NBER WORKING PAPER SERIES

ACHIEVEMENT EFFECTS OF INDIVIDUAL PERFORMANCE INCENTIVES IN A TEACHER MERIT PAY TOURNAMENT

Margaret Brehm
Scott A. Imberman
Michael F. Lovenheim

Working Paper 21598
http://www.nber.org/papers/w21598

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2015

We would like to thank Gary Solon, Jeffrey Wooldridge and seminar participants at CESifo Economics of Education Conference, Michigan State University, University of Arkansas, University of Michigan, and the APPAM Fall Conference for helpful comments and suggestions. This work was generously supported by the National Academy of Education and the Spencer Foundation, and much of this work was completed while Lovenheim was a NAEd/Spencer Postdoctoral fellow. The research reported here was also supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B090011 to Michigan State University. This grant supports MSU's doctoral training program in the economics of education. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We are grateful for the support. We would also like to thank the employees at Houston Independent School District for their help and assistance. This work was done independent of the Houston Independent School District and does not reflect the views of the district or its employees. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Achievement Effects of Individual Performance Incentives in a Teacher Merit Pay Tournament
Margaret Brehm, Scott A. Imberman, and Michael F. Lovenheim
NBER Working Paper No. 21598
September 2015
JEL No. H75,I21,J33,J38

ABSTRACT

This paper estimates the effect of the individual incentives teachers face in a teacher-based value-added merit pay tournament on student achievement. We first build an illustrative model in which teachers use proximity to an award threshold to update their information about their own ability, which informs their expected marginal return to effort. The model predicts that those who are closer to an award cutoff in a given year will increase effort and thus will have higher achievement gains in the subsequent year. However, if value-added scores are too noisy, teachers will not respond. Using administrative teacher-student linked data, we test this prediction employing a method akin to the bunching estimator of Saez (2010). Specifically, we examine whether teachers who are proximal to a cutoff in one year exhibit excess gains in test score growth in the next year. Our results show consistent evidence that teachers do not respond to the incentives they face under this program. In line with our model, we argue that a likely reason for the lack of responsiveness is that the value-added measures used to determine awards were too noisy to provide informative feedback about one's ability. This highlights the importance of value-added precision in the design of incentive pay systems.

Margaret Brehm
Michigan State University
486 W Circle Dr.
110 Marshall-Adams Hall
East Lansing, MI 48824-1038
orourk49@msu.edu

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

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

1 Introduction

The contribution of teachers to student educational achievement has been demonstrated repeatedly by researchers (Rivkin, Hanushek and Kain, 2005; Rockoff, 2004; Kane and Staiger, 2008; Chetty, Friedman and Rockoff, 2014). Consequently, school districts and states increasingly have used teacher incentive pay programs as a tool for improving student performance. The motivation for these programs is to provide monetary incentives for teachers to increase effort and to try new strategies that lead to higher measured achievement. As of 1993, over 12 percent of teachers were covered by a merit pay system (Ballou, 2001). This percentage has increased substantially in recent years as large school districts, such as Denver, CO, Houston, TX, and Minneapolis, MN as well as the states of Florida, North Carolina and Tennessee have implemented such systems. Incentives for teacher performance also play a large role in the Obama Administration's "Race to the Top" Initiative, suggesting that such programs are likely to be an important aspect of education policy in the near future.

While there is some mixed evidence on the overall impact of incentive pay systems,[1] little is known about how particular aspects of incentive design affect student achievement. One of the design features that has received the most attention in the literature is the size of the group used to measure performance. At one extreme, many incentive pay programs, particularly in the US, are group-based at the school or grade level. In these systems, the average performance of the school or grade is used to make award determinations. At the other extreme are individual-based incentive pay programs in which teachers are given awards based only on the performance of their own students. There is growing evidence that basing awards on large groups reduces the effectiveness of merit pay due to the free-rider problem (Muralidharan and Sundararaman 2011; Goodman and Turner 2013; Imberman and Lovenheim 2015). The existence of substantial free-rider effects suggests that individual-based incentive pay systems may be more effective.

Prior work on how individual incentive pay affects student outcomes has not come to a consistent conclusion, however. Quasi-experimental estimates of systems in which at least some of the incentives are at the individual level find a mix of positive and null results (Dee

[1] See Neal (2011) for an overview of the literature.

and Wyckoff, 2013; Sojourner, Mykerezi and West, 2014; Lavy, 2009; and Dee and Keys, 2004). The clearest evidence on the overall effect of individual teacher incentive pay in the United States comes from a randomized controlled trial done in the Metro Nashville Public Schools (Springer et al. 2010).[2] Teachers were randomly assigned to be eligible for merit pay, which was awarded based on their value-added rank. The study finds no evidence that students assigned to treated teachers exhibited higher test score growth.[3]

These studies focus on estimating average effects of exposure to a teacher incentive pay program. While clearly a parameter of interest, average effects may miss important variation across teachers in the specific incentives they face under these programs that are driven by performance feedback. Individual incentives necessarily give feedback to teachers on their performance, particularly relative to other teachers. In a system with sharp award thresholds, this feedback allows teachers to update their beliefs about their likely proximity to an award threshold: those who are close to the award thresholds face stronger incentives to change their behavior than those far away, who are unlikely to affect their award determination through changes in effort.[4] Average estimates of individual incentive pay likely do not accurately characterize program effects because they ignore the potentially large variation in incentive strength across teachers.

In order for performance feedback to operate as a mechanism through which individual teacher incentive pay influences the effort levels of certain teachers, the feedback needs to be informative. That is, the performance measure needs to give teachers accurate information about their relative productivity. This can be a particular problem when value-added models are used as the performance measure, which is increasingly the case in individual incentive pay systems. Value-added estimates have been show to exhibit substantial noise when using one year of data for an individual teacher (Guarino et al., 2015).[5] This is a design feature of

[2] International experimental evidence has found positive effects in India (Muralidharan and Sundararaman, 2011), but the Indian education system is starkly different from systems in developed countries, making it difficult to compare results.

[3] A smaller scale experiment conducted in Chicago Heights (Fryer et al., 2012) finds similarly little impact. Intriguingly, Fryer et al. (2012) do find large impacts from incentives set up to take advantage of loss aversion.

[4] This feedback mechanism also can impact teacher performance by providing information on the effectiveness of any responses teachers are making to the incentive pay program.

[5] While it is standard practice to calculate teacher value-added using multiple years of data for evaluation purposes, in a teacher incentive pay system this is very uncommon because typically the goal of these systems

teacher incentive pay programs and of value-added models more generally that has received very little attention in the literature. Thus, especially for lower-grade teachers who teach a small number of students, it may not be feasible to obtain precise estimates of teacher value-added for individual incentive pay systems to impact teacher behavior (and thus student achievement).

