due to school switching, dropouts, and students exiting 11^{th} grade. Since the best method for linking Stanford tests to teachers is not entirely clear, we provide robustness checks from models that link students to

[23] Results are similar without weighting and are provided in the Appendix Table A-4.

[24] Since reading scores contribute towards award determination, teacher shares for ELA teachers are calculated as the number of students that teacher has in ELA courses divided by the total number of students in ELA & reading courses in the grade.

the fall teacher of year t and the spring teacher from year $t-1$ as well as estimates that use only the fall semester teachers to identify award impacts. These results are provided in Appendix Table A-4 and show that our conclusions are robust to the specific manner in which we match teachers and students.

Since HISD had an individual award system for high school in 2005-06, we drop this year from our main analysis as it is unclear whether this should be considered a treatment or comparison year. Furthermore, we drop 2006-07 as awards during this year were based on school-wide value-added in a subject rather than grade-level value-added. Nonetheless, we will show later that including these years with 2005-06 as a "pre" year and 2006-07 as a "post" year has little impact on our estimates. We further limit the sample by dropping charter schools and alternative schools as the former tend to be very small and the latter serve special populations. In both cases, this makes these schools relatively incomparable to traditional high schools. We also drop observations for all teachers who instruct fewer than 10 students in a subject as these are likely to be part-time teachers who are ineligible for the awards. Finally, we exclude teachers for whom more than 80% of their students are limited English proficient or more than 80% are special education, because these classes tend to be small and specialized. In all of these cases, we estimate models without the restriction and find results - described in more detail below - that are similar to baseline. Our final sample includes approximately 240,000 student-course observations in 33 high schools with between 263 and 356 teachers in each subject per year.[25]

Table 2 provides summary statistics and exact observation counts for the data, split by subject. In general, student characteristics are similar regardless of the subject. This result is not surprising, as most students are required to take at least one course in math, science, social studies and English/language arts each year. Note that the smaller sample size for English is due to the exclusion of students in reading classes. HISD is a heavily minority district - only 11% of high school students are white. The racial composition is mainly a mix of Hispanic (54%) and black (31%) students. Students in HISD are also relatively low income, with 70%

[25] HISD's teacher identification system changed in 2006-07. Hence, we are not able to match all teachers across pre and post periods and as such we cannot identify how many teachers there are in total. However, from 2006-07 through 2009-10, the period in our data during which teachers had unique identification numbers that were inviolate over time, we have 745 math, 728 language arts, 695 science and 683 social studies teachers.

being economically disadvantaged.[26] Furthermore, 63% of students are classified as being at risk for dropping out, 7% of students in the sample have limited English proficiency and 17% of the sample is classified as gifted. While the gifted population may seem large, it is likely upward biased relative to the underlying population, as a substantial portion of the non-gifted students drop out prior to or during high school. In Panel [B] we see that, on average, teachers are responsible for between 12% to 14% of students in a subject-grade, and there are between 12 - 15 teachers in each grade and subject. Note that the mean share is not equal to the inverse of the number of teachers because teachers with rates of LEP or special education students over 80%, who generally have smaller classes, are dropped from the sample.

6 Empirical Methodology

Our theoretical model indicates that when a teacher is responsible for a small portion of a group, an increase in that responsibility as measured by the share of students the teacher instructs should generate increases in achievement. The goal of this analysis is to identify whether teachers who are responsible for a larger share of students increase test scores more post-ASPIRE than pre-ASPIRE. If students were randomly assigned to classrooms, we could simply compare teachers with higher and lower shares after program implementation. However, since students sort non-randomly into classrooms we need to control for underlying characteristics of students and teachers that might be correlated with their teachers' shares. We accomplish this task via a difference-in-differences specification.

We use administrative data from HISD on student test scores, student demographics and teacher assignments as described in Section 5 to estimate the following model:

$$\begin{aligned} A_{isgjt} = & \beta_0 + \beta_1 Share_{sgjt} + \beta_2 Share_{sgjt} * Post_t + \\ & \sum_t \sum_g \gamma_{gt} A^{pre}_{isgjt} \times Year_t \times Grade_g + X'_{it}\Phi + \lambda_{gt} + \nu_{jt} + \varepsilon_{isgjt}, \end{aligned} \tag{8}$$

where A_{sigjt} is test score in subject s of student i in grade g with teacher j in year t, $Share$

[26] Economically disadvantaged means that a student qualifies for free-lunch, reduced-price lunch, or some other Federal or state anti-poverty program.

is the proportion of students teacher j teaches in year t, grade g and subject s, $Post$ is a dummy variable equal to 1 if the incentive pay program is in effect (2006-07 and later), and A^{pre}_{isgjt} is lagged student test score. In order to avoid conditioning on scores that could have been influenced by ASPIRE, we condition on each student's 2004-05 achievement score for 2005-06 and later. For 2003-04 and 2004-05, we use once lagged achievement.[27] Since the role of our lagged achievement measure may change by year and grade level, we interact A^{pre}_{isgjt} with year-by-grade indicators. The vector X contains student demographic characteristics, such as race, gender, participation in special education, participation in gifted and talented programs, limited English proficiency, and whether the student is economically disadvantaged, along with a quartic in total enrollment in the school. In addition to these controls, equation (8) contains grade-by-year fixed effects (λ_{gt}) and school-by-year fixed effects (ν_{jt}). We estimate this model separately for math, English, science and social studies tests. Because of the likelihood that errors are correlated across students within schools and within schools over time, all estimates in the analysis are accompanied by standard errors that are clustered at the school level.[28]

The coefficient of interest in equation (8) is β_2, which shows how the effect of teacher share shifts when the incentive pay program is implemented. In order to interpret β_2 as a causal estimate, we must control for the non-random sorting of students into classes with different teacher shares. It is important to emphasize that we control for lagged student test scores. To the extent that these scores pick up fixed differences in student academic ability, any residual selection would have to be a function of student test score growth, not student test score levels. We also control for $Share$, which estimates the effect of teacher share on student test scores in the absence of ASPIRE. Students in classes in which the teachers teach a large proportion of students may perform better if the teacher is of higher quality or could perform worse if the large volume of students causes her to under-perform. The parameter β_1 picks up this

[27] Results are similar if we use 2002-03 achievement as the lagged score for all years and grades and are provided in Appendix Table A-4.

[28] Clustering standard errors may still cause one to over-reject null hypotheses when the number of clusters is small (Cameron, Gelbach and Miller, 2008; Bertrand, Duflo and Mullainathan, 2004). Using monte-carlo simulations, Bertrand, Duflo and Mullainathan (2004) show only very small over-rejection rates with 20 clusters and Cameron, Gelbach and Miller find similar results with 30 clusters. These simulations suggest that clustering our standard errors at the school level will not be problematic for the purposes of hypothesis testing, as we have 33 clusters.

underlying relationship between teacher share and student achievement and thus β_2 provides a difference-in-differences estimate.

Conditional on the school-by-year and grade-by-year fixed effects in addition to the other controls in the model, the identifying variation in $Share$ comes from several sources. The first is year-to-year differences in share within teachers over time. The share of students for whom a given teacher is responsible may vary from year to year due to population variation, idiosyncratic demand differences for specific subjects across cohorts, and teacher turnover. The variation in $Share$ in equation (8) also comes from differences in teacher share across teachers in a given subject and grade within each school.[29] Although it is not possible to know perfectly why different teachers are responsible for different proportions of students, for our identification strategy to be valid what must be true is that the reason for these differences does not change when the program is implemented.

Thus, in order for β_2 to provide an unbiased estimate of responses to stronger merit pay incentives, it must be the case that students with different test score growth patterns are not differentially sorting post-ASPIRE relative to pre-ASPIRE into classrooms with teachers who teach a larger (or smaller) share of students. Note that the value added methodology by which the awards are administered uses statistical adjustments based on multiple years of prior achievement. Thus, a principal would not be able to manipulate class assignment to maximize award receipt by simply sorting students based on raw achievement or growth rates. Further, it is important to stress that due to school accountability rules principals already had strong incentives to maximize total achievement prior to the implementation of ASPIRE. Nonetheless, if administrators did try to reallocate high performing students to high share teachers in response to the incentive program, we would expect that the information principals use to sort students is correlated with student demographics and achievement. If this is true, teacher share interacted with being in a post award period would be related to the demographic characteristics and prior achievement scores of students. In Table 3, we present balancing tests

[29] In Appendix Table A-2, we provide results from an analysis of variance for teacher share in 2006 and later. After accounting for observables and all fixed effects in our model, the results indicate that, depending on the subject, between 40% and 58% of the remaining variance in teacher share is across teachers while the rest is within teachers.

that show the correlation between our key explanatory variable and demographic characteristics of students. In particular, we estimate regressions of the following form:

$$x_{isgjt} = \alpha_0 + \alpha_1 Share_{sgjt} + \alpha_2 Share * Post_{sgjt} + \lambda_{gt} + \nu_{jt} + \varepsilon_{isgjt}, \tag{9}$$

where x is a specific student characteristic and all other variables are as previously defined. In Table 3, we show estimates of α_2 that test whether shifts in teacher share surrounding the implementation of the incentive pay program are correlated with shifts in student observable characteristics.

The estimates in Table 3 suggest there were no significant changes in the relationship between student demographics and teacher share when ASPIRE was implemented. We test whether there are "impacts" on gender, race, economic status, at-risk status, special education, LEP, and gifted and talented status. We also examine the "impact" of ASPIRE on pre-treatment achievement levels and gains (one-year growth in test scores). In no case are these estimates significant at the 5% level and only one, LEP status for science exams, is significant at even the 10% level. The one potentially troublesome estimate is for science achievement. While not statistically significant, it is nonetheless large and indicates that teachers with higher shares tend to get higher achieving students in science. While this result may give us some pause in the interpretation of the science results, it is nonetheless comforting that we see no similar estimates in any of the three other subjects, and in fact the math and English point estimates have negative signs. Further, we stress that we control for lagged achievement and other student observables in all of our models, which helps address the potential sorting in science.

Another identification concern is that, even if student sorting did not change as a function of share when ASPIRE was implemented, teacher shares could have adjusted in response to the awards. For example, a principal may decide that, in order to maximize award receipt, she will increase shares for good teachers while decreasing shares for low-performing teachers. While principals have very limited ability in HISD to fire teachers due to low value-added, this goal could be achieved by assigning teachers in core subjects to teach in non-core subjects instead or to teach lower-share core classes. Such re-assignment is likely to be difficult, however, as by high school most teachers specialize in specific subjects and have high levels of specific human capital in those subjects, which makes it costly for them to switch. Also, as mentioned above, due to accountability pressures, the principal already had an incentive to maximize group achievement by assigning the best teachers the highest shares before ASPIRE.[30]

Nonetheless, we can check on the empirical relevance of this theory by assessing whether there is a change after implementation of ASPIRE in how teachers of varying quality are assigned shares. We link teachers over time to calculate teacher value added for a subset of teachers. Unfortunately, while we have unique teacher id's for 2006-07 and later, prior to 2006-07 the teachers were not linked over time or as they changed schools by ID numbers. For these years, we have teacher names and gender but we were unable to acquire teacher name data (along with teacher characteristics) for the 2007-08 academic year. Further, these data from 2008-09 and 2009-10 are incomplete, with missing information for many teachers. Thus, in order to get a reasonable sample of teachers linked over time, we create a sample that links teachers by name and gender from 2002-03 through 2006-07 and then matches to their 2006-07 id number. Teachers from 2007-08 and later are then matched to earlier years via the id number, and we restrict the sample to those who are in the data in 2006-07. We note though that this method has two key limitations. First, we are left with a select sample of teachers who were in HISD in 2006-07. Second, matching by names leaves us with some teachers with the same name who will be grouped together and some teachers who change names, mostly due to marriage, who will be identified as two separate teachers.

With these caveats in mind, we calculate teacher value-added using data from prior to the implementation of ASPIRE. In particular, we estimate the following model for each subject, applying the weights described in Section 5:

$$Y_{isgjt} = \gamma_0 + \gamma_1 Y_{isgj,t-1} + X'_{it}\Phi + \lambda_{gt} + \nu_{jt} + \varepsilon_{isgjt}, \tag{10}$$

[30] Note that if principals re-organized shares across teachers in a way that increased aggregate test scores, this would be a positive causal effect of the program. However, it would be coming through altering teacher assignments rather than through increasing teacher effort.

