Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System

Scott A. Imberman*
Michigan State University and NBER

Michael F. Lovenheim
Cornell University and NBER

March 2013

Abstract

We estimate the impact of incentive strength on student achievement under a group-based teacher incentive pay program. Awards are based on the performances of students within a grade, school and subject, providing substantial variation in group size. Using the share of students in a grade-subject enrolled in a teacher's classes as a proxy for incentive strength, 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. Mean science estimates are small and are not statistically significant. We also find substantial heterogeneity by share. 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. Calculations based of these estimates show large positive effects overall of group incentive pay on achievement.

KEYWORDS: Teacher Incentive Pay, Free Riding, Teacher Productivity
JEL CLASSIFICATION: I21, J33, J38, H41

---

*Imberman: Department of Economics, Marshall-Adams Hall, 486 W Circle Dr Rm 110, East Lansing, MI 48824, imberman@msu.edu. Lovenheim: Department of Policy Analysis and Management, 135 Martha Van Rensselaer Hall, Ithaca, NY 14853, mfl55@cornell.edu. We wish to thank seminar participants at Aarhus University, CESifo, Cornell University, the Institute for Research on Poverty at the University of Wisconsin, NBER Summer Institute, Purdue University, Teachers' College, Tilburg University, University of Copenhagen, University of Virginia, and University of Houston for helpful comments and suggestions. We further thank Kiel Albrecht, Jack Barron, Aimee Chin, Judy Scott-Clayton, Steve Coate, Steven Craig, Steve Rivkin, Jonah Rockoff, Gary Solon and Lesley Turner for helpful comments and suggestions. Finally, we would like to thank the employees at the Houston Independent School District for their help and assistance. All errors, omissions and conclusions are our own.

1 Introduction

Teacher incentive pay has become an increasingly popular education policy throughout the world. However, at least in a developed country context, the results from several recent randomized controlled trials suggest that linking teacher pay to their students' academic performance does little to raise student achievement (Fryer, 2013; Goodman and Turner, 2013; Fryer, Levitt, List, and Sadoff, 2012; Springer, et. al., 2010).[1] The lack of effects found in these studies is consistent with teachers being unresponsive to financial incentives for improving student outcomes. Alternatively, it could be the case that the design of the incentive schemes limited their ability to raise test scores. Currently, little is known about how the specific design aspects of teacher incentive pay programs, such as the group size, the outcome measure being incentivized, and the payment structure, influence the effectiveness of these systems in raising student academic achievement.[2] It is critical to understand how these design aspects impact their effectiveness in order to construct incentive pay programs that provide properly powered and salient incentives for workers.

We focus on a particularly important aspect of group-based teacher incentive pay design – the size of the incentivized group – in order to shed needed light on the role of a core design feature. Due to the stated desire of teachers to foster an environment of cooperation and collaboration, group-based incentive pay systems that pay teachers based on grade- or school-specific performance on standardized exams in a given subject are the most common in education. A potential drawback of providing rewards based on group performance, however, is the "$\frac{1}{n}$" problem (Holmstrom, 1982; Kandel and Lazear, 1992), which can lead teachers in larger groups to be less responsive to financial incentives. The "$\frac{1}{n}$" problem is driven by two related mechanisms: free riding and "award salience." Free riding occurs in group-based incentives because each worker has an incentive to reduce his effort and consume more leisure in response to the expected benefit he receives from the effort of others in the group. "Award salience" is a related but distinct mechanism, whereby the strength of the effort incentive is reduced as the

[1] Fryer, et al. (2012) do find positive effects, but only when teachers have to pay back an earlier bonus payment for poor performance. Their traditional incentive pay scheme did not elicit positive effects.

[2] Barlevy and Neal (2012) provide the only theoretical work on optimal teacher incentive pay of which we are aware; their main focus is on how to structure incentive pay systems to avoid teaching to the test.

group size increases, because the expected marginal benefit of an increase in effort falls.[3]

The existence of larger groups also could enhance worker responses to group-based incentives. For example, larger groups may encourage more cooperation and coordination of teaching strategies across teachers or induce teachers to take advantage of efficiencies of scale through technology spillovers and team teaching. Group size may affect peer monitoring as well, which could either increase or decrease the effect of group-based monetary incentives depending on whether peer monitoring is stronger or weaker in larger groups.[4] Taken together, these mechanisms indicate that how individual performance responds to group-based incentives is an empirical question about which little currently is known across labor markets, and especially 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[5] 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.[6]

Our empirical analysis focuses on testing for the existence and extent of group size effects using individual student-level data from before and after implementation of the incentive pay

[3]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.

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

[5]ASPIRE stands for "Accelerating Student Progress, Increasing Results and Expectations." Further details on the program can be found at http://portal.battelleforkids.org/aspire/home.html.

[6]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.

program. Specifically, we examine whether teachers who are responsible for a larger share 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 (i.e., the award salience increases) and free rider incentives decline. 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, although also test for direct effects of group size. 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-differences 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 by examining direct measures of sorting and by showing the robustness of our estimates to the use of different sources of share variation that are each subject to different potential biases from endogenous sorting.

Our results show evidence that student performance increases more post-ASPIRE among those whose teachers are responsible for a larger share of students. 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 the students in the higher-share grade. These results point to teacher productivity increasing as financial incentives become stronger. Furthermore, despite large effects on the incentivized math exam, we find no impact of ASPIRE on a non-incentivized math test taken by all students.

The change in the relationship between teacher share and student test scores when ASPIRE was enacted is unlikely to be linear due to the fact that free rider effects are much larger at lower shares and that the marginal return to effort is decreasing in effort. Hence, we estimate local linear regression models that allow for the changing effect of share due to ASPIRE to vary non-parametrically over the distribution of share. Indeed, we find evidence of much heterogeneity: 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. 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. Notably, the New York City school-based teacher incentive pay experiment (Fryer, 2013; Goodman and Turner, 2013) used group sizes that imply average shares well below 0.2.[7]

In order to test whether our results are driven by variation in incentive strength or by variation in peer monitoring and cooperation that are more likely a function of the number of teachers in each group, we also estimate local linear regression models controlling for both teacher share and the number of teachers in the group. Conditional on share, we find little evidence of a role for department size except in English. These estimates point to little scope for cooperation and monitoring to play a significant role in driving group size effects in teacher incentive pay systems.

Recent work on teacher incentive pay has created doubts about whether teachers respond to financial incentives at all, especially in more developed countries. Hence, a major contribution of our analysis is to establish that teachers are responding to incentive pay as a function of the share of students they teach. Our estimates also allow us to calculate the implied average

[7] While our results are suggestive that the lack of effects found in the group-based incentive pay literature are due to large group sizes, they cannot speak to the ineffectiveness of individual incentive pay (Fryer et. al., 2012; Springer et al. 2010). We discuss the implications of our findings for the broader teacher incentive pay literature as well as some mechanisms that could lead individual incentive pay to poorly incentivize teachers in Section 7.

effect of the ASPIRE program on student test scores by multiplying each teacher's share by the estimated change in test scores at that share. Using the local linear regression estimates, we find that achievement increased by between 3 percent (in English) and 10 percent (in social studies) of a standard deviation in HISD high schools as a result of the teacher incentives. These effects provide evidence that when structured correctly, group-based teacher incentive pay systems can have large positive effects on student academic achievement. Overall, the results from this analysis indicate that design features matters a lot in determining how effective an incentive system is in increasing productivity. These findings 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 group size effects in incentive pay programs. In Section 3, we describe the HISD incentive pay program, and our data are discussed in Section 4. We present our empirical methodology in Section 5. All results are discussed in Section 6, and Section 7 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.[8] Lavy (2002) studies a school-wide performance incentive program in Israeli Public Schools 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

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

less responsive to the implementation of the award system. Lavy (2009) studies another incentive pay tournament in Israel that was individual-teacher based and finds that test-taking, passing, and high school exit math test scores increased significantly due to teachers' 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 are to the educational environment we study.[9] 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, these estimates are difficult to generalize to a more developed country context due to large cross-country differences in educational systems.

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, 2013; Goodman and Turner, 2013) and individual incentive pay in Nashville, Tennessee (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 enact a set of human resource reforms, including incentive pay that is based on a wide array of outcomes that vary across schools. They find small but significant positive effects of opting into this system, but the design of the program does not allow them to disentangle the effects due to the human resource reforms from the incentive pay impacts. 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.[10] Finally, Fryer,

[9]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.