In this paper, we examine whether the specific incentives teachers face under an individual-based, value-added rank-order tournament affects measured student performance. We begin by specifying an illustrative theoretical model that highlights the role of performance feedback and measurement precision in driving heterogeneous teacher responses to merit pay. In the model, teachers are unsure of their ability and use their proximity to an award cutoff in the current year to inform their likelihood of being proximal in the next year's tournament. We show that teachers close to award thresholds should be the most responsive to the incentive pay system but that this responsiveness declines with the amount of noise in the value-added estimates. Intuitively, if the value-added measure is simply random noise, it tells teachers nothing about their actual ability and they should disregard it when making current effort decisions. The theoretical model contributes to our understanding of incentive pay design by highlighting the role measurement precision plays in driving teacher responses to incentives when teachers use the achievement measures as feedback to update their beliefs about their own ability. To our knowledge, this paper is the first to highlight these important design considerations related to individual teacher incentive pay programs.

We then take the central prediction of the model to the data: teachers who are closer to award thresholds in the current year should exhibit higher student growth in the subsequent year because their expected marginal return to effort is higher. This is the first study to look at this type of heterogeneous response by teachers to incentive pay.[6] Our empirical analysis uses 3^{rd}-8^{th} grade data from the Houston Independent School District's (HISD) incentive pay system, called ASPIRE (Accelerating Student Progress, Increasing Results and Expectations). The ASPIRE program is a rank-order tournament based on teachers' value-added scores on

is to reward teachers for their current year's performance.

[6] A number of papers have looked at a similar phenomenon with regards to school accountability systems. They have found mixed evidence that teachers and schools focus on students who are close to passing thresholds (called 'triaging') on exams (Ladd and Lauen 2010; Neal and Schanzenbach 2010; Reback 2008).

subject-specific standardized tests in mathematics, reading, language arts, science and social studies. In the rank-order tournament, teachers in the top quartile of the district-wide value-added distribution for each subject earn the top award amount, teachers in the 50^{th} to 75^{th} percentiles earn a lower amount, and teachers who have below-median value-added scores do not earn an award for that subject. The award amounts are substantial: in 2008-2009, the maximum top quartile award was $7,700 and the maximum award for being above the median was $3,850.[7]

Our empirical approach uses a variation on a technique developed in Saez (2010) to test for "bunching" near a tax notch (i.e., a change in the marginal tax rate) to estimate the elasticity of taxable income.[8] The idea behind this estimator is that there is some counterfactual relationship between current value-added rank and future test score performance that, in the absence of the award incentives, should be smooth across the current value-added rank distribution. However, in the presence of the award incentives, test scores should be higher than one would predict using this counterfactual relationship in the area surrounding the award threshold. We therefore predict future test scores as a function of a flexible polynomial in the distance to an award cutoff in the current year and student characteristics, excluding teachers with value-added scores close to the cutoff. We then use model estimates to generate predictions of what the achievement of students who had the excluded teachers *would have been* in the absence of award incentives in the threshold area; the difference between the actual and predicted test scores within this area identifies the effect of the individual incentives on teachers who are proximal to an award threshold.[9]

Our results indicate that teachers do not respond to the individual incentives imbedded within the ASPIRE incentive pay system. Teachers near award thresholds in one year do not exhibit excess growth in test scores in the following year, and this null result is robust to changes

[7] These awards include a 10% bonus for perfect teacher attendance.

[8] This method has been used subsequently by several researchers to estimate tax responsiveness (Kleven et al. 2011; Sallee and Slemrod 2012; Chetty, Friedman and Saez 2013; Kleven and Waseem 2013; Bastani and Selin 2014). To our knowledge, we are the first to apply this technique to the study of education in general and teacher incentive pay in particular.

[9] This is different from the Saez (2010) method as he tries to estimate deviations from the predicted values in terms of the density of the tax distribution - how many taxpayers there are close to the cutoff. In our context, the outcome is not whether there are more teachers near a cutoff but rather whether the actions of teachers who are near the cutoff result in higher performance than expected.

in the bandwidth that defines which teachers are proximal to an award as well as to the type of polynomial used to estimate the counterfactual. We also do not find any evidence of an effect for teachers with different characteristics, such as experience and race/ethnicity. These findings are surprising given the clear predictions of the model and the fact that Imberman and Lovenheim (2015) find sizable effects among high school teachers who face a similar program but at the group level. We argue that these results can be reconciled by the fact that the individual value-added estimates are very noisy. Indeed, we show that the year-to-year transitions in value-added rank in the areas surrounding the cutoffs is essentially uniformly distributed. This implies that value-added estimates in one year tell teachers very little about their relative ability, and as our model predicts they ignore the information.

Hence, our preferred interpretation of the evidence is not that individual incentives are ineffective, *per se*, but when performance measures contain too much noise to provide meaningful feedback to teachers, responses will be small. This paper highlights the roles of teacher feedback and measurement precision as key design features in an individual incentive pay program that have not received attention in prior research. Our results therefore provide important guidance for the optimal design of teacher incentive pay programs. Nonetheless, we caution that we cannot identify the *overall* impact of the incentive program, and it remains possible that teachers respond homogenously across the value-added distribution in a way that leads to higher overall performance.

The rest of this paper is organized as follows: Section 2 describes the previous literature. Section 3 provides a theoretical model of how teachers may respond to tournament based incentive pay. Section 4 describes the ASPIRE program, section 5 provides a description of the data, and the estimation strategies are presented in Section 6. Section 7 shows our results, and Section 8 concludes.

2 The ASPIRE Incentive Pay Program

The ASPIRE program is a rank-order tournament based on teachers' value-added scores on subject-specific standardized tests in math, reading, language arts, science and social studies that was implemented by HISD in 2006-2007. We focus on the individual teacher based tournaments that are used for 3^{rd} through 8^{th} grade teachers who teach core subjects - math, reading, English, science and social studies.[10] Each teacher who teaches a 3^{rd} through 8^{th} grade self-contained (e.g. a single teacher for all major subjects) class is entered into a rank-order value-added tournament for each grade and subject she teaches. Thus, teachers can win a separate award for every individual tournament in which they are entered. This feature of the program means that self-contained classroom teachers in elementary schools typically compete in all five subject-specific tournaments in the grade in which they teach. Other teachers are "departmentalized," meaning they teach only certain subjects to students. They are entered into tournaments with all other departmentalized teachers in the same subject and school type. For example, a departmentalized math teacher in an elementary school (grades 3-5) would compete against all other departmentalized math teachers who teach in an elementary school in Houston. Departmentalized math teachers in middle school (grades 6-8) are entered into a separate tournament.

Award determinations are based on each teacher's place in the subject-specific and, for self-contained classroom teachers, grade-specific value-added distribution.[11] Value-added is calculated for each teacher using the Education Value-Added Assessment System (EVAAS) by the SAS Corporation. 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). Teacher value-added is calculated by estimating a linear mixed model that regresses student achievement on a set of weighted indicators for each teacher the student had in the current and prior two years along with indicators for subject, grade and year. No other controls are included. By construction, the model shrinks estimates so that those with fewer observations are attenuated towards the grade-subject-year mean.

While in this study we use the student level test scores as our outcome measure, it nonethe-

[10] High school teachers compete in grade-school-subject groups, and these tournaments are studied in Imberman and Lovenheim (2015).