For each grade, we use the standardized score on the exam that would eventually be used to determine ASPIRE awards. After estimating equation (10) using data from 2003-04 and 2004-05, we generate residuals for each student-course linkage and average over all (weighted) observations for each teacher. These average residuals are used as the "Teacher Value-Added" dependent variable in Table 3. The last row of Table 3 provides the estimates for the "impact" of teacher share after ASPIRE implementation on teacher value-added. For all four exams, the point estimates are small in magnitude and are not statistically different from zero at conventional levels. Further, there is no consistent pattern in the point estimates, with English being positive and the other three subjects negative.[31] Thus, we find no evidence that teachers with higher pre-ASPIRE performance were being given higher shares when the incentive pay program was implemented.

To develop more evidence on whether principals are altering teacher shares endogenously in response to ASPIRE, we compare teacher share distributions before and after ASPIRE is implemented. If there is a push to place more students with the high performing teachers in core subjects, we would expect to see a shift in the teacher share distribution towards having more teachers with large teacher shares. Figure 3 provides these distribution plots in each subject during the pre-ASPIRE (2003-04 to 2004-05) and post-ASPIRE (2007-08 to 2009-10) periods. In all four subjects, the distributions are very similar across time periods, with little evidence of any shift towards higher teacher shares. Hence, these results, combined with the teacher value-added results in Table 3, suggest that teacher assignments were unlikely to have been adjusted in a way systematically related to teacher share concurrent with program implementation. Furthermore, we show below that our results are similar if we instrument for share using pre-ASPIRE inverse department size, which cannot be influenced by endogenous sorting in response to the program.

[31] In Appendix Table A-1, we also provide estimates of the impacts of share on whether a student is new to the school or was not enrolled in the district in the prior year. In the former case only the math sample shows a significant effect at the 10% level, while only the science sample shows significant effects for the latter. We also look at whether the number of courses taught by a teacher is correlated with $Post * Share$ and only the English estimate is significant at even the 10% level, but the coefficient is positive. Having more courses requires more work on the part of teachers, and so without any effort adjustment achievement should be lower. Thus, we would expect that, if anything, this effect would generate a downward bias in the English estimates.

7 Results

7.1 Baseline Estimates

Before presenting our estimates of equation (8), we examine the correlation between teacher share and achievement by year in order to see whether there are pre-treatment trends and whether a break in any pre-treatment relationship between these variables is evident around 2006-07 when the group incentive pay system started. We estimate models similar to (8) except $Share$ and $Post * Share$ are replaced by interactions of $Share$ with year indicators. Note that while in our main regressions we omit 2005-06 and 2006-07, we include them here to better measure trends. Figure 4 presents estimates of the effect of a 10 percentage point change in teacher share by year, separately by exam. The estimates for math, shown in the first panel, are the most notable. Prior to 2006, teacher share was uncorrelated with student achievement, while after the incentive pay system was enacted teachers who were responsible for more students performed better than those responsible for fewer students. The estimates for English (second panel) also show a clear level shift after 2005. For science and social studies,[32] the year-by-year estimates after 2006 are more mixed. Nonetheless, the figures show that there is no trend in estimated effects of teacher share prior to implementation of ASPIRE, providing support for our difference-in-differences identification strategy. Indeed, F-tests of the joint significance of the pre-ASPIRE years (2003-04 through 2005-06) do not reject the null of equality, with test statistics of 0.3, 0.0, 0.0 and 0.0 for math, English, science and social studies, respectively. Thus, we find no evidence of pre-treatment trends in the share-achievement relationship prior to ASPIRE implementation. In particular, the figure indicates that any falsification test that uses pre-treatment data and involves setting the treatment year to 2005-2006 or before would show no change in the relationship between test scores and share when the false program was implemented. The figure also provides evidence that the ASPIRE program generated a positive shift in the relationship between teacher share and achievement, particularly for math.

Table 4 presents the baseline estimates of equation (8). The estimates in each column of

[32] We do not have data for performance on the state exam in social studies for 2006-07, so we omit that year from the social studies regressions.

each panel come from separate regressions, with controls added sequentially across panels. For brevity, only the estimates for *Share* (β_1) and *Post*Share* (β_2) are shown. Consistent with Figure 4, Table 4 shows that the effect of teacher share on math test scores increases after the incentive pay system went into effect. In the first panel, we include grade-year fixed effects and lagged achievement as controls. In Panel [2], we add student level controls. Then in Panel [3] we add school fixed effects. In Panel [4], we have our preferred specification that replaces school fixed effects with school-by-year fixed effects. The first four columns provide results for the exams that are linked to the incentives. Both math and social studies show similar results in all four specifications. In Panel [4], the math estimate is 0.24 and is significant at the 5% level. It indicates that a 10 percentage point increase in share increases average achievement amongst that teacher's students by 0.024 standard deviations. Similarly for social studies, the estimate is 0.20 and is significant at the 10% level. For English and science, the inclusion of school year fixed effects makes a notable difference, increasing the estimate for English from an insignificant 0.05 to a significant 0.14. For science the opposite occurs, as the school-year fixed effects drop the science estimate from 0.13 to essentially zero. With significant and positive impacts for math, English and social studies, the results indicate that teachers do respond to changes in the share of students for whom they are responsible in the direction predicted by the theoretical analysis in Section 3.

The results in Table 4 also help address whether the bonuses incentivize teachers to focus on specific tests or whether they lead to a general increase in knowledge.[33] We examine whether students in classrooms with teachers who have a higher share post-2006 score higher on the Stanford math exam, which is administered to all students but is not part of the incentive pay system. The last column of Table 4 presents estimates of equation (8) using standardized Stanford math scores as the dependent variable. Focusing on panel [3], the estimate is 0.22. While it is not statistically significantly different from zero at conventional levels, the estimate is close to the estimate for the state math exam. However, when we include school-year fixed

[33] Another possibility is that incentives encourage cheating. For example, Jacob and Levitt (2003) find non-trivial amounts of teacher cheating on standardized tests in Chicago in response to accountability incentives. See Barlevy and Neal (2012) for a discussion of the design of optimal teacher incentive mechanisms that avoid this problem.

effects, the estimate falls considerably, to 0.03. Hence, we find little evidence of impacts on this non-incentivized exam. We note though, that while this could be indicative of teachers focusing specifically on the incentivized exam, it is also the case that TAKS and Stanford do not fully overlap in topics studied. Because the curriculum is targeted towards TAKS, it may be that teachers focus on topics in the curriculum that are not well covered in the Stanford exam.[34] Indeed, our estimates show that Stanford math performance did not decline as a function of share post-ASPIRE, which suggests teachers were not completely shifting their focus to the incentivized exam. That the relationship between Stanford exam scores and share does not shift post-ASPIRE also provides support for our main identification assumption that principals did not sort students differentially into classrooms as a function of share post-ASPIRE. Such a change in sorting should show up on all test scores, not just on the incentivized exams.

In Panel [5], we provide a set of estimates that relies specifically on variation within teachers and years. The unique design of the ASPIRE program leaves many teachers with different incentives across grades, depending on the proportion of students they teach in each grade. For example, a teacher may instruct 50% of 9^{th} grade students but only 20% of 10^{th}. Thus, the teacher will face stronger 9^{th} grade incentives than 10^{th}. This setup provides a different source of identifying variation than our baseline difference-in-difference models, as it leverages differences within teacher in share in the same year by controlling for teacher-by-year fixed effects. Furthermore, these estimates are of interest to the extent that they show teachers shifting focus or effort across grades due to the financial incentives they face. The results in Panel [5] show estimates that are positive and significant for all four incentivized exams, with no impact on the non-incentivized Stanford math exam, mirroring the estimates in Panel [4].[35] The only estimate that is notably different from those in Panel [4] is science, which is now large, positive and significant. These results suggest that teachers do indeed shift focus across grades to the grade in which they have a higher share. They also provide further support for our

[34] Scores on the state math exam and Stanford math have a correlation in our data of only 0.63, which leaves substantial room for differences in outcomes across the two exams.

[35] In Appendix Table A-4, we also provide estimates that use teacher fixed effects and school fixed effects. With this analysis we have to use the restricted sample of teachers in HISD in 2006-07, so they may not be directly comparable to estimates in Table 4. Nonetheless, we obtain similar results with this model, which shows positive effects for all tests that are significant for TAKS math, science and Stanford math.

contention that our estimates are driven by teacher responses to ASPIRE, as it is difficult to tell an alternative story that would lead to within-teacher and year increases in the relationship between share and student achievement post-ASPIRE. For example, these estimates also are suggestive that our results are not being driven by increased resources being given to teachers with higher shares, as it is unlikely that principals can target resources in such a way that teachers can only use them in one grade.[36] Given the similarity of the estimates between Panels [4] and [5], we will use the model in Panel [4] throughout the rest of the analysis unless noted otherwise, as these estimates are considerably more precise.

7.2 Heterogenous Treatment Effects by Teacher Share

Thus far, we have estimated the mean effect of post-ASPIRE teacher share over the entire distribution of shares. These estimates may hide important information, however, as our theoretical model predicts larger effects at lower shares that fall to zero as share rises. To test for heterogeneous responses as a function of share, we estimate local linear regressions of the effect of teacher share post-2006 on achievement at different parts of the share distribution in Figure 5. This method allows us to examine non-parametrically how the effect of teacher share changes when the incentive pay system is implemented.[37] The figure shows point estimates and 95% confidence intervals from a series of regressions of equation (8) centered at each percentage point of the teacher share distribution and restricted to a bandwidth of 0.15 on each side using a rectangular (uniform) kernel. We show regression estimates up to a share of 0.5, as sample sizes become too small at larger shares for reasonable inference. Since 95% of the distribution has a share below 0.4, the standard errors tend to grow considerably at larger shares.[38]

Consistent with the predictions from our theoretical model, Figure 5 shows that the estimate

[36] We stress that if changes in resource targeting were a driver of the effects we find, our estimates still would be showing the causal effect of the incentive pay program on student achievement and how this effect varies with teacher share. For policy purposes, this is the relevant parameter. But, the interpretation of our estimates would differ: instead of being driven by changes in teacher effort, changes in resources also would play a role. While we believe our estimates are most consistent with effort changes by teachers as a function of share post-ASPIRE, our results are valid even in the presence of resource changes across the share distribution in response to ASPIRE.

[37] While there is parametric structure on the linear models we estimate, we impose no structure on the heterogeneity with respect to teacher share.

[38] In Appendix Figure A-1, we provide figures that use a bandwidth of 0.1 instead of 0.15. Although noisier, the basic pattern remains.

for $Share{*}Post$ starts out positive at low shares and then falls to zero for all four subjects.[39] In particular, for a teacher with a share close to zero, the impact on achievement from increasing share by 0.1 would be between 0.05 and 0.09 standard deviations. With the exception of language, all estimates are statistically significantly different from zero from a 0.0 share to a 0.2 share. The point estimates first cross the zero effect line between 0.2 and 0.3 in each subject, including ELA. Hence, Figure 5 shows that achievement increases substantially for teachers who are responsible for small shares of the class as that share increases; that is, the marginal impact of increasing share falls as the teacher's share increases. The effects at low shares are sizable, representing about half to a quarter of the effect of reducing class sizes by seven (Krueger, 1999) and are about the same size as a one standard deviation increase in teacher quality (Rivkin, Hanushek and Kain, 2005; Rockoff, 2004).

The estimates shown in Figure 5 do not lend themselves simply to statistical tests that the effect of share on test scores post-ASPIRE declines with share. Thus, in Appendix Table A-3 we estimate equation (8) separately for teachers with shares above and below 0.15 and then test for the equality of the *Post*Share* coefficients across the two models. As the table demonstrates, the effect of increasing share post-ASPIRE among teachers with shares less than 0.15 is much larger than among teachers with shares more than 0.15. For English, science and social studies, this difference is statistically significant at the 5% level, and for math, while insignificant, the p-value is only 0.11.