[10]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.

et al. (2012) conduct an experiment that gave some teachers individual-based award bonuses and other teachers fixed cash pay-outs prior to the school year that were then required to be returned if performance was low. Consistent with much of the literature, they find no significant impacts from the first group, but they do find improvements from the second group suggesting that loss aversion is a more powerful incentive than standard pay-for-performance. This paper highlights how the particular design of a program matters. It also is 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 (2013). 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 examine whether the effect of the incentives vary with group size. They present evidence that achievement declined slightly in larger schools and may have increased by a small amount in smaller schools. However, given that the groups were all school-wide, they cannot test whether the differences in responsiveness by school size are causally related to group size or whether they are due to school attributes that are correlated with school size. Furthermore, they are unable to 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.[11]

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

[11] 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.

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, scant 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 worker behavior is critical to developing optimal merit pay systems. The 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.

3 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 Texas Assessment of Knowledge and Skills (TAKS). 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.[12] The current department's fixed effect is captured and then adjusted via a Bayesian shrinkage estimator so that estimates for departments with fewer observations are attenuated towards the mean (which equals zero by construction).[13] This adjusted department fixed effect is the department value-added score. The value-added measures are then ranked within grade, subject and year. 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 be eligible for awards in all grades in his subject regardless of whether he teaches each grade, each award is based on grade- and subject- specific performance. For example, if a teacher only teaches 9^{th} grade students in science, her students only contribute to the 9^{th} grade portion of the science award. However, if the 10^{th} grade science teachers in her school win an award, she will receive that award money as well. Despite the fact that teachers may receive bonus money due the actions of teachers in other grades, the

[12] More precisely the model estimates regressions of the form

$$Score_{igst} = \alpha + \sum_{s}^{S}\sum_{j}^{J}\sum_{k=t-2}^{t}(\beta_{jgsk} \times weight_{isk} \times T_{jgsk}) + \gamma_s + \lambda_g + \mu_t + \varepsilon_{igst}$$

where "*score*" is student i's achievement in subject s, grade g and year t. "*weight*" is a measure equal to the inverse of the number of teachers a student has in subject s in year k if grade < 9, otherwise weight equals one. "T" equals one if the student is assigned to teacher or department j in subject s and year k. Finally, γ, λ, μ are fixed effects for subject, grade and year, respectively. Regressions are pooled over multiple years and all students in the district. Note that while departments are used in grades 9 and above, in prior grades students are linked directly to individual teachers and thus the level of estimate of β depends on the grade for year k. See Wright, White, Sanders and Horn (2010) for a detailed technical treatment

[13] Unfortunately, EVAAS does not provide information on how they construct the shrinkage factor.

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 faces monetary incentives to get over an award threshold.[14] Furthermore, the most salient measure of the incentive 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 2006-07 and 2007-08, teachers could earn up to $5,500 from the departmental awards.[15] In addition to the these awards, there are a series of awards for school-wide performance.[16] Each of the school awards are relatively small, ranging from $150 to $750 apiece, hence we do not consider them in our analyses. Nonetheless, they raise the maximum total award a teacher could receive to $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. The maximum award amounts to about 20% of a beginning teacher's total wage compensation, 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 (including elementary and middle schools) was $3,614 in 2009-10. The large bonus amounts relative to base pay suggests there is substantial scope in this system for teachers to respond to financial incentives.

One potential concern 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,

[14] 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 Online Appendix Table A-1.

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

[16] 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 school-wide participation in and performance on Advanced Placement and International Baccalaureate exams.

teachers were asked about their level of understanding of the program parameters.[17] 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 and grade. Since detailed documents that explain the value-added system are easily accessible to teachers online, we believe this condition likely is 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.

4 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 are only 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

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

"Texas Assessment of Knowledge and Skills" (TAKS).[18] 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. For both types of exams, we standardize the scale scores within grade, subject and year to have a mean of zero and a standard deviation of one. 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, and teachers are matched to awards based on a list compiled in 2009 of courses that count for each award.[19]

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 awards. In this case, the student's achievement only would count towards the value-added metric that determines awards once even though the student appears 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.[20]

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.

[18] We do not know whether a given seating of this exam is the first or a retake after failing the first exam. 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. We thus use each student's lowest score in a year as our achievement outcome for the TAKS exam.

[19] 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. 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.

[20] Results are similar without weighting and are provided in the Online Appendix Table A-1.

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.[21]

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. Since there is ambiguity about how to best link Stanford tests to teachers, we provide robustness checks from models that link students to 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. These results are provided in Online Appendix Table A-1 and show that our results 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. For each of these

[21] 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 and reading courses in the grade.

sample restrictions, 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.

Table 2 provides summary statistics and exact observation counts from 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 also are relatively low income, with 70% being economically disadvantaged.[22] 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 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.[23]

5 Empirical Methodology

Our empirical analysis focuses on testing a central implication of the group incentive pay models of Holmstrom (1982) and Kandel and Lazear (1992), that workers should be more responsive to a given monetary incentive when they are responsible for a larger share of the output. However, some unique aspects of teaching, such as the desire for cooperation and peer monitoring, may

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

[23]Note that the mean share is not equal to the inverse of the number of teachers because students in teacher-course cells with rates of LEP or special education students over 80%, which generally are smaller share courses, are dropped from the sample even though their teachers still count in department size calculations. Thus, the remaining students have teachers with larger shares. Also, the total number of teachers in the school does not equal the sum of the department-grade sizes shown, as many teachers teach in multiple grades and subjects.

counteract this prediction. We identify whether teachers who are responsible for a larger share of students increase test scores more post-ASPIRE than pre-ASPIRE using a difference-in-differences model. If students were randomly assigned to classrooms, we could simply compare teachers with higher and lower shares after program implementation. But, 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 use administrative data from HISD on student test scores, student demographics and teacher assignments as described in Section 4 to estimate the following model:

$$\begin{aligned} A_{isgjt} =& \beta_0 + \beta_1 Share_{sgjt} + \beta_2 Share_{sgjt} * Post_t + \\ & \sum_t \sum_g \gamma_{gt} A^{pre}_{isgjt} \times Year_t \times Grade_g + X'_{it}\Phi + \lambda_{gt} + \nu_{jt} + \varepsilon_{isgjt}, \end{aligned} \tag{1}$$

where A_{sigjt} is test score in subject s of student i in grade g with teacher j in year t, $Share$ 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.[24] 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. In addition to these controls, equation (1) 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.[25]

[24]Results are similar if we use 2002-03 achievement as the lagged score for all years and grades and are provided in Online Appendix Table A-1.

[25]Clustering standard errors still may cause one to over-reject null hypotheses when the number of clusters

The coefficient of interest in equation (1) 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. Critically, we also control for $Share$, which estimates the underlying relationship between teacher share and student academic achievement in the absence of the incentive pay program, conditional on the extensive set of controls in our model. There are several reasons to believe that there will be a pre-existing correlation between $Share$ and test score growth: principals may generate better teacher-student match quality for teachers with a higher share or might assign the best teachers to teach the higher-share classes. Conversely, teachers with higher shares may perform worse if the larger volume of students negatively impacts her performance. The parameter β_1 picks up this underlying relationship between teacher share and student achievement, and thus our model controls for any underlying endogenous relationship between share and test score growth. The parameter of interest, β_2, is identified off of any change in this relationship when the ASPIRE program comes into place. The main identifying assumption we invoke is not that $Share$ is exogenous (i.e., $\beta_1 = 0$), but that the reason for any change in the relationship between teacher share and student outcomes when ASPIRE is enacted is due to teachers' responses to the incentives they face under the program and not due to changes in the sorting mechanism that drives any pre-ASPIRE correlation between $Share$ and test scores.

In assessing the plausibility of this identification assumption, it is helpful to clarify the sources of variation in teacher share that are used to identify β_2. Conditional on the fixed effects in equation (1), one source of variation in $Share$ comes from 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

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.

specific subjects across cohorts, and teacher turnover. The variation in $Share$ in equation (1) also comes from differences in teacher share across different classes taught by the same teacher in different grades and across different teachers within and across grades.[26] In order to help assess the validity of our identification strategy, we estimate equation (1) using different fixed effects that each isolate different aspects of the identifying share variation. We estimate the model using school-grade-year (i.e., department) fixed effects that only allow share variation across teachers in the same department and year as well as using teacher-year fixed effects that include share variation only within teachers across grades in the same subject and year. To the extent that each of these fixed effects estimates provide similar results to those from equation (1), it will be evidence that our main estimates are not being driven by changes in the mechanism that sorts students to teachers when ASPIRE comes into place, as each of these sources of teacher share variation are subject to biases from very different sorting mechanisms.[27] Finally, we estimate a model that uses $\frac{1}{\text{2004 Department Size}}$ as an instrument for $Share$. By employing only pre-ASPIRE variation in share, this model cannot be biased by any endogenous changes in student or teacher sorting post-ASPIRE. However, given the lower statistical power in this model, it is not our preferred specification.