[11] Self-contained classroom teachers in grades 3-5 are compared to other self-contained teachers in the same grade, separately for each subject. Only elementary school teachers can be self-contained; all middle school teachers are departmentalized.

less is useful to understand how the value-added measure is calculated as this might inform the teachers' behavior. To more clearly see the implications of the model structure, Ballou, Sanders and Wright (2004) point out that the following system of equations must hold:

$$
\begin{aligned}
Y_0^g &= b_0^g + u_0^g + e_0^g \\
Y_1^{g+1} &= b_1^{g+1} + u_1^{g+1} + u_0^g + e_1^{g+1} \\
Y_2^{g+2} &= b_2^{g+2} + u_2^{g+2} + u_1^{g+1} + u_0^g + e_2^{g+2} \\
&\vdots \\
Y_t^{g+t} &= b_2^{g+2} + \sum_0^t (u_k^{g+k}) + e_t^{g+t} \qquad (1)
\end{aligned}
$$

where Y is achievement, b is the district-grade-year mean, u is a teacher effect, g is grade level of the student, t is time, and e is random error. Note that each achievement score incudes each prior teacher's effect on the student, and thus the teacher effects "stack" over time. By recursively solving backwards from period T, Ballou, Sanders and Wright show that the teacher effect is calculated as

$$
u_T^g = (Y_T^g - Y_{T-1}^{g-1}) + (b_T^g - b_{T-1}^{g-1}) + (e_T^g - e_{T-1}^{g-1}) \qquad (2)
$$

so that the teacher effect in grade g and year T is the sum of the annual change in the student's achievement, the district-wide achievement, and the student-specific error. In practice, the EVAAS system only includes a students' current and two prior teachers in the model, implicitly assuming that $u_{T-k}^{g-k} = 0 \ \forall k \geq 3$.

There are a couple of important limitations to this model. First, since in equation (2) there is no parameter on Y_{T-1}^{g-1}, the model does not allow for decay in student achievement. Second, Guarino, Reckase and Wooldridge (2014) show that models like this are more subject to bias than simpler lagged dependent variable models. Third, the lack of controls for student characteristics is controversial and may contribute to bias.[12]

To administer the incentive awards, the value-added measures produced by this model are

[12] See Ballou, Sanders and Wright (2004) for a discussion of this issue.

ranked within subject and grade for self-contained teachers or within subject and school type for departmentalized elementary school teachers and for middle school teachers. Each year there is a new set of tournaments, and the value-added analysis is redone incorporating the new data. Teachers who receive value-added scores greater than zero (indicating value-added greater than the district-wide grade-subject mean) and who are above the median value-added in the tournament receive an award. The award doubles if the teacher is within the top quartile of value-added. Table 1 provides details on the ASPIRE system for the 3^{rd} through 8^{th} grade tournaments.

For the time period of this study, awards were based on two types of exams, depending on the grade and subject. In math and reading, the Texas Assessment of Knowledge and Skills (TAKS) exams are used, which were state standardized tests that are also used for school accountability ratings. TAKS exams were used for science in 5^{th} and 8^{th} grades and for social studies in 8^{th} grade as well. Texas did not administer state standardized tests in science and social studies for other grades, nor does it test language arts. Thus, for these subjects and grades, award determination was based on Stanford Achievement Test scores. The Stanford test is a nationally-normed, grade-specific exam that is administered to students throughout the United States. Teacher value-added was then reported in terms of a normal-curve equivalent (NCE) of the state-wide distribution for TAKS and district-wide distribution for Stanford. For ease of comparison to other papers in the literature, we convert the NCE scores into student standard deviation units.

In addition to the teacher award, there are a series of awards for school-wide performance.[13] Each of these awards is relatively small, ranging from $150 to $750 apiece, hence we do not consider them in our analysis.[14] In 2006-07 and 2007-08, teachers could earn up to $5,500 from the individual awards. This amount includes a 10% attendance bonus that is given to

[13] Each year there are four types of campus-wide awards for which teachers are eligible. 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.

[14] Principals and assistant principals also were given awards for school-wide performance. The incentives under these awards were only partially aligned with those of teachers, as they were based on performance by the entire school in each subject, not by each teacher and subject. Importantly, the incentives of principals and vice principals are unlikely to be affected by whether an individual teacher in their school is marginal for an award.

teachers who take no sick days during the year. The maximum total award a teacher could receive was \$8,030. In 2008-09, HISD increased award amounts substantially. The maximum award jumped to \$7,700, with a total maximum award of \$11,330. Given that the base salary for a new teacher with a bachelor's degree in that year was \$44,027, up to 20% of a teacher's total wage compensation was determined by performance bonuses, with up to 14% from the individual award portion. The average award across all core teachers in HISD (including high schools) was \$3,614 in 2009-10. Thus, the large bonus amounts relative to base pay suggests that the awards were highly salient and desired by teachers.

One potential problem with the ASPIRE program is that the use of the EVAAS value-added methodology for determining award receipt might make the award formula complex and difficult for teachers to understand. We note that for us to see a response we need only that the teachers know their approximate position in the distribution and that they understand that increasing the test scores of their students will increase their award likelihood. There is some evidence from surveys conducted by HISD that indicate teachers were well informed and had a good understanding of the system.[15] If becoming marginal for an award induces teachers to learn more about the system, this could be a mechanism underlying our results rather than a potential threat to identification.

As detailed in Table 1, the system is designed such that, within a given year, each teacher is eligible for the same maximum award amount. Since teachers who instruct in multiple core subjects (e.g., self-contained teachers) are entered in multiple tournaments, the award amounts for a given step in the award function are reduced so that they are proportional to the inverse of the number of subjects taught. Thus, the teacher's total award increases when she exceeds a threshold as follows:

$$AwardIncrement_{jt} = \frac{MaxAward_t}{2 \times Subjects_{jt}} \tag{3}$$

For example, in 2009-10 when the maximum individual award, excluding the attendance bonus, is \$7,000, a teacher with one course earns \$3,500 for winning the median award and that increases by another \$3,500 if she wins the top quartile award. On the other hand, a teacher

[15] The survey results can be found at *http://www.houstonisd.org/portal/site/researchaccountability*. See Imberman and Lovenheim (2015) for a discussion of this survey.

who teaches all 5 subjects - which includes all self-contained teachers in elementary schools - receives awards in $700 increments. Thus, over the study period, the marginal value of winning a given award ranges from $500 to $3,500. Figure 1 shows how the average award for a teacher changes as a teacher exceeds an award threshold in a given subject (units on the horizontal axis are standard deviations of the teacher value-added distribution). A teacher who exceeds a median award threshold in a given subject receives an incremental bonus that averages approximately $1,000 while teachers who exceed the top quartile award threshold average approximately $1,500.

3 Theoretical Model

In order to generate predictions about how teacher effort, and thus student performance, should change with a teacher's proximity to an award threshold in the prior year, we develop a model of teacher behavior under an individual incentive pay scheme. The three core components of the model are: 1) teachers differ in their ability, a, which will make some teachers more likely to be close to award margins than others, 2) teachers have imperfect knowledge of their own ability, and 3) teachers do not know exactly where next year's award cutoff will be. The second and third components generate uncertainty, as teachers cannot perfectly predict how close they are to an award cutoff in a given year. Furthermore, this setup highlights the role of performance pay systems in providing teachers feedback about their ability, which has received little attention in prior work.