In Figure 6, we provide local linear regression estimates for the non-incentivized Stanford math exam. As expected, given the estimates in Table 4, there is no significant effect of share on Stanford math throughout the share distribution. In particular, in ranges where in Figure 5 we see significant effects for math, the estimate for Stanford math is virtually zero. These results indicate that there are no spillovers from improvements in TAKS into the Stanford test due to larger incentive strength. Nonetheless, it remains to be seen whether this is due to "teaching to the test" or because the Stanford exam does not cover the material on which teachers focus.

Although we model the teacher response as a function of the share of the students they teach,

[39] While there appears to be an uptick for math starting at around 0.3, the lack of precision at this range prevents us from being able to test whether this is a true effect. Indeed, except for a small range around 0.4, these estimates are not statistically significantly different from zero at the 5% level.

there also is a potential direct role for the department size. For example, if teachers monitor each other's performance, the number of teachers in each department should be directly related to teacher effectiveness.[40] In Figure 7, we provide local linear regressions of equation (8) with the addition of a variable for the number of teachers in the department ($DepartmentSize$) and the interaction of $DepartmentSize$ with being in the ASPIRE period ($Post * DepartmentSize$). The left column shows the impact estimates for teacher share while the right column shows impact estimates for department size. Department size only has an independent significant effect on language scores post-ASPIRE. Nonetheless, inclusion of these additional controls strengthens the share results. In particular, the estimates for language are now statistically significantly different from zero at shares between 0.0 and 0.2 and social studies effects stay positive at slightly higher levels of teacher share. Most importantly, however, is that the graphs show the same downward sloping relationship between the effect estimate and share as in Figure 5. These results indicate that teachers are responding directly to incentive strength rather than being influenced by other factors associated with different department sizes, which is a rather unique finding in this literature. Furthermore, these results show that teacher share is a stronger proxy for incentive strength than is department size, which is what has been used previously to examine group-size effects in education and in other labor markets. Our linked student-teacher data thus allow us to identify how teachers respond to group-based incentives with much more precision than has been feasible in previous analyses.

The estimates in Figures 5 and 7 are particularly important as they provides us with a measure of the achievement maximizing (optimal) group size. The results suggest that, assuming all teachers in a group have equal teacher shares, the optimal group size is between 3 and 5 teachers.[41] However, there are a few notes of caution here. First, the teachers in the groups we measure do not have the same shares, so if how students are allocated across teachers within the group matters for our calculations, then the estimates may not extend to the equal distribution

[40] If the incentive pay program leads to increased monitoring of higher-share teachers, then the *Post*Share* coefficients could be picking up monitoring as well. This would be one potential mechanism that would lead to higher effort among higher-share teachers post-ASPIRE.

[41] The 3 teacher group size is calculated using the share where the local linear estimates first crosses the zero effect line for social studies in Figure 7 (0.35), while the 5 teacher group size is calculated using the corresponding estimate for language in Figure 7 (0.20). All other estimates first cross the zero effect line between these two values.

case. Second, the groups in this study are all teachers in a grade who teach a given subject - the department. It is not clear that the achievement maximizing group size would remain the same if smaller groups are generated within departments. Nonetheless, these estimates do give us a starting point for thinking about optimal group size. To our knowledge, no other paper has been able to provide even rough estimates of such a parameter. Finally, and most importantly, even if these values do not precisely identify the optimal group size, the results in Figures 5 and 7 clearly show that if group sizes are large there are efficiency gains from reducing group size under a group incentive scheme. For example, these estimates suggest that school-level incentive pay programs may have little effect on teacher behavior because each teacher's school-level share is so low that the incentive she faces is quite weak.

7.3 Robustness Checks

As discussed in Section 6, the interpretation of the shift in the effect of teacher share after program implementation as causal is predicated on our extensive set of fixed effects and student background controls being sufficient to account for any changes in the underlying relationship between teacher share and achievement growth coincidental with ASPIRE implementation. In Table 5, we present a series of robustness checks that help shed light on the validity of this assumption.

First, we add school-grade fixed effects to the regressions, which has little impact on the estimates. We then control for the number of students each teacher teaches in Panel [2]. A teacher who has more students may be able to benefit from economies of scale in responding to the awards. Including this variable has a negligible effect on our estimates, however, suggesting that our results are not driven by economies of scale.

HISD has a number of charter and alternative schools. Teachers in these schools are eligible for the incentive pay awards, but we exclude them from our main analysis because of the difficult selection problems associated with these schools, given that teachers, administrators and students in these schools likely differ substantially from those in traditional public schools. When we include these schools, the estimates are attenuated for math, English and social studies, although they remain positive and statistically significant.

Throughout the analysis, we have excluded school years 2005-2006 and 2006-2007 because in those years the incentive pay system differed substantially from the subject-grade-specific tournaments of later years. As in the previous panel, when we add them back into the sample our estimates become attenuated - which is not surprising as we are essentially adding measurement error. Nonetheless, the estimates are qualitatively similar to baseline. In Panel [5], we relax our restriction on the minimum number of students teachers can have to be included in the regressions. The results change little. In Panel [6], we add back in teacher-courses with more than 80% Special Education or 80% LEP and find results that are in-line with those shown in Table 4. In Panels [7] and [8], we drop all special education and LEP students, respectively, and find results similar to baseline.[42]

Finally, we check the robustness of our estimates to potential sorting of students and teachers in response to the program by estimating difference-in-difference models using $\frac{1}{2004 DepartmentSize}$ to specify treatment intensity. This model tests whether departments that were larger (and thus had lower teacher shares) in the pre-ASPIRE period experienced larger increases in test scores when the program was implemented. Panel [A] of Appendix Table A-5 contains these estimates, and they are very similar to those shown in Table 4. While the estimates are less precise than in the baseline models, which underscores the value of using the share variation we have in our main estimates, the similarity of the point estimates in this model to the baseline model suggests limited scope for bias from endogenous student and teacher sorting in reaction to ASPIRE. In Panel [B], we use $\frac{1}{2004 DepartmentSize}$ and $Post * \frac{1}{2004 DepartmentSize}$ as instruments for $Share$ and $Post * Share$ and also find estimates that are, on the whole, similar to our main results. These estimates serve to complement the results shown above by providing evidence from an identification strategy that is not subject to the same potential biases from post-ASPIRE selection. Together with our baseline estimates, Table A-5 suggests teachers

[42] In the appendix we also estimate models that allow for different estimates by grade. In no case is an estimate in a given grade significantly different from other grades. We also estimate models that match students to their spring of the prior year and fall of the current year teachers instead of matching to spring and fall of current years. We further estimate models only matching student to their fall of current year teachers and, that use 2002-03 scores as the lagged score for all observations, and that are unweighted. All these models are similar to baseline. Finally we estimate models with teacher and school fixed-effects (but no school-year FE) and find similar estimates to baseline with the exception of Stanford math which is positive and significant in this model.

responded to the incentives they faced under ASPIRE and were more responsive when they were responsible for a larger proportion of the output.

8 Conclusion

Numerous school districts and states have implemented programs linking teacher compensation to student exam performance. Despite their widespread popularity, the evidence on the effectiveness of these programs is mixed. Particularly troublesome is that recent experimental analyses have found little impact of incentive pay on achievement (Fryer, 2011; Goodman and Turner, 2011; Springer,et al., 2010). One potential explanation for these findings is that these programs are not designed in a way that induces teachers to respond to the incentives. Unfortunately, while Barlevy and Neal (2012) provide useful theoretical analyses, there is a severe lack of empirical analysis into the optimal design of such programs. This paper takes a first step in understanding the role of program design in the development of incentive pay programs by testing for individual teachers' responses to incentive strength in a group-based teacher incentive pay program in the Houston Independent School District. The program we study, called ASPIRE, provides a unique opportunity to examine how teachers respond to free-riding, award salience and peer-monitoring incentives embedded within the program, since high school teachers are provided cash awards based on the performance of all students in the teacher's grade-school-subject cell. The cash awards are large, accounting for up to 14% of a teacher's total wage compensation. This is a useful program for studying teachers' responses to incentive strength since, unlike in cases where awards are determined on a school-wide basis, there is substantial variation in the share of students within a grade-subject that a teacher instructs. This "share" value is directly related to incentive strength because, as the share increases, the potential impact of teachers on award receipt increases as well. We use this teacher share as a proxy for incentive strength and evaluate difference-in-difference models that estimate the shift in the relationship between achievement and teacher share when the teacher incentive pay program is implemented. In addition to informing optimal program design, our analysis establishes whether teachers respond to the incentives at all. This is important to study as evaluations of overall programs cannot distinguish between whether the specific program is poorly designed or the more general problem that teachers simply may not respond to incentive pay.

Our study establishes that teachers do indeed respond to incentives when they are strong enough. In particular, we find evidence that student achievement increases in response to stronger group incentives, which we interpret as coming from increases in teacher effort. That is, teachers' effort increases as their contribution to the probability of award receipt increases. On average, our preferred estimates indicate that a 10 percentage point increase in teacher share increases math and social studies achievement by 0.02 standard deviations, while language scores increase by 0.014 standard deviations. There is no effect on science scores. However, these pooled estimates hide a substantial amount of heterogeneity. Using local linear regression techniques we find that, at very low levels of teacher share, math, language, science, and social studies achievement increases by 0.05 to 0.09 standard deviations for each 10 percentage point increase in teacher share post-ASPIRE. This treatment effect fades out as teacher share increases and reaches zero at teacher shares between 0.2 and 0.3. These results are consistent with our theoretical model and are indicative of substantial free-riding or response to award salience when teachers are responsible for small portions of the relevant student population. Furthermore, the results provide a basis for estimating achievement maximizing group size. If students are distributed equally amongst teachers in a group, and assuming there are no effects from splitting teachers within a department into separate groups, the estimates indicate that there are benefits to reducing group size until the teachers are in groups of 3 to 5.

More importantly however, is that our analysis suggests that the design of group teacher incentives has important implications for productivity. In particular, the results indicate that when implementing group incentive pay it is better to provide awards on the basis of small groups and that there is substantial potential for schools with group awards to improve productivity by reducing group size. This is an important finding because most group incentive pay schemes use the school as the group level. Our results suggest that a group that large, and even groups based on all teachers in a grade-subject, are likely less effective than smaller groups with a handful of teachers.

References

[1] Ahn, Thomas, 2011. "The Missing Link: Estimating the Impact of Incentives on Effort and Effort on Production Using Teacher Accountability Legislation." University of Kentucky, mimeo.

[2] Barlevy, Gadi and Derek Neal, 2012. "Pay for Percentile." *American Economic Review* 102(5): 1805-31.

[3] Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan, 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119(1): 249-275.

[4] Cameron, Colin A., Jonah B. Gelbach and Douglas L. Miller, 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90(3): 414-427.

[5] Fryer, Roland G., 2011. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." NBER Working Paper No. 16850.

[6] —, Steven D. Levitt, John List and Sally Sadoff. 2012. "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." NBER Working Paper No. 18237.

[7] Gaynor, Martin and Mark V. Pauly, 1990. "Compensation and Productive Efficiency in Partnerships: Evidence from Medical Groups Practice." *Journal of Political Economy* 98(3): 544-573.

[8] Glewwe, Paul, Nauman Ilias, and Michael Kremer, 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2(3): 205-227.

[9] Goodman, Sarena F. and Lesley J. Turner, 2011. "Teacher Inventive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." Columbia University, mimeo.

[10] Hamilton, Barton H., Jack A. Nickerson, Hideo Owan, 2003. "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation." *Journal of Political Economy* 111(3): 465-497.

[11] Holmstrom, Bengt, 1982. "Moral Hazard in Teams." *The Bell Journal of Economics* 13(2): 324-340.

[12] Jackson, C. Kirabo, 2010. "A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program." *Journal of Human Resources* 45(3): 591-639.

[13] Jackson, C. Kirabo, 2012. "Do College-Prep Programs Improve Long-Term Outcomes?" National Bureau of Economic Research Working Paper No. 17859.

[14] Jacob, Brian and Steven Levitt, 2003. "Rotten Apples: An Investigation of The Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118(3): 843-877.

[15] Kandel, Eugene and Edward P. Lazear, 1992. "Peer Pressure and Partnerships." *Journal of Political Economy* 100(4): 801-817.

[16] Krueger, Alan B, 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2): 497-532.

[17] Ladd, Helen F., 1999. "The Dallas School Accountability and Incentive Program: an Evaluation of its Impacts on Student Outcomes." *Economics of Education Review* 18(1): 1-16.