Ultimately, it is not possible in our setup to know perfectly why share varies across teachers or within teachers over time. In order for β_2 to provide an unbiased estimate of responses to stronger merit pay incentives, however, 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. It is important to emphasize that the principals do not have direct monetary incentives for re-sorting students to maximize the likelihood of teacher award receipt, as they do not receive money from the teacher-based awards.[28] Furthermore, simply reshuffling students to match those with higher

[26]In Online 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 over time.

[27]For example, the teacher-year fixed effects estimates would be biased if principals sorted students such that post-ASPIRE, higher-growth students were being put differentially in the teacher's higher-share class relative to her lower-share class. The school-grade-year fixed effects estimates would be biased by principals sorting higher-growth students to the teachers with higher share in the department, post-ASPIRE.

[28]Principals and assistant principals were eligible for awards, but they are based on school-wide rather than

underlying test score growth to higher-share teachers would not affect the likelihood of award receipt, because department-average test scores would be unaffected by such changes. We thus would expect any changes in student-teacher sorting to be due to either increasing the match quality between higher-share teachers and students or due to increasing share among higher-productivity teachers post-ASPIRE. If principals did alter student-teacher match quality more for higher-share teachers or re-organized shares across teachers in a way that increased aggregate test scores, this still would be a positive causal effect of the program on department-average test scores, although it would be coming through altering student and teacher assignments rather than through increasing teacher effort. Thus, even in the case of shifting student-teacher matching, we are still identifying the causal effect of ASPIRE on student test scores and how this effect varies with teacher share. Additionally, there was a school accountability regime in place in Houston throughout the entire study period, and the accountability incentives principals faced were unchanged by ASPIRE. If principals could have re-organized teachers and students to increase aggregate output, they likely already would have done so in response to the accountability incentives, as they are much stronger than any incentives they face under ASPIRE.

In order to gain insight into the extent to which our findings are driven by changes in teacher effort or changes in student-teacher sorting, we examine whether there is any evidence that student composition or measured academic achievement shifted as a function of share post-ASPIRE. In Table 3, we present balancing tests that show the correlation between our key explanatory variable and demographic characteristics of students. In particular, we estimate regressions of the following form:

$$x_{isgjt} = \alpha_0 + \alpha_1 Share_{sgjt} + \alpha_2 Share * Post_{sgjt} + \lambda_{gt} + \nu_{jt} + \varepsilon_{isgjt}, \qquad (2)$$

where x is a specific student characteristic and all other variables are as previously defined. Table 3 shows 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

on department-grade performance. The incentives under these awards thus were only partially aligned with those of teachers.

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 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. We further stress that we control for lagged achievement and other student observables in all of our models, which helps address the potential sorting in science. Indeed, in the last row we show estimates of the impact of share on pre-ASPIRE achievement controlling for lagged achievement and find that the science estimate drops substantially and remains insignificant, as do the estimates for the other exams as well.

As discussed above, teacher shares also 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, due to accountability pressures, the principal already had an incentive to maximize group achievement by assigning the best teachers the highest shares before ASPIRE.

We address this concern in a few ways. First of all, below we show that our estimates are robust to the inclusion of teacher-year fixed effects. These estimates use variation in share within teachers but across grades. If our estimates were were due to the sorting described here, we would not find any effects within teacher-year. Second, we also show estimates below from two-stage least squares model using pre-ASPIRE inverse department size as an instrument for share. This instrument is unrelated to ex-post resorting.[29] Third, if such sorting was occurring, we would expect the effects to show up on both incentivized and non-incentivized exams. Below we show that, in the case of math where we have data on an exam that was not subject to incentives, there is no change in the impact of share on achievement post-ASPIRE.

Finally, if principals alter teacher shares endogenously in response to ASPIRE, there should be a shift in the teacher share distribution towards having more teachers with large teacher shares. Figure 1 provides teacher share distributions 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. These results are inconsistent with adjustments in teacher assignments that were systematically related to teacher share concurrent with program implementation.

6 Results

6.1 Baseline Estimates

Before presenting our estimates of equation (1), 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 (1) 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

[29] In Online Appendix Table A-3, 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 (5% level) 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 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.

to better measure trends. Figure 2 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 also show a clear level shift after 2005. For science and social studies[30], 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 earlier 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 (1). The estimates in each column of each panel come from separate regressions. In the first panel, we include grade-year and school fixed effects as well as the full set of student demographic and lagged test score controls discussed in the previous section. In Panel [2], we add school-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 both specifications. In Panel [2], which is our preferred model, 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 post-ASPIRE. 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

[30]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.

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. This result indicates that there are some unobserved school-level test scores shocks that are correlated with share post-ASPIRE. That they move in opposite directions in English and science is suggestive that these shocks are idiosyncratic and are not driven by the treatment. We favor the school-year fixed effect model in Panel [2] due to the fact that it controls for such secular variation. With significant and positive impacts for math, English and social studies in Panel [2] of Table 4, the baseline results indicate that teachers do respond to changes in the share of students in a positive direction.

We also provide the estimates on $Share$ in Table 4 that show how teacher share and student test scores were correlated prior to ASPIRE. As explained above, our identification strategy is unaffected by any pre-existing correlation between achievement and share, but it still is of interest to examine the strength and sign of any pre-ASPIRE relationship among these variables. Pre-ASPIRE, teacher share was only weakly positively correlated with math and English test scores, conditional on the controls in the model. The correlations are positive and larger in magnitude for science and social studies. Thus, for math and English, the ASPIRE program generated a positive relationship between test scores and teacher share, while for social studies it strengthened a pre-existing positive relationship.

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.[31] 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 shows little evidence of spillovers to the non-incentivized test. While the coefficient is 0.22 in Panel [1], it is not statistically significantly different from zero at conventional levels and it falls considerably once school-year fixed effects are included. This finding could be indicative of teachers focusing specifically on the incentivized exam, but

[31] 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.

it also is the case that the TAKS and Stanford exams have limited topical overlap. Since 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.[32] 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.

6.2 Estimates Using Different Sources of Teacher Share Variation

As discussed in Section 5, the empirical setup does not allow us to fully explain why share varies across teachers or within teachers over time. It thus is important to explore the sensitivity of the results to different sources of share variation that each are subject to potential biases from different sorting mechanisms. Our preferred estimates in panel [2] of Table 4 are identified off of a few sources of variation, though in all cases the estimates are based on variation in the post-ASPIRE period relative to the pre-ASPIRE period in order to exploit the implementation of the award system. Differences in share across teachers within departments and year is one important component of the identifying variation. The second core variation source is differences in share distributions in a given year within schools but across grades. Third, differences in share within teachers and year but across grades contribute to identification. In order to investigate the role of these multiple sources of variation, in Table 5 we estimate models that either eliminate or isolate specific sources. Showing that our estimates are robust to using different types of share variation supports the validity of our identification strategy, as it is unlikely that there is a reasonable sorting mechanism that would affect all these sources in the same way.

In the first panel of Table 5, we control for school-grade-year (i.e., department-year) fixed effects as well as the controls included in equation (1). This model is identified only off differences

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

in share between teachers in the same department and year. Thus, it does not use identifying variation from differences in share across grades in a school. Furthermore, it implicitly controls for any direct impact of department size and any department-level shocks that do not work through share. The resulting estimates are similar to those in Panel [2] of Table 4.

In Panel [2] of Table 5, we provide a set of estimates that relies solely on variation within teacher and year by including teacher-year fixed effects in equation (1). 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}. 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. They also allow us to focus on this single source of share variation; the model eliminates any share variation across teachers and variation within teachers over time. The results, shown in Panel [2], are positive and significant for all four incentivized exams, with no impact on the non-incentivized Stanford math exam. These findings mirror the baseline estimates. The only estimate that is notably different from those in Table 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 post-ASPIRE. They also provide further support for the 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 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.[33]

It also is interesting that there is a strong positive correlation between share and student

[33]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.

scores pre-ASPIRE in this model. These results are consistent with principals targeting higher shares towards teachers in the courses for which she is most productive. One concern this raises is whether principals may have exacerbated this behavior post-ASPIRE. We note though that this is not sorting in the sense we should be concerned with - that principals sort higher achieving students to high share teachers. Rather this would be an enhancement of match quality between students and teachers. To the extent ASPIRE leads to better matches, that is a productivity enhancement and our estimates are capturing the effects of that enhanced productivity.