Each teacher has an exogenous ability level, a, and provides effort, e, which are translated into value-added: $V = V(e, a)$. We assume higher effort and ability both lead to higher value-added ($\frac{\partial V}{\partial e} > 0$, $\frac{\partial V}{\partial a} > 0$), the marginal benefit of effort declines with effort ($\frac{\partial^2 V}{\partial e^2} < 0$) and higher-ability teachers have a higher return to effort ($\frac{\partial^2 V}{\partial e \partial a} > 0$). The cost of teacher effort is given by $C = C(e)$, with $C'(e) > 0$ and $C''(e) > 0$. Student test scores also are a function of teacher effort and ability, but the function that maps effort and ability into test scores ($S = S(e, a)$) is allowed to be different than the mapping for value-added.[16]

[16] Although the function form is allowed to differ, we still assume that ($\frac{\partial S}{\partial e} > 0$ and $\frac{\partial S}{\partial a} > 0$).

We focus on a two-period model in which teachers do not have full information about their own ability. However, they have some prior belief about their ability relative to other teachers that they update when they receive new information. Both time periods are assumed to have an incentive pay system in which there is a sharp cutoff based on value-added for award eligibility. That is, in each period, teachers compete in a rank-order tournament in which they win award A if their value-added is above some exogenous threshold, V_t^*, $t \in \{1, 2\}$. We assume the award threshold is exogenous for simplicity, and we note that with several hundred teachers competing in each award tournament, the effects of strategic interactions are likely to be very small while increasing the complexity of the model dramatically. Instead, we model V_t^* as a random variable that represents teachers' beliefs over the location of the award cutoff.

In the first period, teachers have some belief about their ability, a_1, which we assume is exogenously assigned from some common distribution across teachers. Teachers exert effort in period 1, e_1, and based on that effort they receive an award if value added is greater than the award threshold, V_1^*. Then, based on the realization of value-added in period 1, teachers update their beliefs about their ability in the following manner:

$$a_2 = a_1 + \lambda(V_1 - a_1), \text{where } \lambda \in [0, 1]. \tag{4}$$

Equation (4) states that teachers will update their beliefs about their ability using the distance between their period 1 outcome and their prior belief. The parameter λ represents the weight teachers place on the new information. It can be thought of as the signal-to-noise ratio of V_1 or the reliability of the signal about ability contained in the first period value-added outcome. If $\lambda = 0$, then teachers place no weight on the prior year's score, while if $\lambda = 1$, they completely replace their prior belief about their ability with the prior year's value-added estimate.

The perceived probability a teacher wins award A in period 2 based on her information set from period 1 is given by:

$$\begin{aligned} P(A) &= P(V(e_2, a_2) > V_2^*) \\ &= P(V(e_2, a_1 + \lambda(V_1 - a_1)) > V_2^*). \end{aligned} \tag{5}$$

Uncertainty in this model is driven by the fact that teachers do not know the exact cutoff in each period. We assume they have a belief over the distribution of the location of the cutoff that is normally distributed with a mean equal to the prior year's cutoff.[17] This ensures beliefs are distributed symmetrically around the prior year's cutoff, which should represent the best "guess" of where the current year's cutoff is. Let $V_2^* = V_1^* + \psi$, where ψ is a normally-distributed random variable with mean zero and variance σ^2. The likelihood of winning the award in period 2 is given by:

$$\begin{aligned} P(A) &= P(V(e_2, a_1 + \lambda(V_1 - a_1)) > V_1^* + \psi) \\ &= P(V(e_2, a_1 + \lambda(V_1 - a_1)) - V_1^* > \psi) \\ &\equiv \Phi\left(\frac{V(e_2, a_1 + \lambda(V_1 - a_1)) - V_1^*}{\sigma}\right). \end{aligned} \tag{6}$$

Letting $V_2 \equiv V(e_2, a_2)$, each teacher faces the same expected utility function for period 2, given by:

$$\begin{aligned} E(U_2) &= APr(V_2 - V_1^* > \psi) - C(e_2) \\ &= A\Phi\left(\frac{V_2 - V_1^*}{\sigma}\right) - C(e_2). \end{aligned} \tag{7}$$

Let $\widetilde{V_2} \equiv V_2 - V_1^*$. Then the associated first order condition is:

$$A\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{1}{\sigma}\frac{\partial V_2}{\partial e_2} - C'(e_2) = 0 \tag{8}$$

and the second order condition is:

$$\begin{aligned} & A\phi\left(\tfrac{\tilde{V_2}}{\sigma}\right)\tfrac{1}{\sigma}\tfrac{\partial^2 V_2}{\partial e_2^2} + A\phi'\left(\tfrac{\tilde{V_2}}{\sigma}\right)\tfrac{1}{\sigma}\tfrac{\partial V_2}{\partial e_2} - C''(e_2) \\ = \; & A\phi\left(\tfrac{\tilde{V_2}}{\sigma}\right)\tfrac{1}{\sigma}\tfrac{\partial^2 V_2}{\partial e_2^2} - A\tilde{V_2}\phi\left(\tfrac{\tilde{V_2}}{\sigma}\right)\tfrac{1}{\sigma^2}\tfrac{\partial V_2}{\partial e_2} - C''(e_2) < 0, \end{aligned} \tag{9}$$

[17] We model uncertainty in this way because it matches the setup of the ASPIRE program. However, the implications of the model are the same if there is a fixed value-added threshold that is known and one introduces uncertainty into the function that maps effort into value-added. Thus, the theoretical predictions of this model relate to a broader set of individual incentive pay programs with known, fixed cutoffs.

where the negative sign in equation (9) is assumed in order to ensure that the solution to the first order condition leads to a maximum.[18] The goal of the theoretical analysis is to understand how optimal effort changes when one is closer to or farther from the award threshold in the previous period. To illustrate the model's predictions for period 2 effort as a function of proximity to the cutoff in period 1, we calculate how optimal effort in period 2 responds to period 1 value-added. This is equivalent to how optimal effort in period 2 responds to distance to the period 1 cutoff because V_1^* is fixed and known in period 2. Applying the implicit function theorem to equation (8) evaluated at the actual value of V_1 and the associated value of e_2 that makes equation (8) hold (i.e., e_2^*) yields:

$$\begin{aligned}
\frac{\partial e_2}{\partial V_1} &= \frac{\lambda A \frac{1}{\sigma}\{\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{\partial^2 V_2}{\partial e_2 \partial a_2} + \phi'\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{\partial V_2}{\partial e_2}\}}{C''(e_2) - A\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{1}{\sigma}\frac{\partial^2 V_2}{\partial e_2^2} - A\phi'\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{1}{\sigma}\frac{\partial V_2}{\partial e_2}} \\
&= \frac{\lambda A \frac{1}{\sigma}\{\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{\partial^2 V_2}{\partial e_2 \partial a_2} - \frac{\tilde{V_2}}{\sigma}\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{\partial V_2}{\partial e_2}\}}{AC''(e_2) - \phi\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{1}{\sigma}\frac{\partial^2 V_2}{\partial e_2^2} + A\tilde{V_2}\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\frac{1}{\sigma^2}\frac{\partial V_2}{\partial e_2}} \\
&= \frac{\lambda A \frac{1}{\sigma}\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\{\frac{\partial^2 V_2}{\partial e_2 \partial a_2} - \frac{\tilde{V_2}}{\sigma}\frac{\partial V_2}{\partial e_2}\}}{C''(e_2) + A\frac{1}{\sigma}\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\{\frac{\partial^2 V_2}{\partial e_2^2} - A\frac{1}{\sigma}\tilde{V_2}\frac{\partial V_2}{\partial e_2}\}}
\end{aligned} \qquad (10)$$