[18] Lavy, Victor, 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." *Journal of Political Economy* 110(6): 1286-1317.

[19] Lavy, Victor, 2009. "Performance Pay and Teachers' Effort, Productivity and Grading Ethics." *American Economic Review* 99(5): 1979-2021.

[20] Lazear, Edward P. and Sherwin Rosen, 1981. "Rank-Order Tournaments as Optimal Labor Contracts." *Journal of Political Economy* 89(5): 841-864.

[21] Leibowitz, Arleen and Robert Tollison, 1980. “Free Riding, Shirking and Team Production in Legal Partnerships.” *Economic Inquiry* 18: 380-394.

[22] Mas, Alexandre and Enrico Moretti, 2009. “Peers at Work.” *American Economic Review* 99(1): 112-145.

[23] Muralidharan, Karthik and Venkatesh Sundararaman, 2011. “Teacher Performance Pay: Experimental Evidence from India.” *Journal of Political Economy* 119(1): 39-77.

[24] Neal, Derek, 2011. “The Design of Performance Pay in Education” in Eric A. Hanushek, Stephen Machin and Ludger Woessmann (Eds.) *Handbook of the Economics of Education, vol. 4*. North-Holland: Amsterdam.

[25] Newhouse, Joseph P., 1973. “The Economics of Group Practice.” *Journal of Human Resources* 8(1): 37-56.

[26] Prendergast, Candice, 1999. “The Provision of Incentives in Firms.” *Journal of Economic Literature* 37(1): 7-63.

[27] Rivkin, Steven G., Eric A. Hanushek and John F. Kain. “Teachers, Schools and Academic Achievement.” *Econometrica* 73(2): 417-458.

[28] Rockoff, Jonah, 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review* 94(2): 247-252.

[29] Sanders, William L., Arnold M. Saxton and Sandra P. Horn, 1997. “The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment.” In *Grading Teachers, Grading Schools*, J. Millman, ed.: 137-162.

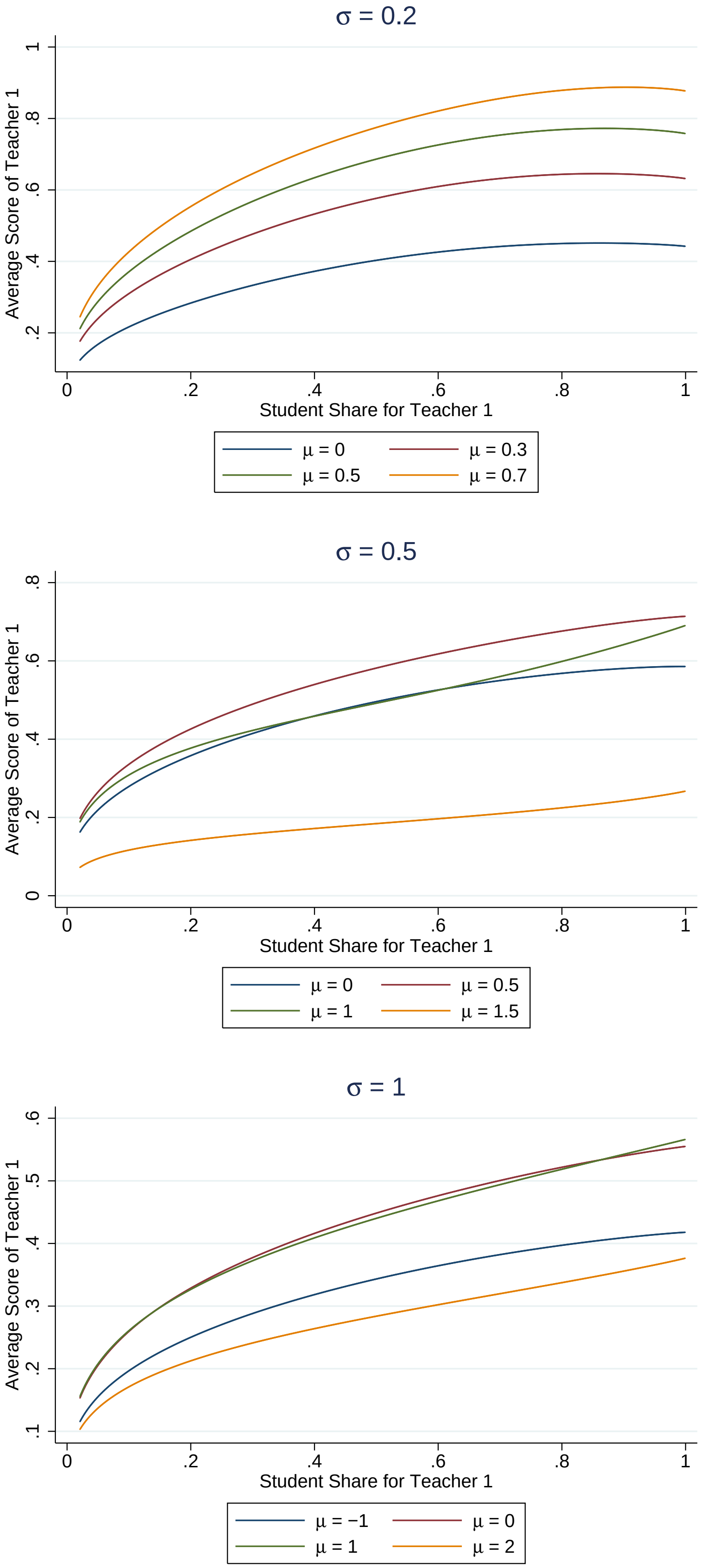
[30] Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper and Brian M. Stecher, 2010. “Teacher Pay For Performance: Experimental Evidence from the Project on Incentives in Teaching.” National Center on Performance Incentives: http://www.performanceincentives.org/data/files/pages/POINT%20REPORT_9.21.10.pdf.

[31] Sojourner, Aaron, Kristine West and Elton Mykerezi, 2011. “When Does Teacher Incentive Pay Raise Student Achievement? Evidence from Minnesota’s Q-Comp Program.” Mimeo.

[32] Wright, S. Paul, William L. Sanders and June C. Rivers, 2006. “Measurement of Academic Growth of Individual students toward Variable and Meaningful Academic Standards.” In *Longitudinal and Value Added Models of Student Performance*, R. W. Lissitz, ed.: 385-406.

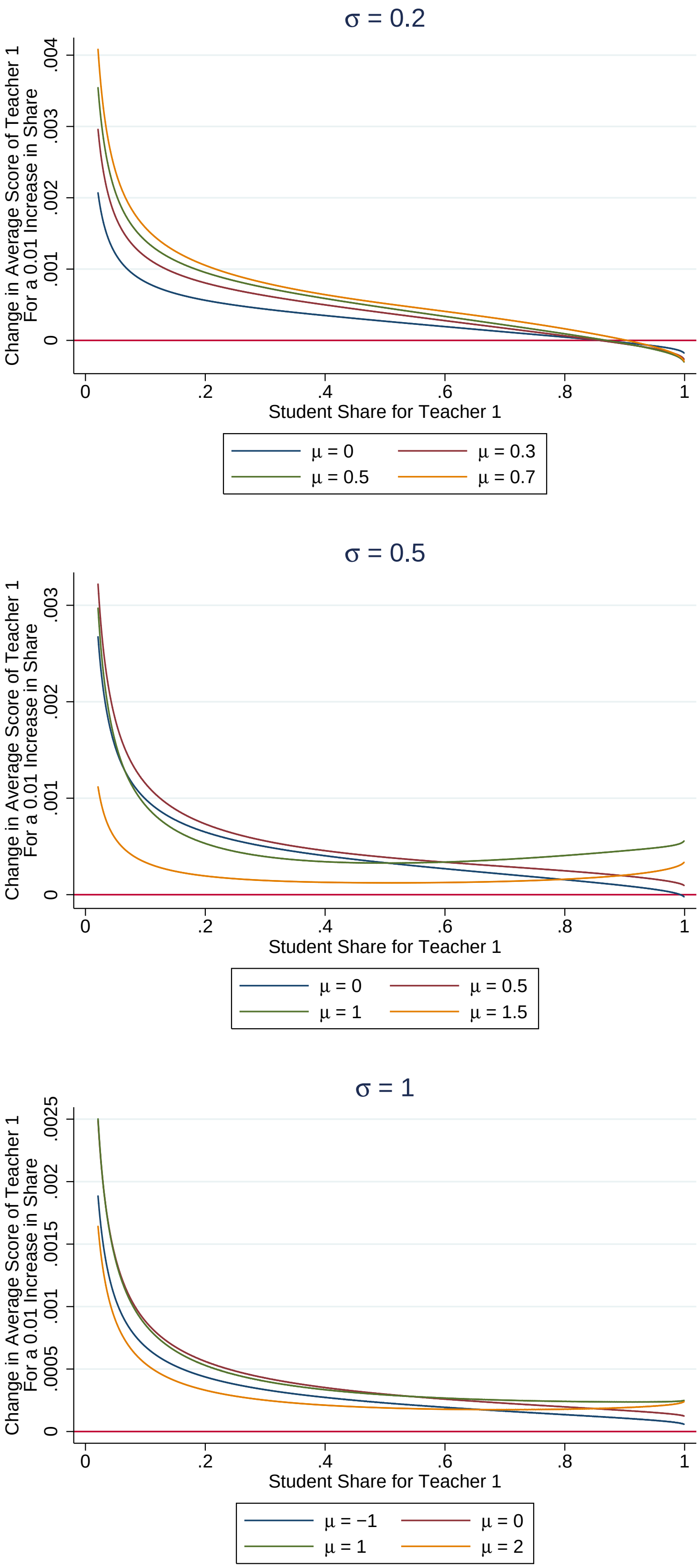
[33] Wright, S. Paul, John T. White, William L. Sanders, and June C. Rivers, 2010. “SAS EVAAS Statistical Models.” Technical report. Available at http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf.

Figure 1: Equilibrium Achievement Under Various Teacher Shares in a Two-Person Group-Based Rank-Order Tournament



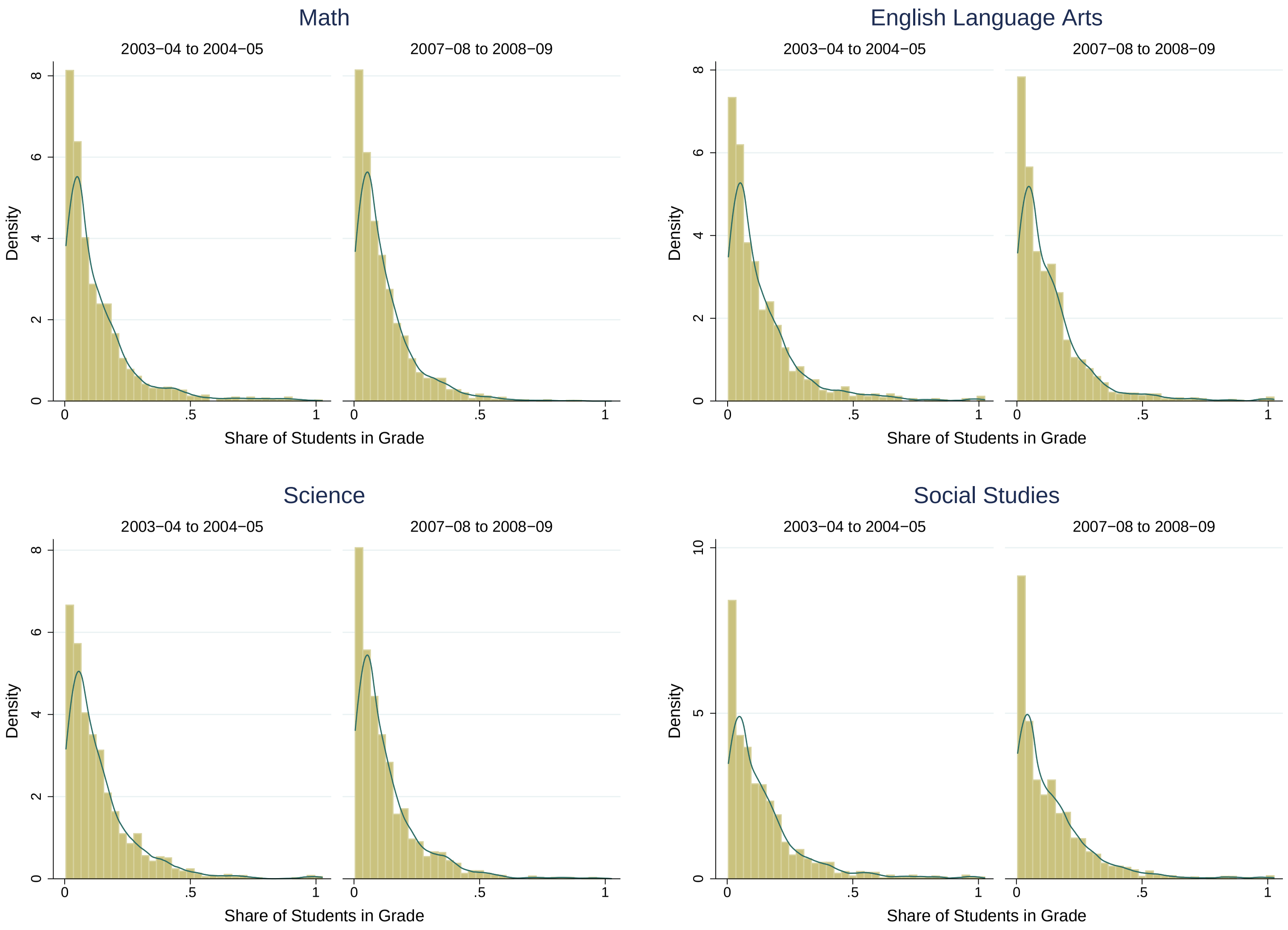
The graphs show the relationship between teacher effort and share (for teacher 1), assuming $S = \sqrt{e}$, $C = \frac{1}{2}e^2$, A=$1,500 and that the score cutoff function is distributed normally. The different panels show how this relationship varies with the standard deviation and mean of the cutoff distribution.