Finally, in Panel [3] of Table 5, we provide estimates that use share variation based only on pre-ASPIRE department-average share. In particular, we instrument $Share$ and $Share * Post$ with $\frac{1}{2004 DepartmentSize}$ and $Post * \frac{1}{2004 DepartmentSize}$. Hence, the estimates are identified solely off of share measures that exist prior to ASPIRE and that are unrelated to teacher quality variation or student-teacher matching changes within a school-grade. This specification also eliminates the possibility that the estimates are identified off of within school and grade sorting. The drawback, however, is that since there is very little within-school variation, we cannot use school fixed effects, and thus the estimates are subject to bias if schools with historically small departments tend to respond more to ASPIRE for reasons other than the impact of average share. Because of this limitation and the imprecision of the estimates, we do not use these as our preferred estimates. Nonetheless, the estimates are on the whole similar to our main results. Math and English estimates are positive and of the same order of magnitude as baseline. Science, on the other hand, becomes positive and significant, which is similar to the results in Panel [2] of Table 5. These results suggest that if anything, our baseline estimates serve as a lower bound for science. For social studies, the estimates are positive but are very imprecise due to a weak first stage on "share." Finally, unlike in our other estimates, we do find significant impacts on the non-incentivized Stanford math exam, but we note that the standard errors are quite large. Together, the results in Table 5 show that our estimates are robust to using different sources of variation and support our claim that our preferred estimates in Table 4 are not biased by changes in how students and teachers are matched when the incentive pay program is implemented.

6.3 Heterogenous Treatment Effects by Teacher Share

Thus far, we have estimated the mean effect of increasing teacher share under a group incentive pay regime over the entire distribution of shares. However, these estimates may hide important information, as the effect of share may be larger among those with lower shares.[34] This prediction comes from the fact that free riding incentives can fall non-linearly with share increases and that the marginal benefit of effort is increasing with effort. Thus, as share increases, effort would be expected to increase at a decreasing rate. 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 3. This method allows us to examine non-parametrically how the effect of teacher share changes when the incentive pay system is implemented.[35] The figure shows point estimates and 95% confidence intervals from a series of regressions of equation (1) 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.[36]

Figure 3 shows evidence of a large amount of heterogeneity in effects as a function of share: the estimate for $Share * Post$ starts out positive at low shares and then falls to zero for all four subjects.[37] 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

[34] In previous versions of this paper, we presented an illustrative theoretical model that shows this prediction comes out of a simple model in which teacher heterogeneity is solely a function of exogenously assigned share. In order to keep the focus of our study on the empirical results, we refer the reader to Imberman and Lovenheim (2012) if they are interested in details of the model.

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

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

[37] 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. Except for a small range around 0.4, these estimates are not statistically significantly different from zero at the 5% level.

each subject, including ELA. Hence, Figure 3 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 3 do not lend themselves simply to statistical tests that the effect of share on test scores post-ASPIRE declines with share. Thus, in Online Appendix Table A-4 we estimate equation (1) 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 4, 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 3 we see significant effects for TAKS math, the estimate for Stanford math is close to zero. These results indicate that there are no spillovers from improvements in TAKS due to larger incentive strength into the Stanford test. Nonetheless, it remains to be seen whether this is due to "teaching to the test" or because of only partial topical overlap between the two exams.

Although we model the teacher response as a function of the share of the students they teach, there is also 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.[38] In Figure 5, we provide local linear regressions of equation (1) with the addition of a variable for the number of teachers in the department ($DepartmentSize$) and its interaction with being in the ASPIRE period ($Post * DepartmentSize$). The left column

[38]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.

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. In fact, it appears that a larger department has a slight positive effect independent of share, suggesting that there are some potential benefits to having a larger group. The inclusion of these additional controls strengthens the share results as well. 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 3. These results indicate that teachers are, for the most part, 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 free-riding effects in education and 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 3 and 5 are particularly important as they provides us information about achievement-maximizing shares. Once teacher shares reach 0.2 to 0.3, there are no more returns to increasing shares (i.e., decreasing group size) in a group-based incentive pay system. These findings imply that group-based incentive pay systems will produce larger test score gains either when sufficiently small groups are used or when students are re-allocated across teachers to avoid having teachers with low shares. In the context of the previous literature on group-based teacher incentive pay, our results suggest that school-level incentive programs may have little effect on teacher behavior because each teacher's share is so low that the incentive she faces is very weak.

6.4 The Effect of ASPIRE on Average Student Test Scores

While the primary goal of this paper is to estimate how teacher responses to financial incentives vary with the strength of those incentives as proxied by *Share*, it also is of high interest to estimate the impact of the program on average test scores. Unfortunately, with only one treated district, we are limited in our ability to identify such average effects. However, we can use our incentive strength estimates to calculate a likely lower-bound effect of the program on student academic achievement. The insight underlying this calculation is that our share estimates provide us with $\frac{\partial Score_{isjgt}}{\partial Share_{sjgt}}$. We thus can recover the main effect of ASPIRE on a given teacher up to an unknown constant by integrating over all shares. That is

$$E\left(\frac{\partial Score_{isjgt}}{\partial ASPIRE}\right) = \alpha_{sjgt} + \int_{k=0}^{Share_{sjgt}} \hat{\beta}_{2,s,k} dk \tag{3}$$

Since we cannot calculate this integral directly, we estimate it using discrete analogs along with the estimated share effects. First, we use the linear model estimated in Panel [2] of Table 4. In this case, the total productivity effect for teacher j in grade g and subject s is

$$E\left(\frac{\partial Score_{isjgt}}{\partial ASPIRE}\right) \approx \hat{\beta}_{2,s} * Share_{sjgt} \tag{4}$$

While this estimate provides a useful baseline, the findings in Figure 3 show substantial heterogeneity by share. Thus, we also provide calculations of total effects using the estimates from the local linear regressions in Figure 3.[39] In this case, the total productivity amounts to calculating the area under the curve in Figure 3 from zero to the share value for a given teacher:

$$E\left(\frac{\partial Score_{isjgt}}{\partial ASPIRE}\right) \approx \sum_{k=0}^{Share_{sjgt}} \hat{\beta}_{2,s,k} * 0.01 \tag{5}$$

Note that in the summation above the steps are of intervals equal to 0.01. The average effect of the program across all students is calculated by taking a weighted sum of (4) or (5) across students-course observations, where the weights are the inverse of the number of courses each

[39]Though they are not shown in the figure due to large standard errors, for these calculations we include teachers with shares above 0.5.

student takes in the given subject and year.

A key limitation of this strategy is that the constant in (3) cannot be recovered, and in fact (4) and (5) assumes that $\bar{\alpha}_{sjgt} = 0$. If in reality $\bar{\alpha}_{sjgt} > 0$, then we can consider our calculation to be a lower bound on the true impact of the program. It is possible though that $\bar{\alpha}_{sjgt} < 0$, which would imply that we overestimate the program effect. Nonetheless, we think that this situation is particularly unlikely. For the constant to be negative, teachers would have to respond negatively to the existence of the program even when they respond positively to the incentives imbedded within the program. There are few plausible scenarios that could generate such behavior.[40]

With this caveat in mind, Table 6 presents the average effect calculations using the effects implied by the point estimates as well as the upper and lower bounds of the 95% confidence intervals from the underlying regressions. The estimates based on pooled difference-in-differences models indicate that ASPIRE increased test scores by 3% to 5% of a standard deviation for math, English and social studies. There is no effect on average science achievement.

Although less precise, the calculations based on the local linear regressions are our preferred estimates because they account for heterogeneity by share. These models provide considerably larger total effect sizes. For math, science and social studies, our results suggest an average effect of ASPIRE of between 7 and 10 percent of a standard deviation, while English test scores increased by 3 percent of a standard deviation. These imply total effects equal to between one-quarter and one-half of the effect of reducing class sizes by 7 students (Kruger, 1999) and are about the same size as a one standard deviation increase in teacher quality (Rivkin, Hanushek and Kain, 2005; Rockoff, 2004). Given that they are likely to be lower-bound estimates, the results in Table 5 suggest that ASPIRE had a large, positive effect on student test scores in Houston, which is in contrast to many of the other recent findings on teacher incentive pay

[40]One possibility is that the incentive pay program generates a culture in the school that is hyper competitive, leading to an overall reduction in achievement even as teachers respond positively to the specific incentives they face. While this is a reasonable theory in the context of an individual incentive regime, we would expect the group-based nature of the award to reduce competitiveness considerably and replace it with cooperative behavior (Hamilton, Nickerson and Owen, 2003). Another possibility is that teachers become jealous of others who win awards when they do not, which, in turn, affects their productivity. The fact that awards are grouped by department and that teachers win money if other grades in the same subject and school win an award likely mutes this effect.