By the second order condition assumption, the denominator in equation (10) is positive, but the equation cannot be signed in general. In order to develop the testable predictions from this model that will inform our empirical strategy, we focus on how optimal period 2 effort responds to changes in period 1 value-added around the period 1 cutoff. Equation (10) is maximized when $\tilde{V_2} = 0$, as at this point $\phi\left(\frac{\tilde{V_2}}{\sigma}\right)\{\frac{\partial^2 V_2}{\partial e_2 \partial a_2} - \frac{\tilde{V_2}}{\sigma}\frac{\partial V_2}{\partial e_2}\}$ attains its largest value. To see this, first note that $\phi\left(\frac{\tilde{V_2}}{\sigma}\right)$ is maximized when $\tilde{V_2} = 0$ since the mean of ψ is 0. Then consider the case where $\tilde{V_2} = \epsilon < 0$. In this case, $\phi(\epsilon) < \phi(0)$. Thus, a sufficient condition for $\tilde{V_2} = 0$ to maximize equation (10) is

$$\frac{\partial^2 V_2}{\partial e_2 \partial a_2} - \epsilon\frac{\partial V_2}{\partial e_2} < \frac{\partial^2 V_2}{\partial e_2 \partial a_2} - 0\frac{\partial V_2}{\partial e_2}, \qquad (11)$$

[18] The partial effects in equations (8) and (9) are with respect to V rather than $\tilde{V}$ because V_1^* is a constant in period 2.

which simplifies to

$$-\epsilon \frac{\partial V_2}{\partial e_2} < 0. \tag{12}$$

Equation (12) holds since $\frac{\partial V_2}{\partial e_2} < 0$ by assumption and $\epsilon < 0$, which implies $\frac{\partial e_2}{\partial V_1}(\epsilon) < \frac{\partial e_2}{\partial V_1}(0)$. On the other side of the cutoff, the same argument works in reverse with $\epsilon > 0$ as we move from $\tilde{V}_2 = \epsilon$ to $\tilde{V}_2 = 0$. Thus, when one moves away from the period 1 cutoff in either direction, optimal effort declines.

The intuition for this result is straightforward: on average, the cost of effort is unrelated to value-added but the expected return to effort is larger the closer one is to an award threshold. Thus, effort is maximized at the period 1 threshold and declines as one moves away from it in either direction. The empirical implication of this prediction is that period 2 effort, and hence period 2 test scores, should be largest for teachers who have a value-added score in a range surrounding the period 1 cutoff. Put differently, teachers who are close to a cutoff should be inclined to increase their effort more than teachers who are far from a cutoff. This prediction forms the basis for our empirical approach.

The parameter λ plays an important role in the predictions of this model as well. When $\lambda = 0$, equation (10) equals zero. As λ increases, equation (10) becomes larger in absolute value, reaching its maximum at $\lambda = 1$. The intuition for this result is that when λ is close to zero, prior year's value-added gives teachers a very noisy and uninformative signal as to what their relative ability is. At the limit, if value-added is just random noise, then λ will equal zero and teachers will be unresponsive in period 2 to information received in period 1. As the signal value of period 1 value-added grows (i.e., λ approaches 1), teachers respond more to the period 1 outcome. The implication of this prediction is that teachers can be unresponsive to the merit pay system if the assignment variable is too noisy to give teachers adequately-powered signals about how to respond to the monetary incentives they face. Nonetheless, we note that while teachers would not respond differentially across the distribution of prior value-added scores in this scenario, they could still respond positively (or negatively) to the incentive pay system as a whole.[19]

4 Data

The data for this analysis come from administrative student-teacher matched records that we obtained through a special agreement with HISD for school years 2006-2007 through 2009-2010. For each teacher, we have information on which subjects and grades were taught in each year, value-added in each subject and grade taught, school in which the teacher works, and demographic information such as race/ethnicity, gender, experience, and highest degree earned. The teachers are linked over time through an anonymized identification number. Student data includes test scores for the Texas Assessment of Knowledge and Skills (TAKS) and the Stanford Achievement Test (SAT). Four subjects are covered for each exam - math, reading, science and social studies. Stanford exams also include a language portion. We also have data on student demographic characteristics and we use gender, race/ethnicity and economic disadvantage as controls in some specifications.

Students and teachers are matched at the subject level so there are up to five records on each student-year. While the incentive system covers grades 3 through 8, we drop 3^{rd} grade students and teachers so as to ensure that we have access to a lagged test score for all students. In 2009, HISD changed how they placed middle school teachers into tournaments and so for that year we only include elementary teachers. Exams are standardized within the district by subject grade and year. In each subject and grade we use the exam that is used for the calculation of awards. For TAKS this includes all grades for math and reading, grades 5 and 8 for science, and grade 8 for social studies. The Stanford exam is used for language in all grades, science in grades 4, 6 and 7, and grades 4 through 7 in social studies.

Table 2 provides means and standard deviations of the variables used in this study at the student-subject-year and teacher-subject-year levels. We also split the sample in Table 2 by the teacher's place in the award distribution to see how characteristics vary by award type. In general, higher-VA teachers have students with slightly higher socio-economic status and overall

[19] We lack a control group that would allow us to test for an overall effect of this program on teacher performance.

the district is highly economically disadvantaged (80%) and heavily minority (30% black and 60% Hispanic). Teachers who receive larger awards have higher achieving students, which is expected given current achievement is the key input into value-added. When we look at teacher characteristics, award receipt appears to have little relationship with teacher gender, education or experience. Mechanically, value-added scores also increase with higher award receipt and, as expected, we see evidence of mean reversion when looking at the change in award receipt. What is particularly interesting and relevant to our theoretical analysis is the standard deviations of both current value-added and the annual within-teacher change in value-added. These are further highlighted in Figure 2, where we plot the distribution of value-added scores and the change in value-added scores by teacher-subject-year. The standard deviation in value-added across the full sample is 1.96, while the standard deviation in the change in value-added is 2.12. We argue below that the breadth of these distributions provide suggestive evidence that the lack of response we find to the incentive pay system is likely due to the substantial noise in the value-added measures.

5 Empirical Models

Using the administrative teacher-subject-year level data described in the previous section, we test whether students assigned to teachers who are closer to an award cutoff in a given year have higher test score growth. To do this, we adapt the bunching estimator first used by Saez (2010) to estimate the elasticity of taxable income. The original application of this technique examines excess density around a change in the marginal tax rate. Instead, we estimate whether there is excess test score growth in an area surrounding the prior year's test score cutoff.