Figure 2: The Effect of Changing Share on Equilibrium Achievement in a Two-Person Group-Based Rank-Order Tournament



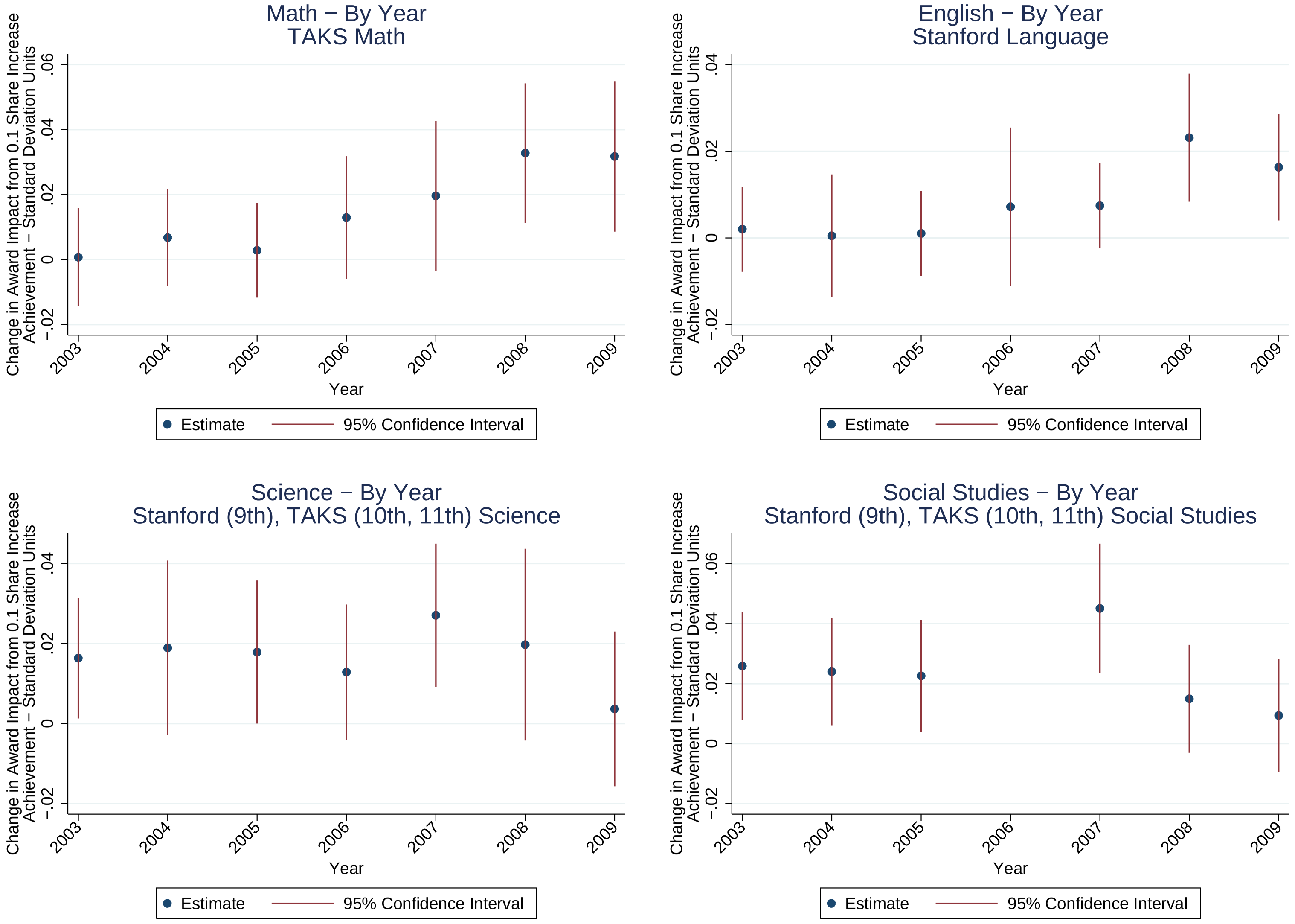
The graphs show derivatives of the curves in Figure 1, calculated empirically for each 0.001 change in share. The different panels show how this derivative varies with the standard deviation and mean of the cutoff distribution.

Figure 3: Distribution of Teacher Shares During Pre- and Post-Incentive Pay Periods



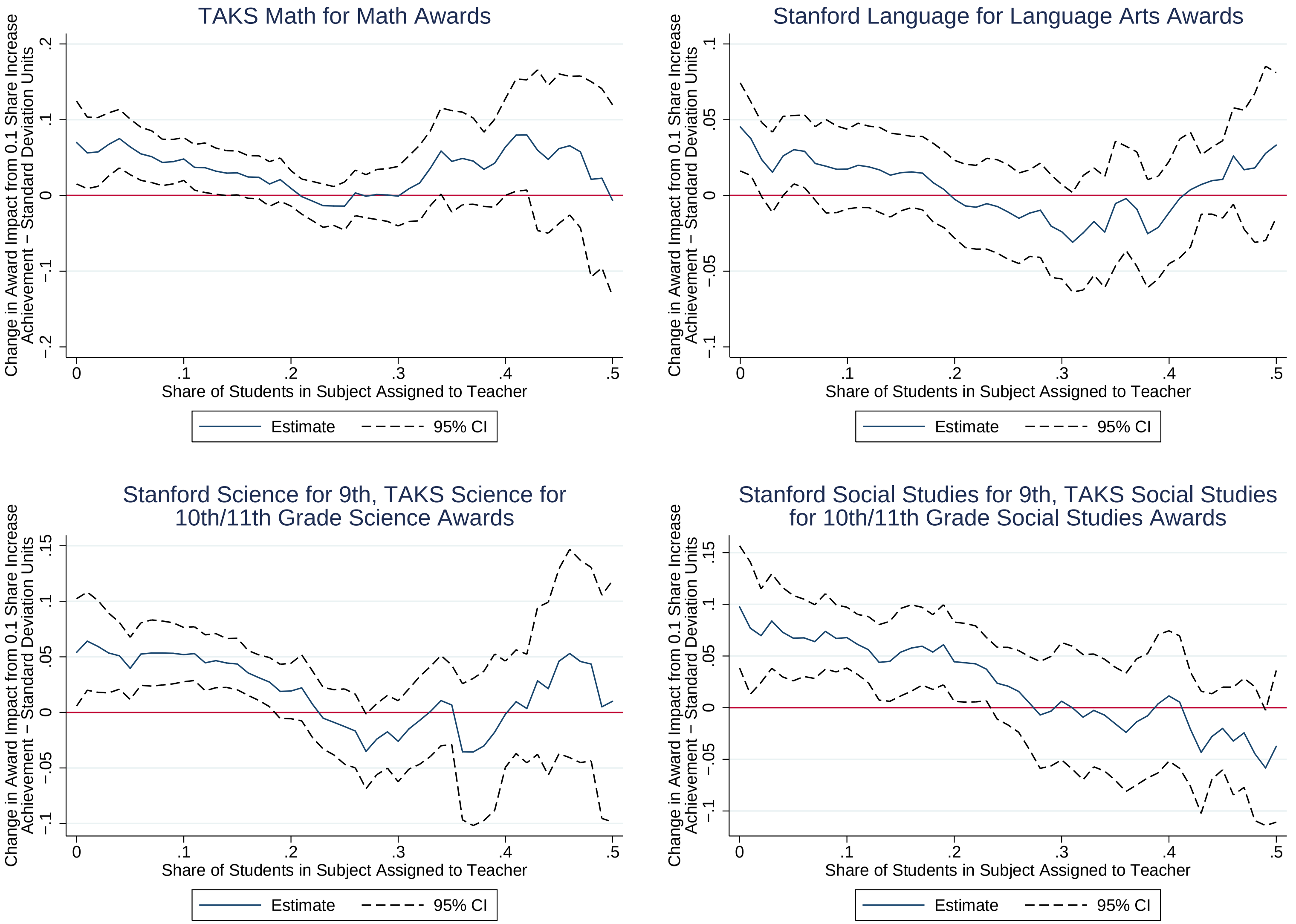
Graphs show distribution of unweighted teacher shares of students. The teacher is the unit of observation. Teachers with fewer than 10 students in a subject are dropped.

Figure 4: Effects of Teacher Share by Year



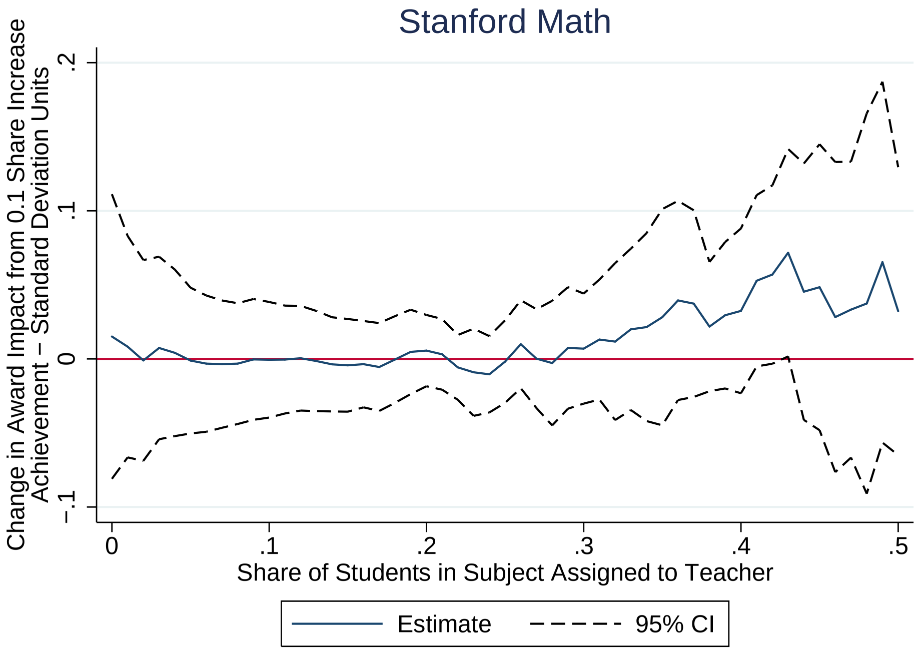
Data for social studies in 2006-07 are unavailable. Each point shows the average effect in a given year of raising the proportion of students a teacher is responsible for by 0.1 on standardized student test scores. These estimates come from models that include school-year and grade-year fixed effects as well as controls for school-grade-specific enrollment, lagged student test scores and student demographics. The bars extending from each point show the 95% confidence interval of each estimate that is calculated from standard errors clustered at the school level.