(Neal, 2011).

We can provide additional context to these estimates by doing some back-of-the-envelope cost-benefit calculations. To calculate benefits, we use the estimated impact of teacher value-added on earnings from Chetty, Rockoff and Friedman (2011) and assume that the achievement gains induced by ASPIRE provide the same value. They estimate that a one standard deviation increase in teacher value-added corresponds to a 0.1 standard deviation increase in student achievement and $186 per-student in additional earnings. Using the fact that teachers in our data average between 76 and 91 students depending on the subject, we calculate returns of $4,200 per teacher for English and $17,200 for social studies, with math and science falling in-between. These benefits far exceed the costs of up to $1,950 per-teacher.[41]

6.5 Robustness Checks

As discussed in Section 5, 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 7, we present a series of robustness checks that shed further light on the validity of this assumption.

First, we control for the number of students each teacher teaches in Panel [1]. 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

[41]The cost estimate is based on the 2009-10 payment scale with 25% of teachers receiving the median award (max $3,500) and 25% receiving the top-quartile award (max $7,000). We assume that each teacher receives the 10% attendance bonus and that average department size is the same for winners and losers.

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 [4], 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 [5], we add 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 [6] and [7], we drop all special education and LEP students, respectively, and find results similar to baseline.[42]

7 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, 2013; Goodman and Turner, 2013; Fryer et al., 2012; 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 examination of 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

[42]In Online Appendix Table A-1 we 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 teacher from spring of the prior year and from fall of the current year instead of matching to the spring and fall of current year teacher. We further estimate models only matching student to their fall of current year teachers, that use 2002-03 scores as the lagged score for all observations, and that are unweighted. All these results are similar to baseline.

respond to free-riding, award salience and collaboration/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 estimate difference-in-difference models that identify 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 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. 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 indicative of substantial free-riding or response to award salience when teachers are responsible for small portions of the relevant student population.

We also calculate lower-bound estimates of the effect of ASPIRE on average student test scores. We find the program had a large, positive effect on student performance in all subjects, with test scores increasing by between 3 and 10 percent of a standard deviation depending on the subject. Through a back-of-the-envelope calculation based of estimated returns to teacher value-added in Chetty, Friedman and Rockoff (2011), we show that the benefits from the department-based incentive portion of ASPIRE far exceed the costs. Thus, this program demonstrates that if incentives are structured properly, it is possible not only to raise student achievement but also to do so in a cost-beneficial manner.

Our results suggest strongly 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. Thus, ASPIRE could have been even more productivity enhancing than it was if it had used groups that minimized the number of teachers with shares below 0.2. Nonetheless, we acknowledge that it is unclear if our results would carry over to cases where the groups are not "natural," i.e. they are not based on standard groupings like subject-grade, or school.

It is tempting to extrapolate from our results that individual-based incentive awards should generate large positive effects on achievement, which is at odds with the findings from existing experimental research (Fryer, et. al. 2012; Springer et al., 2010). However, we believe such an extrapolation would be inappropriate, as this relies on out-of-sample predictions based on strong assumptions. In fact, there are a number of reasons to suspect the individual context is quite different from the group context, and thus our findings likely are not in conflict with the experiment-based individual awards results. First, individual awards may foster a competitive environment that is counter-productive if teaching efficiency benefits from cooperation and spillovers across teachers. Second, teacher value-added, on which most incentive awards are based, has the drawback of being very imprecise at the individual teacher level. This almost certainly mutes teacher responsiveness to awards based on these measures. Thus, there is an inherent tradeoff in the design of value-added based incentive pay systems between the precision of the value-added estimates and the potential for free riding in group-based incentive schemes. That we find teacher shares of 0.2-0.3 minimize free riding suggests that group-based incentives centered around small groups could provide a balance between the need for precision in the evaluation measure and the need to have sufficiently powerful incentives. Future work focusing on this tradeoff would be of high value in informing optimal teacher incentive pay design.

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., 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics* 31(2).

[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, 2013. "The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." *Journal of Labor Economics* 31(2).

[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] Imberman and Lovenheim, 2012. "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System." NBER Working Paper No. 18439.

[13] 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.

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

[15] 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.

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

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

[18] 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.

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

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

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

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

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

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

[25] 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.

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

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

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

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

[30] 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.

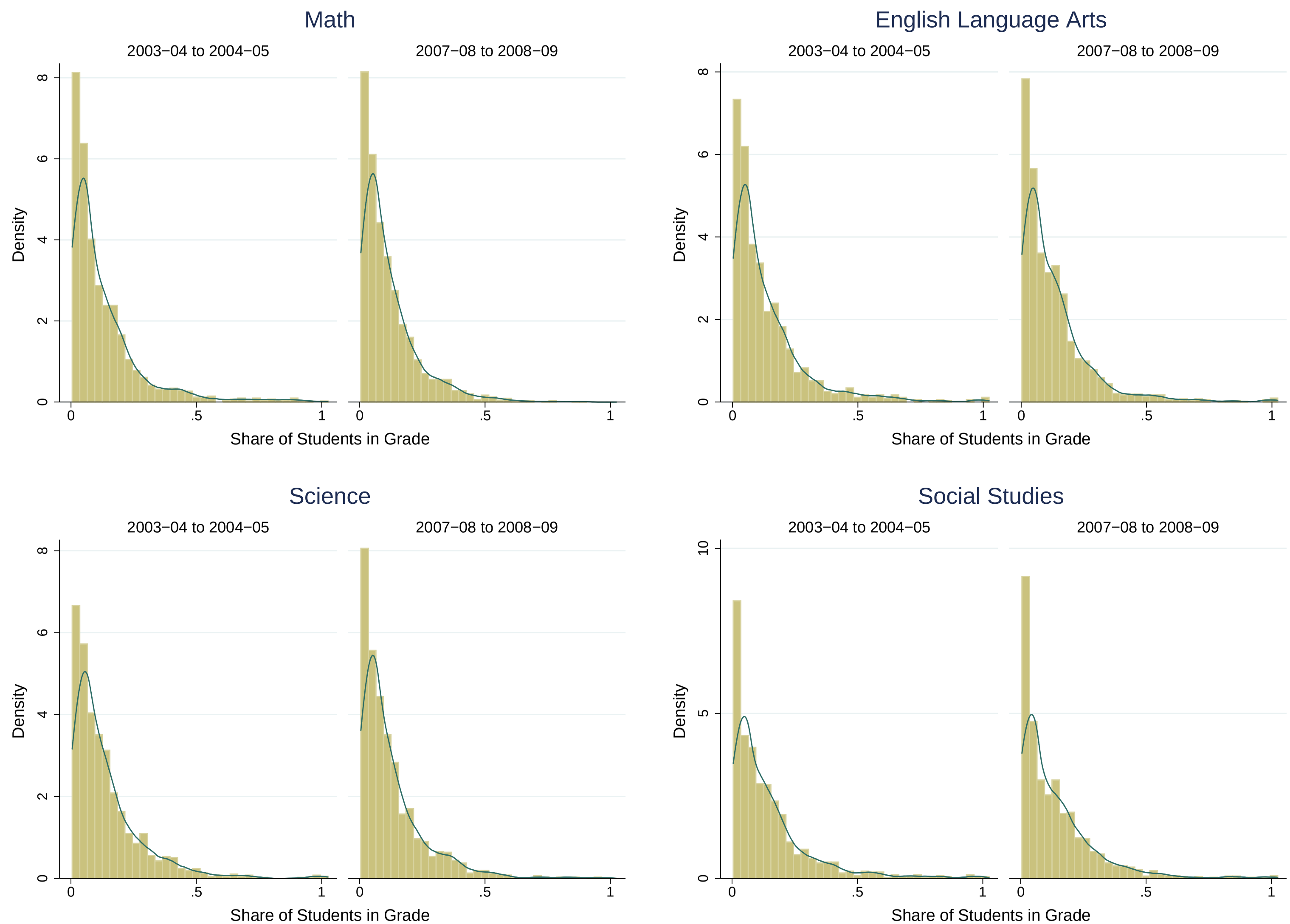
[31] Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper and Brian M. Stecher, 2010. "Teacher Pay For Performance: Experimental Evidence from the Project on Incentives in Teaching." National Center on Performance Incentives: http://www.performanceincentives.org/data/files/pages/POINT%20REPORT_9.21.1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

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

[33] 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.