We first estimate a model that allows us to predict subsequent test scores as a function of flexible distance to an award cutoff in the current year, excluding those who are "local" to an award cutoff. Specifically, we estimate the following model:

$$S_{ijty+1} = \alpha + \beta S_{ijty} + g(VA_{jty} - VA^{*}_{ty}) + \delta X_{ijty} + \psi_y + \phi_t + \epsilon_{ijty}, \tag{13}$$

where S is the test score of student i assigned to teacher j in merit pay tournament t and year $y+1$. We control for current year's test score, year fixed effects (ψ_y), and tournament fixed effects (ϕ_t). Note that because all tournaments are subject specific, the tournament fixed effects also control for subject differences in the models in which we pool across subjects. Equation (13) also contains a vector of student demographic controls (X) that includes student race/ethnicity, grade fixed effects, gender, and whether the student is economically disadvantaged (i.e., eligible for free or reduced price lunch or another state or federal anti-poverty program). The term $g(V_{jty} - V^*_{ty})$ is a polynomial in the distance between the current period value-added and the award cutoff. We estimate this model separately for the distance to the median and 75^{th} percentile cutoffs.

Estimating equation (13) requires us to specify who is "local" to the cutoff, the order of the polynomial, and the estimation sample outside of the proximal cutoff area. Because there is no way to know *ex ante* how close to the cutoff teachers need to be to be affected by the ASPIRE incentives, we use a sequentially widening donut around each cutoff of 1.0, 1.5 and 2.0 value-added points. Thus, for example, we exclude all students with teachers who are within 1 value-added point of a cutoff (0.5 points on either side of the cutoff) and estimate equation (13). Recall that value-added is calculated in terms of student-level test score standard deviations, so this ostensibly excludes all students with teachers whose value-added places them within a half of a student standard deviation of an award cutoff. Widening the donut allows us to test whether our results are sensitive to which teachers we consider local to an award margin.

To determine the specific form of $g(\cdot)$, we use a Bayesian Information Criterion (BIC). The BIC chooses the polynomial that best fits the data with a penalty for adding extra terms. It is based on the Aikaike Information Criterion (AIC) but adjusts the penalty term for the number of observations. We thus choose the polynomial that minimizes the BIC, setting a maximum of six terms. Finally, we restrict the model to use observations that are either 2.5 value-added units above or below the cutoff (i.e., a range of 5 value-added points). We do this to avoid using observations that are very far away from award margins to make test score predictions. We will show later that estimates of equation (13) on this sample provide extremely accurate test score predictions outside of the local treatment area.

For our empirical test, we want to estimate whether students who have teachers with value-added inside the local award margin donut have higher than predicted test scores. We thus use equation (13) and predict test scores for each student located in the donut. We term this prediction $\hat{S}$. Then, we calculate the average difference between actual and predicted test scores for each student as our measure of the extent to which teachers respond to the incentives they face under the ASPIRE program:

$$\mathrm{TE} = \frac{1}{N_d} \sum_{i=1}^{N_d} S_{ijty+1} - \hat{S}_{ijty+1}, \tag{14}$$

where N_d is the number of students in the local treatment donut. We estimate equation (14) pooled across all subjects as well as separately by subject. In order to obtain standard errors for statistical inference, we block boostrap the entire estimation process (equations (13) and (14)) at the teacher-year level using 500 replications. The standard errors shown in the results below are the standard deviations of TE over these bootstrap replications.

The main identification assumption in our estimation procedure is that students are not sorting on unobservable attributes to teachers in a way that would place some students who would otherwise be outside of the donut into the donut and vice-versa. We view such sorting as unlikely because students and parents do not know where the new cutoff is. While there may be some sorting across the cutoff, as long as the sorting occurs within the donut it will not bias our estimates. Furthermore, we do not allow the polynomial to vary across the cutoff. The model therefore assumes that there is not a large discontinuity at the cutoff in student test scores. That is, the relationship between value-added and subsequent test scores should move somewhat smoothly through the award threshold. In order to provide some evidence on the validity of this assumption, we estimate a regression discontinuity model of winning the award in a given year on subsequent test score outcomes. This model is similar to equation (13), but we allow the running variable to differ across the cutoff and include an indicator for whether the teacher is above the cutoff.

Table 3 presents regression discontinuity estimates pooled across subjects using bandwidths of 0.5, 0.75 and 1.0 and a linear spline in distance to the cutoff.[20] At most, there is a small relationship between award receipt and subsequent award performance. For the low award, the estimates are negative but quite small in absolute value and are not statistically significantly different from zero at even the 10% level. For the higher award, we do see some evidence of a positive effect, but only for the larger bandwidths. Furthermore, these estimates are small, at about 2% of a standard deviation, and they only are significant at the 10% level. On the whole, these results support the way in which we have modeled the running variable in equation (13) as any jump in subsequent test score performance at the award cutoff is small.[21]

In Appendix Table A-1, we estimate the regression discontinuity model using teacher and student observable characteristics as the dependent variable. These estimates test for the validity of the RD design in this setting, and they provide indirect evidence on the likelihood of sorting into the local treatment window. Whether a teacher wins an award is far more salient than how close a teacher is to an award threshold. Hence, if there is little sorting with respect to this observable feature of the ASPIRE system, it is unlikely there will be differential sorting into the local treatment donut. Appendix Table A-1 shows little evidence that teacher or student characteristics jump at the award threshold. There is one statistically significant estimate out of 20 in the table, and on the whole the coefficients are close to zero. It therefore does no appear that students sort according to teacher award determination, nor are teachers with different characteristics more likely to win an award. Online Appendix Figure A-1 reinforces this conclusion by showing the distribution of value-added relative to the cutoff.[22] For the top

[20] We have conducted extensive tests for the sensitivity of these results to the use of higher order polynomials in the running variable and to the use of different bandwidths. Our results are not sensitive to these modeling assumptions.

[21] A question arises as to why future test scores may respond to a teacher winning an award. One possibility is that winning an award generates an increase in job satisfaction, which in turn could improve student performance. Links between job satisfaction and productivity have been shown in other sectors (Bockerman and Ilmakunnas, 2012; Jones et. al., 2009). Judge et. al., 2001 provides for a review of the early literature. Another possibility is that teachers treat the awards as a signal of their teaching ability. This could affect achievement in two ways. First, teachers may gain direct utility from the positive feedback, which in turn would increase job satisfaction and performance. Research in both economics and psychology show that workers respond to feedback about their performance, though the relationship is not necessarily positive (Ederer, 2010; Ferdor, 1991; Podsakoff and Farh, 1989; Pearce and Porter, 1986; Stone and Stone, 1985; Ilgen, Fisher and Taylor, 1979). The other possible pathway is that a positive signal makes it more likely that a teacher will remain in the profession, leading her to invest in more teacher-specific human capital. Further, it should be noted that Ahn and Vigdor (2014) find evidence in North Carolina that schools as a whole respond to the receipt of school-wide rewards.

[22] Although the awards are based on value-added percentile, teachers can appeal, which can create heaping

award, there is little evidence of excess density right over the cutoff. For the low award, there is a small increase in the density right above the award threshold. But, this appears to be a natural feature of the distribution as the lower value-added cutoff tends to be at zero, which is the mean of the distribution. Tests of discontinuities of the density show no statistically significant changes (McCrary, 2008).