Figure 5: Local Linear Regressions of the Effect of Teacher Share Post-ASPIRE on Student Achievement



Each line shows local linear regression estimates of *Share*Post* from models that include school-year and grade-year fixed effects as well as controls for school-grade-specific enrollment, lagged student test scores and student demographics. Rectangular kernels are used with a bandwidth of 0.15. The dashed lines show the bounds of the 95% confidence interval that are calculated from standard errors that are clustered at the school level.

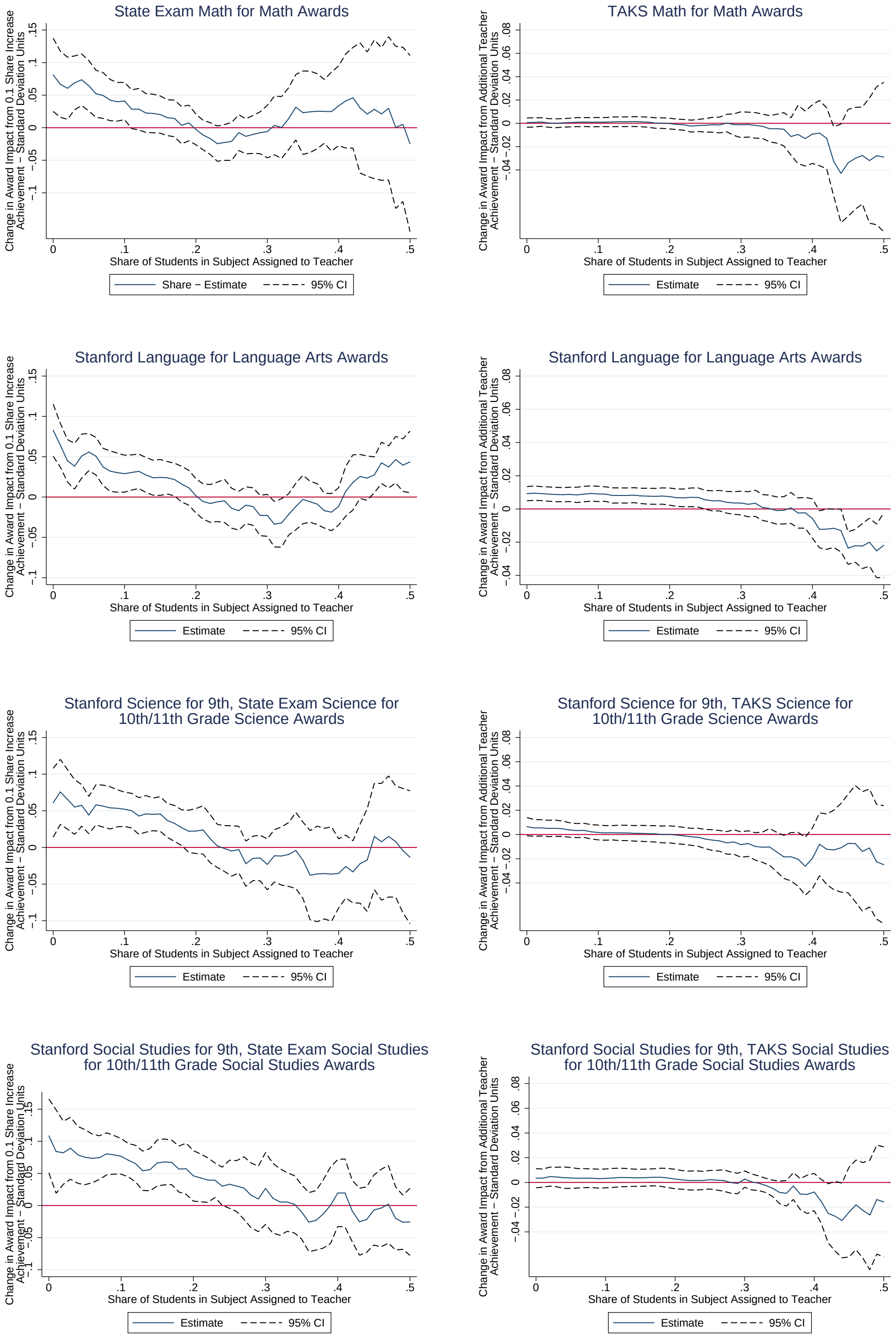
Figure 6: Local Linear Regressions of the Effect of Teacher Share Post-ASPIRE on the Non-Incentivized Math Exam



Each line shows local linear regression estimates of *Share*Post* from models that include school-year and grade-year fixed effects as well as controls for school-grade-specific enrollment, lagged student test scores and student demographics. Rectangular kernels are used with a bandwidth of 0.15. The dashed lines show the bounds of the 95% confidence interval that are calculated from standard errors that are clustered at the school level.

Figure 7: Local Linear Regressions of the Effect of Teacher Share Post-ASPIRE on Student Achievement, Controlling for Department Size

$Post * Share$ $\qquad$ $Post * DepartmentSize$



Each solid line shows local linear regression estimates of *Share*Post* or *Department Size*Post* from models that include school-year and grade-year fixed effects as well as controls for school-grade-specific enrollment, lagged student test scores and student demographics. Each row of figures comes from a separate regression. Rectangular kernels are used with a bandwidth of 0.15. The dashed lines show the bounds of the 95% confidence interval that are calculated from standard errors that are clustered at the school level.

Table 1: **Characteristics of Department Award Portion of the HISD Teacher Incentive Pay Program for 9th to 12th Grade Teachers**

Year	Description	Per-Subject Award For Being in Top 50%	Per-Subject Award For Being in Top 25%	Max Award (with 10% Attendance Bonus)
2006-2007	Separate award for each subject. Determined by department-wide value-added. Must have value-added > 0 to receive award. Compared to departments in same subject in all high schools.	One subject taught: $2500 Two subjects taught: $1250 Three subjects taught: $833	One subject taught: $5000 Two subjects taught: $2500 Three subjects taught: $1666	$5500
2007-2008	Separate award for each subject. Determined by department-wide value-added within grade. Must have value-added > 0 to receive award. Compared to departments in same subject and grade in all high schools (grades 9 - 11 only). All teachers in department receive award regardless of which grades they teach.	One subject taught: $833 per grade Two subjects taught: $417 per grade	One subject taught: $1667 per grade Two subjects taught: $833 per grade	$5500
2008-2009	Separate award for each subject. Determined by department-wide value-added within grade. Must have value-added > 0 to receive award. Compared to departments in same subject and grade in all high schools (grades 9 - 11 only). All teachers in department receive award regardless of which grades they teach.	One subject taught: $1167 per grade Two subjects taught: $833 per grade	One subject taught: $2333 per grade Two subjects taught: $1167 per grade	$7700
2009-2010	Separate award for each subject. Determined by department-wide value-added within grade. Must have value-added > 0 to receive award. Compared to departments in same subject and grade in all high schools (grades 9 - 11 only). All teachers in department receive award regardless of which grades they teach.	One subject taught: $1167 per grade Two subjects taught: $833 per grade	One subject taught: $2333 per grade Two subjects taught: $1167 per grade	$7700

Table 2: Descriptive Statistics

Panel [A]: Student Characteristics:

Variable	Math Students	English Students	Science Students	Social Studies Students
Asian	0.04 (0.20)	0.04 (0.20)	0.04 (0.19)	0.04 (0.19)
Black	0.29 (0.46)	0.31 (0.46)	0.31 (0.46)	0.31 (0.46)
Hispanic	0.55 (0.50)	0.53 (0.50)	0.54 (0.50)	0.54 (0.50)
White	0.11 (0.32)	0.12 (0.32)	0.11 (0.31)	0.11 (0.31)
Economically Disadvantaged	0.70 (0.46)	0.69 (0.46)	0.70 (0.46)	0.70 (0.46)
At Risk	0.62 (0.48)	0.61 (0.49)	0.63 (0.48)	0.63 (0.48)
Special Education	0.05 (0.22)	0.07 (0.25)	0.09 (0.28)	0.09 (0.28)
Limited English Proficiency	0.07 (0.17)	0.03 (0.18)	0.07 (0.26)	0.07 (0.26)
Gifted & Talented	0.17 (0.38)	0.17 (0.38)	0.16 (0.37)	0.16 (0.37)
Observations	241,694	230,099	240,572	243,161

Panel [B]: Teacher Characteristics:

Variable	Math Teachers	English Teachers	Science Teachers	Social Studies Teacher
Teacher Share	0.12 (0.13)	0.13 (0.14)	0.13 (0.13)	0.14 (0.15)
Department Size	13.6 (6.8)	15.4 (7.7)	11.9 (5.3)	12.2 (5.6)
Observations	3,518	2,902	3,281	3,053

Source: HISD administrative data from 2003-2009. Standard deviations are shown in parentheses below the means.

Table 3: OLS Estimates of the Relationship Between Student Background Characteristics and a Teacher's Share Post-ASPIRE

	Test Subject:			
		English &		Social
Independent Variable	Math	Language	Science	Studies
Female	0.020	-0.046	0.010	0.034
	(0.038)	(0.064)	(0.044)	(0.039)
White	0.084	0.034	0.009	0.040
	(0.051)	(0.039)	(0.030)	(0.051)
Black	-0.035	-0.039	0.009	0.001
	(0.051)	(0.043)	(0.047)	(0.039)
Hispanic	-0.056	0.027	-0.036	-0.044
	(0.039)	(0.046)	(0.042)	(0.043)
Economically Disadvantaged	0.010	-0.006	-0.007	-0.002
	(0.048)	(0.054)	(0.031)	(0.058)
At Risk	0.005	-0.028	-0.111	-0.085
	(0.106)	(0.123)	(0.086)	(0.096)
Special Education	-0.014	0.025	0.018	-0.006
	(0.020)	(0.032)	(0.038)	(0.023)
Limited English Proficiency	0.032	0.017	0.022	0.056*
	(0.036)	(0.026)	(0.027)	(0.029)
Gifted & Talented	-0.007	0.090	0.053	0.135
	(0.105)	(0.092)	(0.073)	(0.092)
Achievement Levels†	-0.091	-0.090	0.400	0.212
	(0.178)	(0.229)	(0.237)	(0.195)
Observations	241,694	224,044	240,472	242,001
Achievement Value Added††	-0.105	0.079	0.135	0.081
	(0.108)	(0.094)	(0.117)	(0.100)
Observations	224,167	205,995	219,566	220,010
Pre-ASPIRE Teacher Value-Added†††	-0.007	0.008	-0.060	-0.053
	(0.061)	(0.075)	(0.058)	(0.078)
Observations	127,161	124,929	123,975	138,627

† We use the most recent pre-program (2004 and earlier) lagged achievement. For math and English in the exam that determines the awards (TAKS and Stanford, respectively.) For science and social studies the TAKS exam is not given in every grade, so we use Stanford.

†† Value-added regressions include the the most recent lagged achievement from 2003 and earlier as a regressor interacted with indicators for current grade and year.

† † † Teacher value-added is calculated using 2004-05 and 2005-06 data for the subset of teachers in HISD in 2006-07 as these are the only teachers we can link across both pre and post-incentive pay periods. Value-added is calculated as the mean residual for each teacher from a regression of student achievement in the relevant exam for each subject on lagged achievement, gender, ethnicity, economic disadvantaged, at-risk, special education , LEP, gifted, and grade-by-year and school fixed effects.