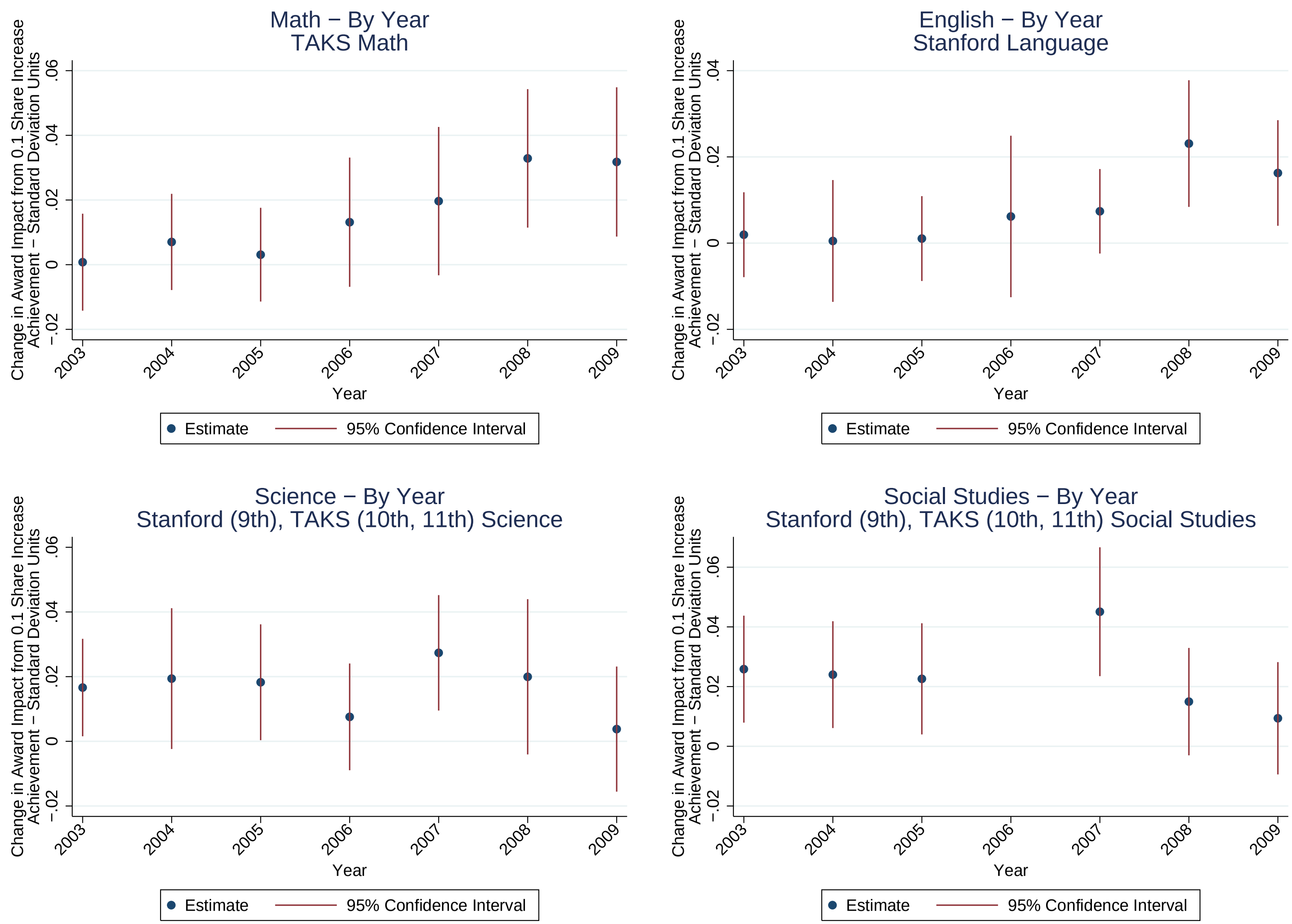
[34] Wright, S. Paul, John T. White, William L. Sanders, and June C. Rivers, 2010. "SAS EVAAS Statistical Models." Technical report. Available at http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf.

Figure 1: Distribution of Teacher Shares During Pre- and Post-Incentive Pay Periods



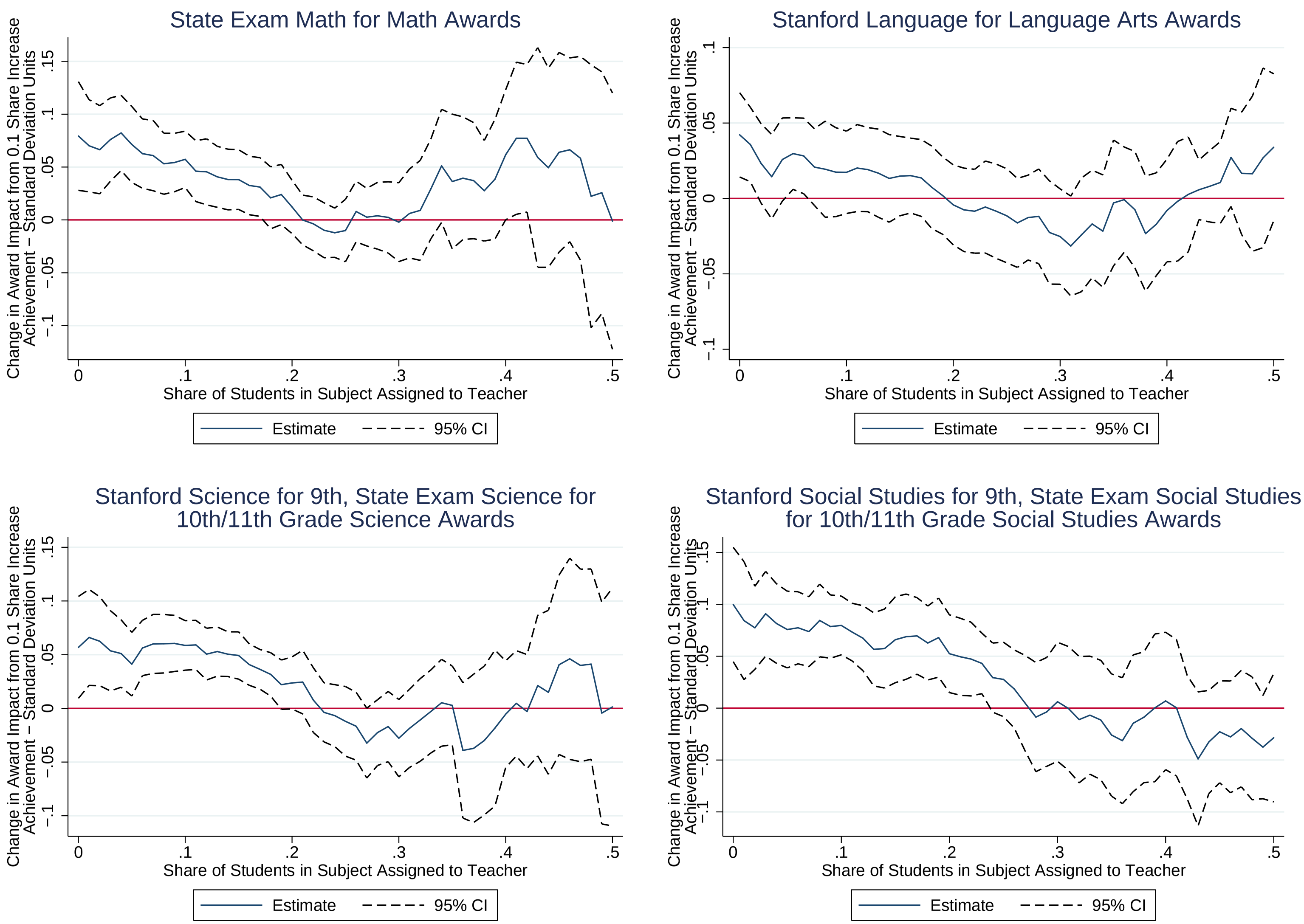
Graphs show distribution of unweighted teacher shares of students. The teacher is the unit of observation. Teachers with fewer than 10 students in a subject are dropped.