6 Results

We first present the results graphically in Figure 3, as the graphical results provide a clear demonstration of our main findings. Figure 3 shows the distribution of value-added surrounding the cutoff (left y-axis) as well as the predicted and actual test score distributions in year $y + 1$ both outside and inside the local treatment area using a donut of 1.5 value-added points. In both panels, the predicted and actual test scores are virtually identical outside of the treatment area. This indicates that our model has very good within-sample prediction and explains the data with considerable accuracy. As such, it is evidence that the out-of-sample predictions we make in the donut range will also be highly accurate estimates of what achievement would have been had those teachers been far from an award margin. As discussed in Section 5, the treatment effect is calculated as the area between the actual and predicted test scores within the local treatment donut. Figure 3 shows that these curves are indistinguishable, meaning that the treatment effects are essentially zero.

Figure 3 presents results pooled across subjects, but there could be important differences by subject. In Figure 4, we show predicted and actual test scores in year $y + 1$ as a function of distance to the median cutoff separately by subject. Figure 5 contains the same information but relative to the top quartile cutoff. There is little evidence that the actual test scores are higher than the predicted for either cutoff or for any subject within the local treatment area.

Figures 3 and 4 suggest there is no effect of being close to an award margin on subsequent performance. The estimation results in Table 4 reinforce this conclusion. Here, we show estimates of equation (14) using the three different definitions of being local to a cutoff. We

right above the cutoff.

present estimates by award level as well as pooled results and results by subject. The cross-subject pooled estimates are very close to zero and are not sensitive to the size of the local treatment donut. In columns (2) and (5), we can rule out positive effects larger than 1% of a standard deviation at the 5% level. Outside of science for the top quartile award, there also is no systematic evidence of a positive effect for any subject. While there are some positive and negative estimates that are statistically significant at the 5 or 10 percent level, the estimates are not robust to altering the size of the local treatment area. For science, there is a positive effect of about 2% of a standard deviation across specifications for the high award. While this provides some evidence of teacher responsiveness to performance incentives, the estimated effect is nonetheless very small. Given that we only find an impact for one subject at one award threshold and that this estimate, while statistically significant, is economically insignificant, we conclude that on the whole there is very little evidence that teachers proximal to an award threshold exhibit larger performance gains in the subsequent year.

While the overall results are essentially zero, it is possible that they hide impacts for certain types of teachers. In Table 5, we examine whether there is any evidence of heterogeneity across types of teachers or students taught. First, we explore the role of teacher experience. Since novice teachers may have less knowledge of their ability, they may be more affected by the feedback from value-added estimates. However, we do not see any effect of being close to a performance cutoff for teachers of any experience level. Given research that suggests that males are more responsive to tournaments than females (Gneezy, Niederle and Rusticini, 2004) we investigate gender effects. While the sample sizes for males are relatively small, male and female teachers nonetheless appear equally unresponsive. This is consistent with findings of little gender difference in incentive pay in Israel (Lavy 2008). Another possibility is that teachers may respond more to information early in the program since the information is more novel. Hence, we also look at differential responses by the year of the program but find no evidence of responses to being on an incentive margin in any year of the program. Finally, since teachers with more students will have more precise measures of value-added, these teachers may be more inclined to incorporate the information into their behavior. We therefore split teachers by the average number of students they teach within each subject over the years we observe them. The results indicate that teachers with more students in a subject do not exhibit larger responses.[23] We highlight that the degree to which each teacher's value-added is informative of ability may not be the relevant parameter, however, as teachers may be responding to how noisy these measures are on average.

7 Discussion

Our results show clearly that teachers who are proximal to an award cutoff in a given year do not have higher test score growth in the subsequent year. This is somewhat surprising, since the theoretical model predicts that such teachers face stronger incentives than those father away from the threshold. There are two potential explanations for the lack of results. The first is that teachers are unresponsive to financial incentives. They may already be putting forth a maximal amount of effort, or they may not know how to increase student achievement. Further, it is possible they are not knowledgeable about the financial incentives available to them under the ASPIRE program. We do not favor this interpretation, however. In related work, we found teachers in sufficiently small groups responded to the incentives they faced under the ASPIRE program (Imberman and Lovenheim 2015). While this prior study focused on high school teachers, the basic structure of the award program was the same. Relatedly, it is unlikely that high school teachers knew about the program but elementary and middle school teachers did not. The sizable bonuses available under ASPIRE further act to make the program salient.

The second potential explanation for our results comes directly from the model: if the value-added measure contains a sufficient amount of noise, teachers should not infer any signal from it about their likely location relative to the cutoff in the next year. That is, if λ is close to zero, being proximal in one year does not tell one anything about the likelihood of being proximal in the next year. Our preferred interpretation of the results is that the EVAAS value-added measure, when applied to one year of teacher data,[24] leads to noisy value-added estimates that

[23] Figures A-2 through A-5 present graphical results for these groups.

[24] The EVAAS model includes multiple years of data, but the value-added measure estimates the contribution of teachers to student test score gains in each year rather than averaged over several years. It is in this sense

did not allow teachers to infer much about their true ability. There are two pieces of evidence to support this interpretation. Panel (b) of Figure 2 shows the distribution of the within-teacher change in value-added. There is a wide distribution of the change in value-added, with a standard deviation equal to 2.1 This standard deviation is larger than standard deviation of value-added across teachers (panel a), suggesting that there is much volatility in teacher value-added changes over time.

Table 6 presents further evidence that the value-added measure used in the incentive pay system contains substantial noise. The table shows the joint distribution of value-added quintiles in years t and $t + 1$. Thus, it shows the likelihood of being in each quintile in the next year conditional on being in a given quintile in the current year. Only 29% of the distribution is on-diagonal. Particularly in quintiles 2, 3 and 4, which are the relevant quintiles for being proximal to an award threshold, the distribution is almost uniform. It thus is not surprising that teachers were unresponsive to the award incentives, as being close to an award threshold told them little about their likelihood of being close in the subsequent year. The tabulations in Table 6 suggest that λ is close to zero, and as a result teachers do not respond to proximity to an award threshold by increasing effort. While it still is possible that *all* teachers increased their effort due to the ASPIRE program, prior work examining average effects among teachers exposed to a teacher-level value-added merit pay tournament shows that this is unlikely (Springer et al. 2010). Rather, it appears that teachers did not respond much to the individual value-added incentives because the measure used to determine awards was too noisy to provide useful information about the returns to effort. We note that this is entirely consistent with the findings in Imberman and Lovenheim (2015), as the group-based value-added used in the high schools would benefit from more observations that likely reduce the inter-temporal variance within groups. Our results therefore point to the precision of the value-added measure being used as an important design feature in individual merit pay tournaments that has received little attention in previous research.

that the EVAAS model uses only one year of data when calculating value-added.

8 Conclusion

This paper estimates the effect of the specific incentives teachers face under a teacher-based incentive pay program in the Houston Independent School District, called ASPIRE. The ASPIRE program is a rank-order value-added tournament in which teachers compete against others who teach the subject and similar grades across the district. Those with value-added over the median earn an award, and the award amount doubles for those in the top quartile. The awards are substantial in size, which makes it likely teachers knew about and responded to the program.