Notes: Each cell comes from a separate estimation of equation (9) and shows the estimate of α_2, which is the coefficient on $Share * Post$. Regressions also include school-year and grade-year fixed effects along with a quartic in enrollment. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

Table 4: OLS Estimates of the Effect of a Teacher's Share Post-ASPIRE on Student Test Scores

	Test Subject:				
	TAKS	English &		Social	Stanford
Independent Variable	Math	Language	Science	Studies	Math
Panel [1]: Grade-Year Fixed Effects & Lagged Achievement:					
Post*Teacher Share	0.201	0.053	0.121	0.152*	0.150
	(0.128)	(0.088)	(0.079)	(0.086)	(0.163)
Teacher Share	-0.136	0.334**	0.123	0.432***	-0.390***
	(0.120)	(0.148)	(0.144)	(0.154)	(0.133)
Panel [2]: [1] + Individual Controls:					
Post*Teacher Share	0.177*	0.033	0.152**	0.164***	0.144
	(0.097)	(0.075)	(0.061)	(0.060)	(0.138)
Teacher Share	-0.025	0.075	0.126	0.218*	-0.305**
	(0.093)	(0.104)	(0.107)	(0.117)	(0.117)
Panel [3]: [2] + School Fixed Effects:					
Post*Teacher Share	0.215**	0.051	0.131*	0.166**	0.223
	(0.099)	(0.068)	(0.070)	(0.064)	(0.137)
Teacher Share	0.034	0.073	0.099	0.268***	-0.270***
	(0.058)	(0.062)	(0.067)	(0.076)	(0.089)
Panel [4]: [2] + School-Year Fixed Effects:					
Post*Teacher Share	0.238**	0.142***	-0.010	0.200*	0.032
	(0.089)	(0.049)	(0.092)	(0.104)	(0.083)
Teacher Share	0.047	0.004	0.184**	0.268***	-0.154**
	(0.061)	(0.047)	(0.078)	(0.075)	(0.065)
Panel [5]: [2] + Teacher-Year Fixed Effects:					
Post*Teacher Share	0.289**	0.187**	0.363***	0.364**	0.017
	(0.137)	(0.087)	(0.105)	(0.151)	(0.079)
Teacher Share	0.226**	0.024	0.236***	0.478***	-0.106*
	(0.091)	(0.058)	(0.071)	(0.073)	(0.079)
Observations	241,694	224,044	240,472	242,001	239,350

Source: HISD administrative data as described in the text. The math test in the first column is the state administered TAKS math exam. The Language exams are Stanford tests. For 10th and 11th grade science and social studies, the TAKS exams are used, while for 9th grade they are Stanford tests. All estimates are in terms of scale scores standardized across the district within grade and year. Individual controls include student gender, race, at-risk, special education, LEP, and gifted status. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

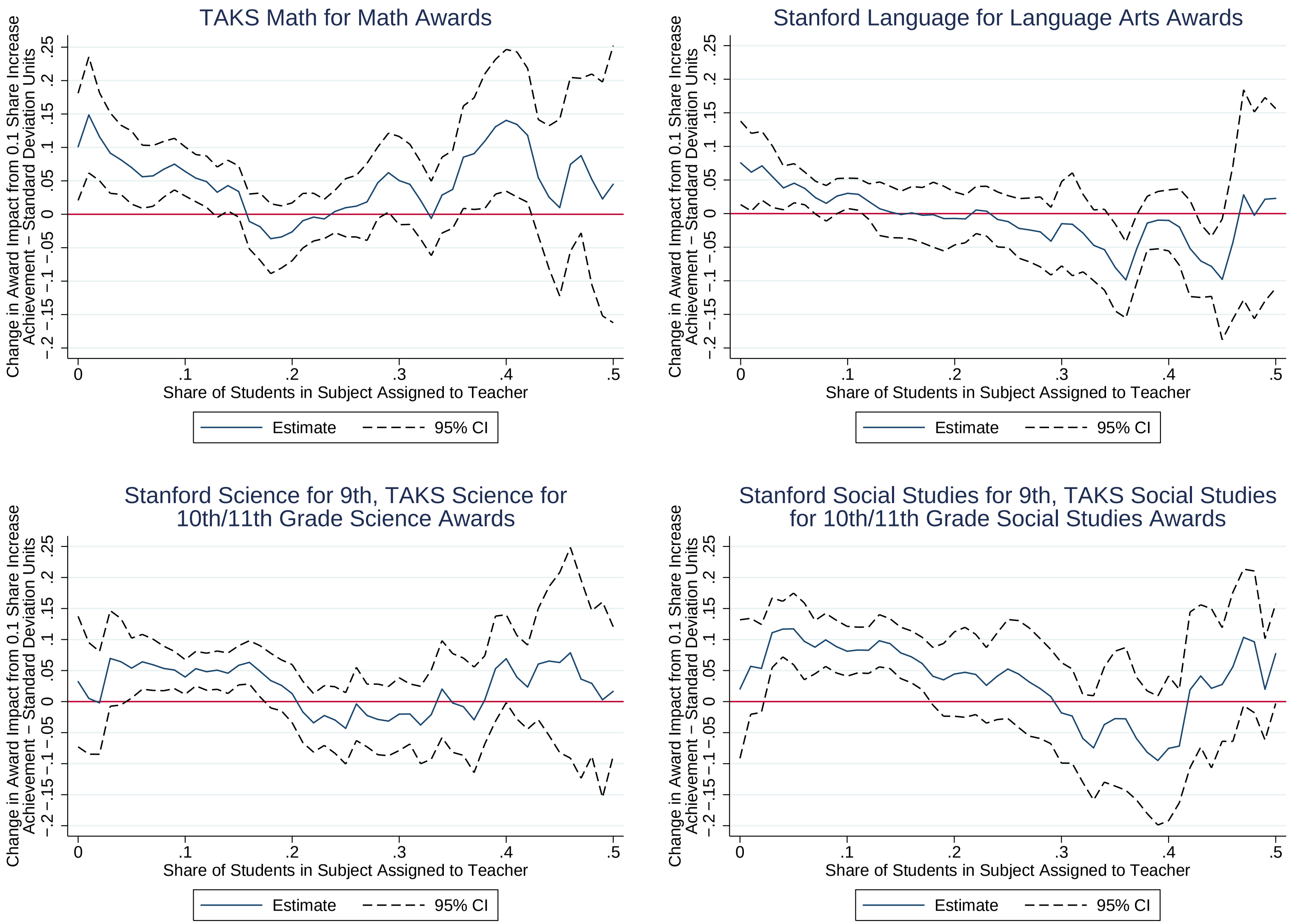
Table 5: Robustness Checks

	Test Subject:				
	TAKS	English &		Social	Stanford
Independent Variable	Math	Language	Science	Studies	Math
[1] Add School-Grade Fixed Effects:					
Post*Teacher Share	0.221**	0.161***	0.091	0.192**	0.033
	(0.089)	(0.054)	(0.084)	(0.091)	(0.080)
Observations	241,694	224,044	240,472	242,001	239,350
[2] Control for # of Students Each Teacher Has:					
Post*Teacher Share	0.236**	0.117**	-0.010	0.179*	0.036
	(0.087)	(0.050)	(0.092)	(0.103)	(0.081)
Observations	241,694	224,044	240,472	242,001	239,350
[3] Include Charters and Alternative Schools:					
Post*Teacher Share	0.144*	0.110**	-0.054	0.169*	0.003
	(0.083)	(0.046)	(0.080)	(0.095)	(0.077)
Observations	254,141	235,200	252,064	252,582	251,198
[4] Include 2005-06 as Pre and 2006-07 as Post Years:					
Post*Teacher Share	0.197**	0.051	-0.043	0.131*	0.101*
	(0.071)	(0.040)	(0.079)	(0.076)	(0.058)
Observations	377,248	346,518	374,473	343,467	373,150
[5] Include Teachers with Fewer than 10 Students:					
Post*Teacher Share	0.245***	0.127***	-0.004	0.203*	0.028
	(0.088)	(0.045)	(0.088)	(0.103)	(0.081)
Observations	243,792	226,082	242,136	243,967	241,414
[6] Keep Classrooms with > 80% Special Ed or LEP:					
Post*Teacher Share	0.179*	0.040	0.096	0.142**	0.229
	(0.097)	(0.067)	(0.066)	(0.060)	(0.136)
Observations	244,699	238,429	242,386	243,658	244,540
[7] Drop Special Education Students:					
Post*Teacher Share	0.251***	0.139***	0.047	0.187*	0.038
	(0.086)	(0.048)	(0.083)	(0.107)	(0.089)
Observations	229,817	209,091	219,536	220,670	223,778
[8] Drop LEP Students:					
Post*Teacher Share	0.235**	0.151***	0.004	0.194*	0.048
	(0.088)	(0.049)	(0.100)	(0.104)	(0.087)
Observations	225,565	216,705	223,344	224,395	222,925

Source: HISD administrative data as described in the text. The math test in the first column is the state administered TAKS math exam. The English and Language Arts exams are Stanford tests. For 10^{th} and 11^{th} grade science and social studies, the TAKS exams are used, while for 9^{th} grade they are Stanford tests. All estimates are in terms of standardized scores. Controls include student gender, race, at-risk, special education, LEP, and gifted status along with lagged achievement interacted with grade-year indicators and grade-year and school-year fixed effects. To ease presentation we do not show the estimate for the "teacher share" main effect. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

Appendix

Figure A-1: Local Linear Regressions of the Effect of Teacher Share Post-ASPIRE on Student Achievement - Bandwidth of 0.1



Each line shows local linear regression estimates of *Share*Post* from models that include school-year and grade-year fixed effects as well as controls for school-grade-specific enrollment and lagged student test scores. Rectangular kernels are used with a bandwidth of 0.1. The dashed lines show the bounds of the 95% confidence interval that are calculated from standard errors that are clustered at the school level.

Table A-1: Additional Tests of "Impact" of Teacher Share on Student and Teacher Characteristics

	Test Subject:			
		English &		Social
Independent Variable	Math	Language	Science	Studies
Student is New to School (Grades 10, 11 Only)	-0.046* (0.026)	-0.008 (0.052)	-0.018 (0.030)	0.034 (0.034)
Observations	144,955	134,576	143,319	145,723
Student was Not Enrolled in District in Prior Year †	0.002 (0.009)	-0.021 (0.015)	-0.030** (0.012)	-0.015 (0.015)
Observations	241,694	224,964	240,544	242,196
Number of Courses Taught	1.113 (1.036)	2.463* (1.300)	-0.149 (0.931)	0.797 (0.539)
Observations	241,694	224,044	240,472	242,001

[†] Since the data are restricted to students having achievement data from 2004, these students would have been in the district prior, left and then returned; i.e. returning dropouts.

Each cell comes from a separate estimation of equation (8) and shows the estimate of α_2, which is the coefficient on $Share * Post$. Each independent variable is a dummy variable that indicates whether a student falls into the given category except where noted. Regressions also include school-year and grade-year fixed effects along with a quartic in enrollment. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

Table A-2: Analysis of Variance of Teacher Share Between and Within Teachers - 2006 and Later

	Test Subject:			
		English &		Social
	Math	Language	Science	Studies
[1] Raw Variance				
Between Teachers	69%	86%	77%	81%
Within Teachers	31%	14%	23%	19%
[2] Residual Variance - No School-Year FE				
Between Teachers	50%	69%	58 %	59%
Within Teachers	50%	31%	42%	41 %
[3] Residual Variance - With School-Year FE				
Between Teachers	40%	58%	46%	44%
Within Teachers	60%	42%	54 %	56 %

Percentages are calculated by conducting one-way ANOVA and then calculating the ratio of between teacher and within teacher sum-of-squares to total sum-of-squares. The first row uses raw Teacher Share across teachers. The second row uses the residuals from a regression of Teacher Share on student gender, race, at-risk, special education, LEP, and gifted status along with lagged achievement interacted with grade-year indicators and grade-year fixed effects.The third row uses residuals from a regression including the controls in panel [2] plus school-year fixed effects.

Table A-3: Tests of Heterogeneity by Share

Independent Variable	TAKS Math	English & Language	Science	Social Studies	Stanford Math
	Panel A: Estimates for Shares ≤ 0.15				
Post*Share	0.699**	0.453***	0.540**	1.172***	0.151
	(0.279)	(0.148)	(0.246)	(0.294)	(0.490)
Share	0.272**	-0.162	0.468**	0.748***	-0.311
	(0.121)	(0.104)	(0.185)	(0.255)	(0.287)
Observations	128,014	101,336	119,081	89,008	126,824
	Panel B: Estimates for Shares > 0.15				
Post*Share	0.213	0.075	-0.302**	-0.150	0.139
	(0.127)	(0.062)	(0.141)	(0.152)	(0.117)
Share	-0.089	-0.008	0.221	0.098	-0.231**
	(0.077)	(0.075)	(0.146)	(0.099)	(0.098)
Observations	113,680	122,708	121,391	152,993	112,526
	Panel C: Tests of Differences for Post*Share Between (A) and (B)				
Chi^2	2.57	5.68	8.85	17.96	0.00
P(Chi^2)	0.11	0.02	0.00	0.00	0.98

Source: HISD administrative data as described in the text. The math test in the first column is the state administered TAKS math exam. The Language exams are Stanford tests. For 10^{th} and 11^{th} grade science and social studies, the TAKS exams are used, while for 9^{th} grade they are Stanford tests. All estimates are in terms of scale scores standardized across the district within grade and year. Individual controls include student gender, race, at-risk, special education, LEP, gifted status and year fixed effects. Standard errors clustered at the school level are in parentheses: **,* indicates statistical significance at the 5% and 10% levels, respectively.