Figure 2: Effects of Teacher Share by Year



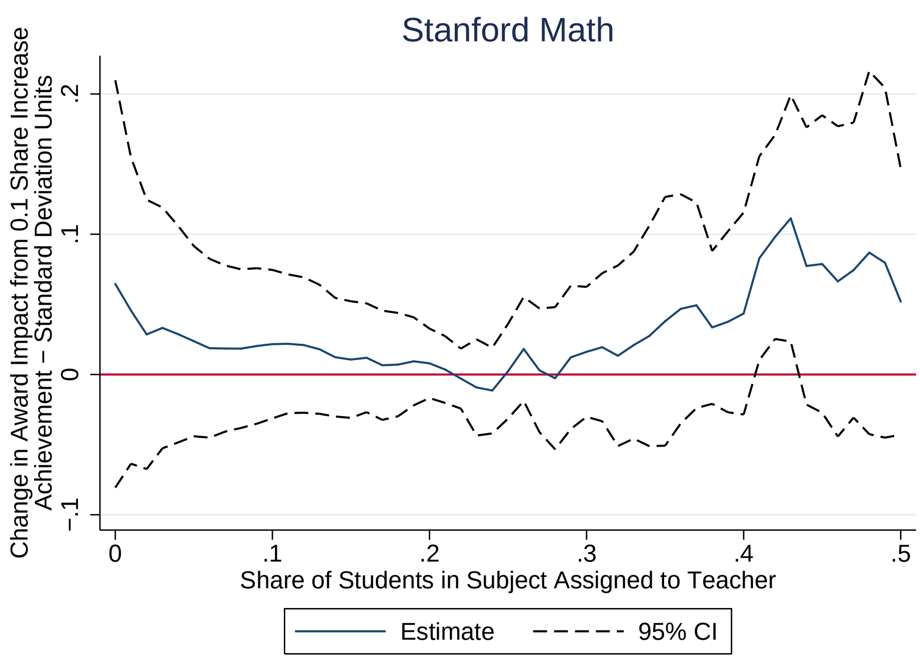
Data for social studies in 2006-07 are unavailable. Each point shows the average effect in a given year of raising the proportion of students a teacher is responsible for by 0.1 on standardized student test scores. These estimates come from models that include school-year and grade-year fixed effects as well as controls for 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 3: 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 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.

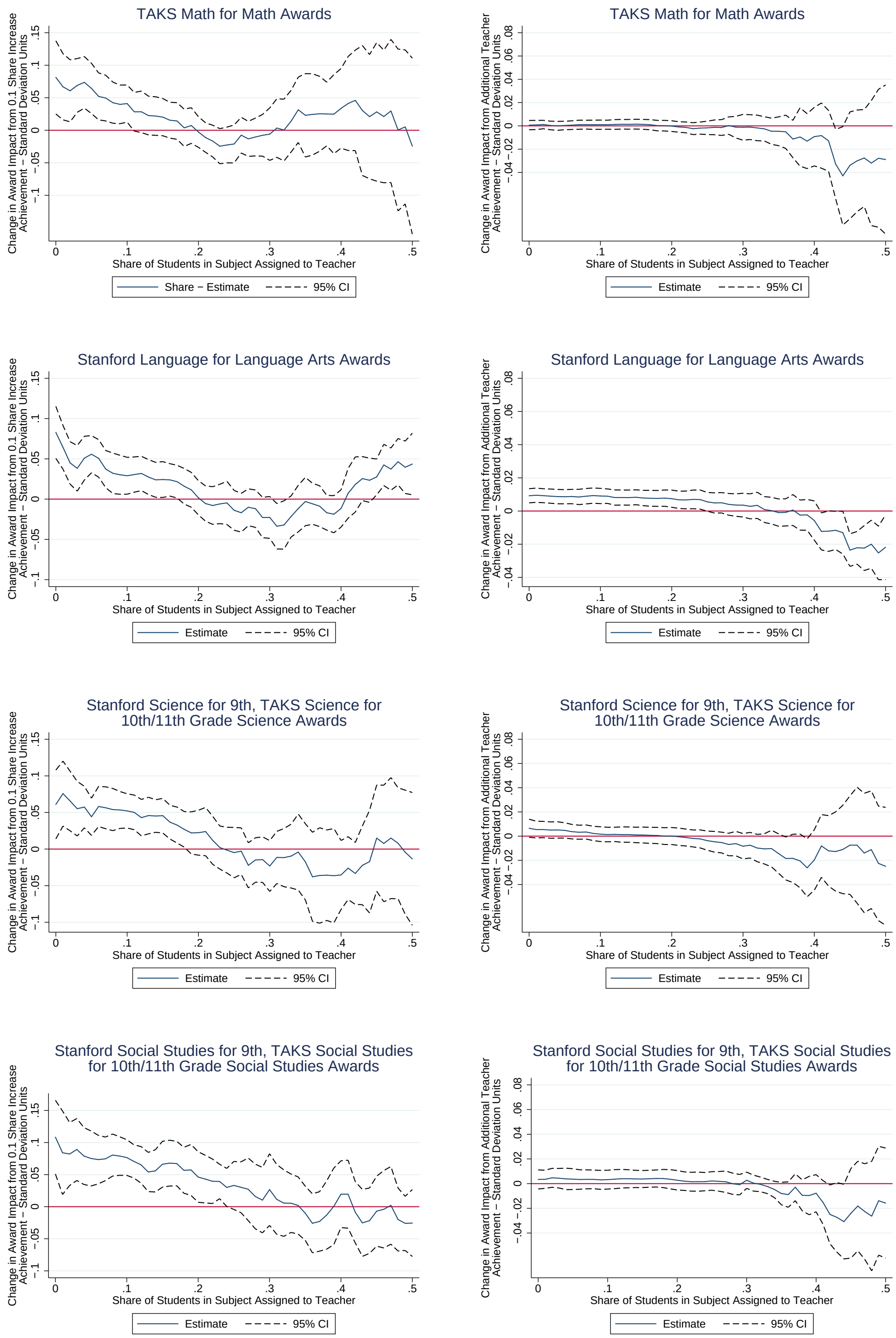
Figure 4: Local Linear Regressions of the Effect of Teacher Share Post-ASPIRE on the Non-Incentivized Math Exam



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 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 5: 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 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 9^{th} to 12^{th} 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	0.04	0.04	0.04
	(0.20)	(0.20)	(0.19)	(0.19)
Black	0.29	0.31	0.31	0.31
	(0.46)	(0.46)	(0.46)	(0.46)
Hispanic	0.55	0.53	0.54	0.54
	(0.50)	(0.50)	(0.50)	(0.50)
White	0.11	0.12	0.11	0.11
	(0.32)	(0.32)	(0.31)	(0.31)
Economically Disadvantaged	0.70	0.69	0.70	0.70
	(0.46)	(0.46)	(0.46)	(0.46)
At Risk	0.62	0.61	0.63	0.63
	(0.48)	(0.49)	(0.48)	(0.48)
Special Education	0.05	0.07	0.09	0.09
	(0.22)	(0.25)	(0.28)	(0.28)
Limited English Proficiency	0.07	0.03	0.07	0.07
	(0.17)	(0.18)	(0.26)	(0.26)
Gifted & Talented	0.17	0.17	0.16	0.16
	(0.38)	(0.38)	(0.37)	(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)	(0.14)	(0.13)	(0.15)
Department-Grade Size	13.6	15.4	11.9	12.2
	(6.8)	(7.7)	(5.3)	(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: Estimates of the Relationship Between Student Background Characteristics and a Teacher's Share Post-ASPIRE

	Test Subject:			
Independent Variable	Math	English & Language	Science	Social 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

† We use the most recent pre-program (2004 and earlier) lagged achievement. For math and English, we use the exam that determines award eligibility (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.