Table A-4: Additional Robustness Checks

	Test Subject:				
	TAKS	English &		Social	Stanford
Independent Variable	Math	Language	Science	Studies	Math
[1] Interactions with Grade Level:					
Post*Teacher Share	0.268**	0.153**	-0.189	0.140	-0.008
	(0.121)	(0.057)	(0.116)	(0.190)	(0.101)
Post*Teacher Share*10th	-0.075	-0.029	0.319	0.162	0.087
	(0.182)	(0.052)	(0.213)	(0.247)	(0.078)
Post*Teacher Share*11th	-0.020	-0.013	0.250	0.024	0.037
	(0.128)	(0.052)	(0.268)	(0.243)	(0.114)
Observations	219,430	197,560	202,017	203,793	208,282
[2] Assign Students to Spring of t - 1 and Fall of t Teachers for Grade/Subjects: that Use Stanford					
Post*Teacher Share	-	0.156***	-0.003	0.220**	0.029
	-	(0.051)	(0.109)	(0.109)	(0.074)
Observations	-	124,412	195,532	195,968	139,534
[3] Assign Students to Fall of t Teachers Only for Grade/Subjects that Use Stanford:					
Post*Teacher Share	-	0.146**	0.011	0.240**	0.023
	-	(0.054)	(0.104)	(0.115)	(0.087)
Observations	-	112,468	190,166	192,491	118,672
[4] Use 2002-03 Score as Lagged Achievement for All Years:					
Post*Teacher Share	0.174*	0.160***	0.107	0.223*	0.032
	(0.093)	(0.055)	(0.083)	(0.117)	(0.094)
Observations	219,430	197,560	202,017	203,793	208,282
[5] Unweighted Regressions:					
Post*Teacher Share	0.281***	0.140**	-0.012	0.210*	0.057
	(0.090)	(0.051)	(0.098)	(0.106)	(0.080)
Observations	241,694	224,044	240,472	242,001	239,350
[6] Add Teacher Fixed Effects (School FE Instead of School-Year FE):					
Post*Teacher Share	0.332***	0.068	0.345***	0.174	0.355***
	(0.103)	(0.092)	(0.121)	(0.104)	(0.119)
Observations	175,119	169,076	177,575	186,210	173,420

Source: HISD administrative data as described in the text. The math test in the first column is the state administered TAKS math exam. The English and Language Arts exams are Stanford tests. For 10th and 11th grade science and social studies, the TAKS exams are used, while for 9th grade they are Stanford tests. All estimates are in terms of standardized scores. Controls include student gender, race, at-risk, special education, LEP, and gifted status along with lagged achievement interacted with grade-year indicators and grade-year and school-year fixed effects. To ease presentation we do not show the estimate for the "teacher share" main effect. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

Table A-5: Estimates of the Effect of ASPIRE on Student Test Scores Using 2004 Department Size and Share

Panel A: Differences in Differences

Independent Variable	TAKS Math	English & Language	Science	Social Studies	Stanford Math
Post*$\frac{1}{\text{2004 Dept. Size}}$	0 .413**	0 .147	0 .406*	0 .293*	0 .753**
	(0.155)	(0.255)	(0.227)	(0.161)	(0.335)
$\frac{1}{\text{2004 Dept. Size}}$	-0.100	1 .303**	0 .897	0 .605*	-0.108
	(0.370)	(0.522)	(0.581)	(0.330)	(0.311)

Panel B: IV Using $\frac{1}{\text{2004 Dept. Size}}$ as an Instrument for *Share*

Independent Variable	TAKS Math	English & Language	Science	Social Studies	Stanford Math
Post*Teacher Share	0 .383**	0 .120	0 .270**	0 .581	0 .719**
	(0.153)	(0.140)	(0.130)	(0.875)	(0.351)
Teacher Share	-0.062	1 .056	0 .924*	6 .149	0 .227
	(0.508)	(0.606)	(0.512)	(8.708)	(0.435)

Source: HISD administrative data as described in the text. School fixed-effects are excluded as there is too little variation within schools in the instrument to generate precise estimates with these controls. The math test in the first column is the state administered TAKS math exam. The Language exams are Stanford tests. For 10^{th} and 11^{th} grade science and social studies, the TAKS exams are used, while for 9^{th} grade they are Stanford tests. All estimates are in terms of scale scores standardized across the district within grade and year. Individual controls include student gender, race, at-risk, special education, LEP, gifted status and year fixed effects. Standard errors clustered at the school level are in parentheses: **,* indicates statistical significance at the 5% and 10% levels, respectively.

Appendix B
Closed-Form Solution for Comparative Statics of Theoretical Model

Definitions:

$$\begin{aligned}
\bar{S} &= \theta S(e_1) + (1-\theta)S(e_2) - \mu \\
F = F(\bar{S}), F' &= F'(\bar{S}), F'' = F''(\bar{S}) \\
S_1 = S(e_1), S_1' &= S'(e_1), S_1'' = S''(e_1) \\
S_2 = S(e_2), S_2' &= S'(e_2), S_2'' = S''(e_2) \\
C_1 = C(e_1), C_1' &= C'(e_1), C_1'' = C''(e_1) \\
C_2 = C(e_2), C_2' &= C'(e_2), C_2'' = C''(e_2)
\end{aligned} \tag{11}$$

First order conditions:

$$G_1 = A\theta F' S_1' - C_1' = 0 \tag{12}$$

$$G_2 = A(1-\theta)F' S_2' - C_2' = 0 \tag{13}$$

Take derivatives of G_1 wrt θ, e_1 and e_2:

$$\frac{\partial G_1}{\partial \theta} = AF' S_1' + A\theta F'' S_1' S_1 \tag{14}$$

$$\frac{\partial G_1}{\partial e_1} = A\theta^2 F'' \left(S_1'\right)^2 + A\theta F' S_1'' - C_1'' \tag{15}$$

$$\frac{\partial G_1}{\partial e_2} = A\theta(1-\theta) F'' S_1' S_2' \tag{16}$$

Take derivatives of G_2 wrt θ, e_1 and e_2:

$$\frac{\partial G_2}{\partial \theta} = - AF'S_1' + A(1-\theta)F''S_2'S_2 \tag{17}$$

$$\frac{\partial G_2}{\partial e_1} = A\theta(1-\theta)F''S_1'S_2' \tag{18}$$

$$\frac{\partial G_2}{\partial e_2} = A(1-\theta)^2F''\left(S_2'\right)^2 + A(1-\theta)F'S_2'' - C_2'' \tag{19}$$

Implicit function theorem implies that:

$$\frac{\partial e_1}{\partial \theta} = \frac{\begin{vmatrix} -\frac{\partial G_1}{\partial \theta} & \frac{\partial G_2}{\partial e_1} \\ \\ -\frac{\partial G_2}{\partial \theta} & \frac{\partial G_2}{\partial e_2} \end{vmatrix}}{\begin{vmatrix} \frac{\partial G_1}{\partial e_1} & \frac{\partial G_2}{\partial e_1} \\ \\ \frac{\partial G_1}{\partial e_2} & \frac{\partial G_2}{\partial e_2} \end{vmatrix}} \tag{20}$$

$$\tag{21}$$

$$\frac{\partial e_1}{\partial \theta} = \frac{-\frac{\partial G_1}{\partial \theta}\frac{\partial G_2}{\partial e_2} + \frac{\partial G_2}{\partial \theta}\frac{\partial G_2}{\partial e_1}}{\frac{\partial G_1}{\partial e_1}\frac{\partial G_1}{\partial e_2} - \frac{\partial G_2}{\partial e_1}\frac{\partial G_2}{\partial e_2}} \tag{22}$$

Solve for elements of equation (22):

$$\begin{aligned} -\frac{\partial G_1}{\partial \theta}\frac{\partial G_2}{\partial e_2} =& - A^2(1-\theta)^2 F' S_1' F'' \left(S_2'\right)^2 \\ & - A^2(1-\theta)\left(F'\right)^2 S_1' S_2'' \\ & + AF'' S_1' C''(e_2) \\ & - A^2\theta(1-\theta)^2 \left(F''\right)^2 \left(S_2'\right)^2 S_1' S_1 \\ & - A^2\theta(1-\theta)^2 F' F'' S_2'' S_1' S_1 \\ & + A\theta F'' S_1' S_1 C_2'' \end{aligned} \tag{23}$$

$$\begin{aligned} \frac{\partial G_2}{\partial \theta}\frac{\partial G_2}{\partial e_1} =& - A^2\theta(1-\theta) F' F'' \left(S_1'\right)^2 S_2' \\ & - A^2\theta(1-\theta)^2 \left(F''\right) \left(S_2'\right)^2 S_1' S_2 \end{aligned} \tag{24}$$

$$\begin{aligned} \frac{\partial G_1}{\partial e_1}\frac{\partial G_1}{\partial e_2} =& A^2\theta^3(1-\theta)(F'')^2(S_1')^3 S_2' + \\ & A^2\theta^2(1-\theta) F'' F' S_1'' S_1' S_2' - \\ & A\theta(1-\theta) F'' S_1' S_2' C_1'' \end{aligned} \tag{25}$$

$$\begin{aligned} -\frac{\partial G_2}{\partial e_1}\frac{\partial G_2}{\partial e_2} =& - A^2\theta(1-\theta)^3(F'')^2(S_2')^3 S_1' - \\ & A^2\theta(1-\theta)^2 F'' F' S_2'' S_1' S_2' + \\ & A\theta(1-\theta) F'' S_1' S_2' C_2'' \end{aligned} \tag{26}$$

$$\tag{27}$$

Solution for numerator:

$$\begin{aligned} -\frac{\partial G_1}{\partial \theta}\frac{\partial G_2}{\partial e_2} + \frac{\partial G_2}{\partial \theta}\frac{\partial G_2}{\partial e_1} = & - A^2(1-\theta)S_1' \Big[(1-\theta)F'F''\left(S_2'\right)^2 + \left(F'\right)^2 S_2'' \\ & + \theta(1-\theta)\left(F''\right)^2\left(S_2'\right)^2 S_1 + \theta F'F''S_2''S_1 \\ & + \theta F'F''S_1'S_2' + \theta(1-\theta)\left(F''\right)^2 S_2'S_2\Big] \\ & + AC_2''S_1'F''\left[1+\theta S_1\right] \end{aligned} \tag{28}$$

Solution for denominator:

$$\begin{aligned} \frac{\partial G_2}{\partial \theta}\frac{\partial G_2}{\partial e_1} - \frac{\partial G_1}{\partial \theta}\frac{\partial G_2}{\partial e_2} = & A\theta(1-\theta)F''S_1'S_2' \times \\ & \big[A\theta^2F''(S_1')^2 + A\theta F'S_1'' - C_1'' \\ & A(1-\theta)^2F''(S_2')^2 + A(1-\theta)F'S_2'' + C_2''\big] \end{aligned} \tag{29}$$

Full closed form solution:

$$\frac{\partial e_1}{\partial \theta} = \frac{-A^2(1-\theta)S_1'\left[(1-\theta)F'F''\left(S_2'\right)^2+(F')^2S_2''+\theta(1-\theta)(F'')^2\left(S_2'\right)^2S_1+\theta F'F''S_2''S_1+\theta F'F''S_1'S_2'+\theta(1-\theta)(F'')^2S_2'S_2\right]}{A\theta(1-\theta)F''S_1'S_2'\left[A\theta^2F''(S_1')^2+A\theta F'S_1''-C_1''A(1-\theta)^2F''(S_2')^2+A(1-\theta)F'S_2''+C_2''\right]}$$