Notes: Each cell comes from a separate estimation of equation (2) and shows the estimate of α_2, which is the coefficient on $Share * Post$. Regressions also include school-year and grade-year fixed. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

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

Independent Variable	TAKS Math	English & Language	Science	Social Studies	Stanford Math
	Test Subject:				
Panel [1]: School & Grade-Year 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 [2]: [1] + 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)
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 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 are included in all specifications and include student gender, race, at-risk, special education, LEP, gifted status and lagged student test scores interacted with grade-by-year indicators. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

Table 5: Estimates of the Effect of a Teacher's Share Post-ASPIRE on Student Test Scores using Alternative Sources of Share Variation

Independent Variable	TAKS Math	English & Language	Science	Social Studies	Stanford Math
	Test Subject:				
Panel [1]: School-Grade-Year Fixed Effects:					
Post*Teacher Share	0.211**	0.125**	0.097	0.209**	0.033
	(0.097)	(0.058)	(0.097)	(0.091)	(0.095)
Teacher Share	0.052	0.073	0.116	0.279***	-0.179**
	(0.068)	(0.049)	(0.091)	(0.063)	(0.073)
Panel [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)
Panel [3]: 2SLS Using $\frac{1}{\text{2004 Dept. Size}}$ as Instrument for *Share*					
First Stage - Post $*$ Share:					
Post*$\frac{1}{\text{2004 Dept. Size}}$	1.09***	2.19***	1.51***	1.29***	1.09***
	(0.09)	(0.32)	(0.14)	(0.22)	(0.09)
First Stage - Share:					
$\frac{1}{\text{2004 Dept. Size}}$	0.76***	1.27***	1.10***	0.20	0.76***
	(0.11)	(0.36)	(0.33)	(0.20)	(0.10)
Second Stage:					
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)
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 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. All models include grade-year fixed-effects and the following individual controls: student gender, race, at-risk, special education, LEP, gifted status and lagged student test scores interacted with grade-by-year indicators. Panels [1] and [2] also include school fixed effects. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

Table 6: Calculations of Average Impact of Group Incentive Pay in HISD

	A. Based on Pooled Difference-in-Differences Estimates			B. Based on Local-Linear Difference-in-Differences Estimates		
	Estimate	Upper Bound	Lower Bound	Estimate	Upper Bound	Lower Bound
	(1)	(2)	(3)	(4)	(5)	(6)
Math (TAKS)	0.044	0.077	0.012	0.096	0.168	0.024
English (Stanford Language)	0.031	0.053	0.010	0.030	0.084	-0.024
Science (Stanford & TAKS)	-0.002	0.036	-0.040	0.072	0.134	0.009
Social Studies (Stanford & Taks)	0.049	0.099	-0.001	0.104	0.238	-0.030
Math (Stanford)	0.006	0.036	-0.024	0.014	0.098	-0.069

Local-linear based calculations integrate the total effect of incentive pay for each teacher from the set of marginal effects of increased share from the local linear regressions shown in Figure 5. Pooled OLS based calculations take the estimates from Panel [2] of Table (4) and multiplies them by the teacher's share in that subject, grade and year. The estimated total effects are then averaged across all students in grades 9 - 11 from 2007-08 through 2009-10, weighted by the inverse of the number of teachers each student has in a subject so that each student counts only once. Columns (1) and (4) uses the point estimates from the regressions as the estimated marginal effect. Columns (2) and (5) use the upper bound of the 95% confidence interval from those regressions while columns (3) and (6) use the lower bound of the 95% confidence intervals.

Table 7: Robustness Checks

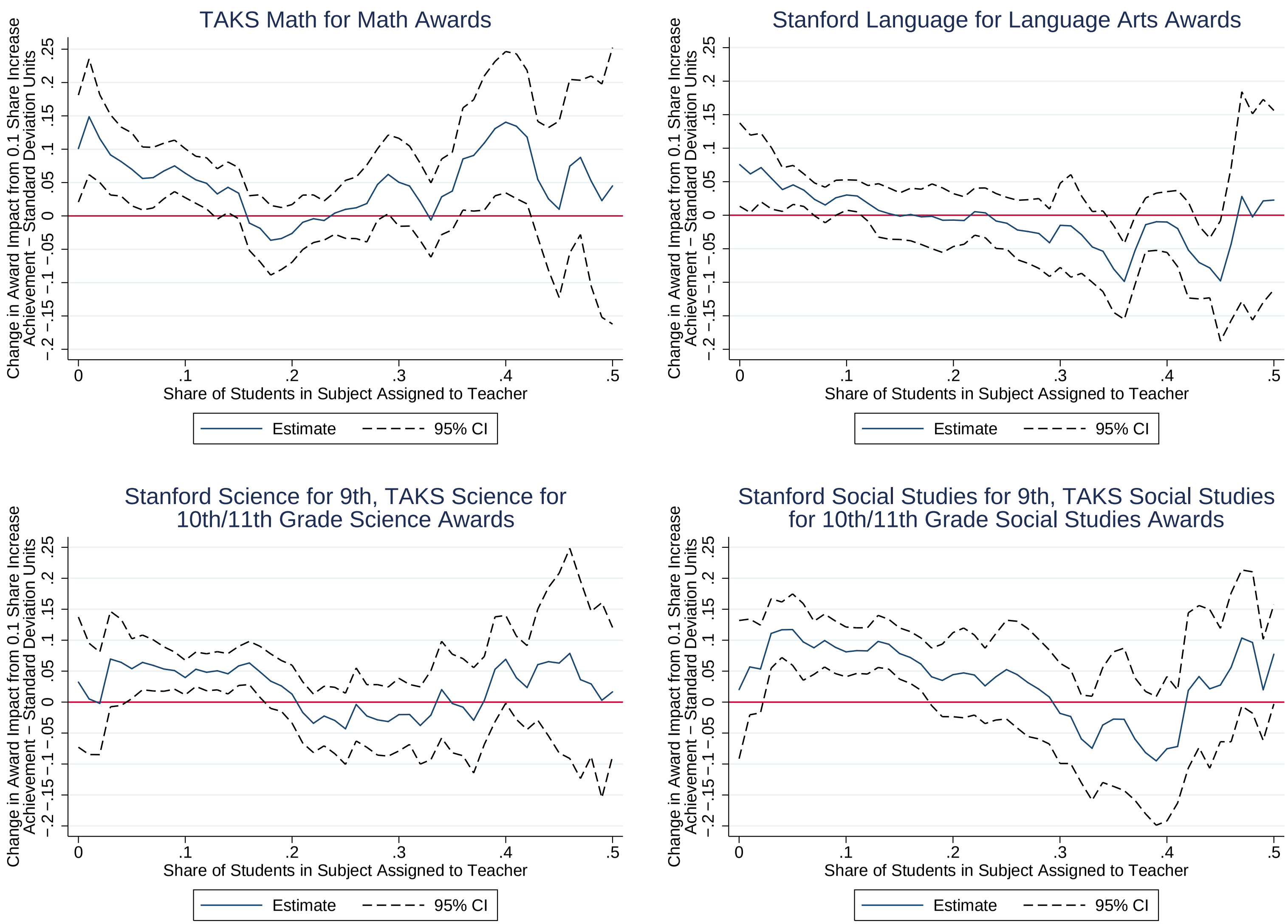
Independent Variable	TAKS Math	English & Language	Science	Social Studies	Stanford Math
			Test Subject:		
[1] Control for # of Students Each Teacher Has:					
Post*Teacher Share	0.236**	0.117**	-0.010	0.179*	0.036
	(0.087)	(0.050)	(0.092)	(0.103)	(0.081)
Observations	241,694	224,044	240,472	242,001	239,350
[2] 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
[3] 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
[4] 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
[5] 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
[6] 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
[7] 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.

Online Appendix

Not for Publication

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 lagged student test scores and student demographics. 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 Robustness Checks

	Test Subject:				
Independent Variable	TAKS Math	English & Language	Science	Social Studies	Stanford 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*10^{th}	-0.075	-0.029	0.319	0.162	0.087
	(0.182)	(0.052)	(0.213)	(0.247)	(0.078)
Post*Teacher Share*11^{th}	-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

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.

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

	Test Subject:			
	Math	English & Language	Science	Social 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: Additional Tests of "Impact" of Teacher Share on Student and Teacher Characteristics

Independent Variable	Math	English & Language	Science	Social Studies
	Test Subject:			
Student is New to School (Grades 10, 11 Only)	-0.046*	-0.008	-0.018	0.034
	(0.026)	(0.052)	(0.030)	(0.034)
Observations	144,955	134,576	143,319	145,723
Student was Not Enrolled in District in Prior Year †	0.002	-0.021	-0.030**	-0.015
	(0.009)	(0.015)	(0.012)	(0.015)
Observations	241,694	224,964	240,544	242,196
Number of Courses Taught	1.113	2.463*	-0.149	0.797
	(1.036)	(1.300)	(0.931)	(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 (2) 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. Standard errors clustered at the school level are in parentheses: ***,**,* indicates statistical significance at the 1%, 5% and 10% levels, respectively.

Table A-4: Tests of Heterogeneous Treatment Effects 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)					
χ^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, grade-year and school-year fixed effects. Standard errors clustered at the school level are in parentheses: **,* indicates statistical significance at the 5% and 10% levels, respectively.