We first build a theoretical model that guides our empirical approach. The intuition for the model follows from the fact that teachers whose ability places them closer to an award threshold are likely to be the most responsive, as they face the highest expected marginal return to effort. However, if teachers do not know their ability, they will infer it from how close they were to the cutoff in the prior year. The degree to which they can make such an inference is proportional to the precision of the value-added measure as a measure of underlying ability. The central prediction of the model that we test empirically is that those closer to an award threshold in the prior year will exert more effort and have higher student test score growth. We take this prediction to the data using student-teacher linked administrative data on all elementary and middle school students in HISD. Our empirical approach adapts the bunching estimator pioneered by Saez (2010) and estimates whether there is excess test score growth among teachers who were near to a cutoff in the prior year. Our results show conclusively that this is not the case: teachers do not respond differentially to the ASPIRE incentives based on their proximity to the cutoff in the prior year.

We argue that the lack of responsiveness among teachers to the incentives they face under the ASPIRE program is driven, at least in part, by the imprecision of the value-added estimates. Thus, they do not appear to be an informative signal of teacher ability, which significantly mutes the incentives embedded in the merit pay system. From a policy perspective, this is an important finding as it highlights that the feedback mechanism inherent in individual incentive pay systems as well as the precision of the value-added estimate are core parameters of interest in the design of incentive pay. Because teachers will use the incentive pay measures as information about their ability to inform their effort decisions, it is critical that these measures provide accurate information about their ability. However, due to small class sizes and the common use of merit pay systems based on a single year of teacher performance, there is a technological constraint that exists in the design of teacher merit pay. Individual-based incentives may not generate achievement gains because it is not possible to calculate value-added measures that are sufficiently precise using one year of a given teacher's outcomes. A potential solution to this problem is to use group-based outcomes that can increase precision. However, as Imberman and Lovenheim (2015) and Goodman and Turner (2013) show, group-based incentive pay can lead to free riding. Taken together, these studies suggest that there is a balance that needs to be struck between the group size and the amount of measurement noise when designing incentive pay systems in order to maximize their effectiveness on student achievement.

References

[1] Ahn, Thomas, and Jacob L. Vigdor, 2014. "When Incentives Matter Too Much: Explaining Significant Responses to Irrelevant Information." tional Bureau of Economic Research Working Paper No. w20321.

[2] Ballou, Dale, 2001. "Pay for Performance in Public and Private Schools." *Economics of Education Review* 20(1): 51–61.

[3] Ballou, Dale, William Sanders and Paul Wright, 2004. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics* 29(1): 37–65.

[4] Bastani, Spencer and Hakan Selin, 2014. "Bunching and Non-bunching at Kink Points of the Swedish Tax Schedule." *Journal of Public Economics* 109(January): 36–49.

[5] Bockerman, Petri and Pekka Ilmakunnas, 2012. "The Job Satisfaction-Productivity Nexus: A Study Using Matched Survey and Register Data" *Industrial and Labor Relations Review* 65(2): 244-62.

[6] Chetty, Raj, Jonathan Friedman and Jonah Rockoff, 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633-2679.

[7] Chetty, Raj, Jonathan Friedman and Emmanuel Saez, 2013. "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings." *American Economic Review* 103(7): 2683-2721.

[8] Dee, Thomas S. and Benjamin J. Keys, 2004. "Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment." *Journal of Policy Analysis and Management* 23(3): 471–488.

[9] Dee, Thomas S. and James Wyckoff, 2015. "Incentives, Selection and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34(2): 267–297.

[10] Ederer, Florian, 2010. "Feedback and Motivation in Dynamic Tournaments." *Journal of Economics and Management Strategy* 19(3): 733–69.

[11] Fryer Jr, Roland G., 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.

[12] Gaurino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge, 2014. "Can Value-Added Measures of Teacher Performance Be Trusted?" *Education Finance and Policy* 10(1): 117-156.

[13] Gaurino, Cassandra M., Michelle Maxfield, Mark D. Reckase, Paul N. Thompson, and Jeffrey M. Wooldridge, 2015. "An Evaluation of Empirical Bayess Estimation of Value-Added Teacher Performance Measures." *Journal of Educational and Behavioral Statistics* 40(2): 190-222.

[14] 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).

[15] Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, 2003. "Performance in Competitive Environments: Gender Differences." *Quarterly Journal of Economics* 118(3): 1049-1074.

[16] Imberman, Scott A. and Michael F. Lovenheim, 2015. "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System." *Review of Economics and Statistics* 97(2): 364–386.

[17] Jones, Derek C., and Takao Kato, 1995. "The productivity effects of employee stock-ownership plans and bonuses: evidence from Japanese panel data." *American Economic Review* (85)3 : 391-414.

[18] Jones, Melanie K., Richard J. Jones, Paul L. Latreille and Peter J. Sloane, 2009. "Training, Job Satisfaction, and Workplace Performance in Britain: Evidence from WERS 2004." *Labour* 23: 139–75.

[19] Judge, Timothy A., Joyce E. Bono, Carl J. Thoresen and Gregory K. Patton, 2001. "The Job Satisfaction-Job Performance Relationship: A Qualitative and Quantitative Review." *Psychological Bulletin* 127(3): 376–407.

[20] Kane, Thomas J. and Douglas O. Staiger, 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper No. 14607.

[21] Kleven, Henrik J. and Mazhar Waseem, 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *Quarterly Journal of Economics* 128(2): 669-723.

[22] Kleven, Henrik J., Martin B. Knudsen, Claus Thustrup Kreiner, Soren Pedersen, and Emmanuel Saez, 2011. "Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark." *Econometrica* 79(3): 651-692.

[23] Ladd, Helen and Douglas L. Lauen, 2010. "Status Versus Growth: The Distributional Effects of School Accountability Policies." *Journal of Policy Analysis and Management* 29(3): 426–450.

[24] Lavy, Victor, 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." *Journal of Political Economy* 110(6): 1286–317.

[25] Lavy, Victor, 2008. "Gender Differences in Market Competitiveness in a Real Workplace: Evidence From Performance-based Pay Tournaments Among Teachers." National Bureau of Economic Research Working Paper No. w14338.

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

[27] McCrary, Justin, 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2): 698–714.

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

[29] 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.

[30] Neal, Derek and Diane Whitmore Schanzenbach, 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability" *Review of Economics and Statistics* 92(2): 263-283.

[31] Pearce, Jone L. and Lyman W. Porter, 1986. "Employee Responses to Formal Performance Appraisal Feedback" *Jounal of Applied Psychology* 71(2):211–18.

[32] Podsakoff, Philip M and Jiing-Lih Farh, 1989. "Effects of Feedback Sign and Credibility on Goal Setting and Performance." *Organizational Behavior and Human Decision Processes* 44(1): 45–67.

[33] Reback, Randall, 2008. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics* 92(5): 1394–1415.

[34] Rivkin, Steven G., Eric A. Hanushek and John F. Kain, 2005. "Teachers, Schools and Academic Achievement." *Econometrica* 73(2): 417–458.

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

[36] Saez, Emmanuel, 2010 "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2(3): 180-212.

[37] Sallee, James M. and Joel Slemrod, 2012. "Car Notches: Strategic Automaker Responses to Fuel Economy Policy." *Journal of Public Economics* 96(11-12): 981-999.

[38] 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.

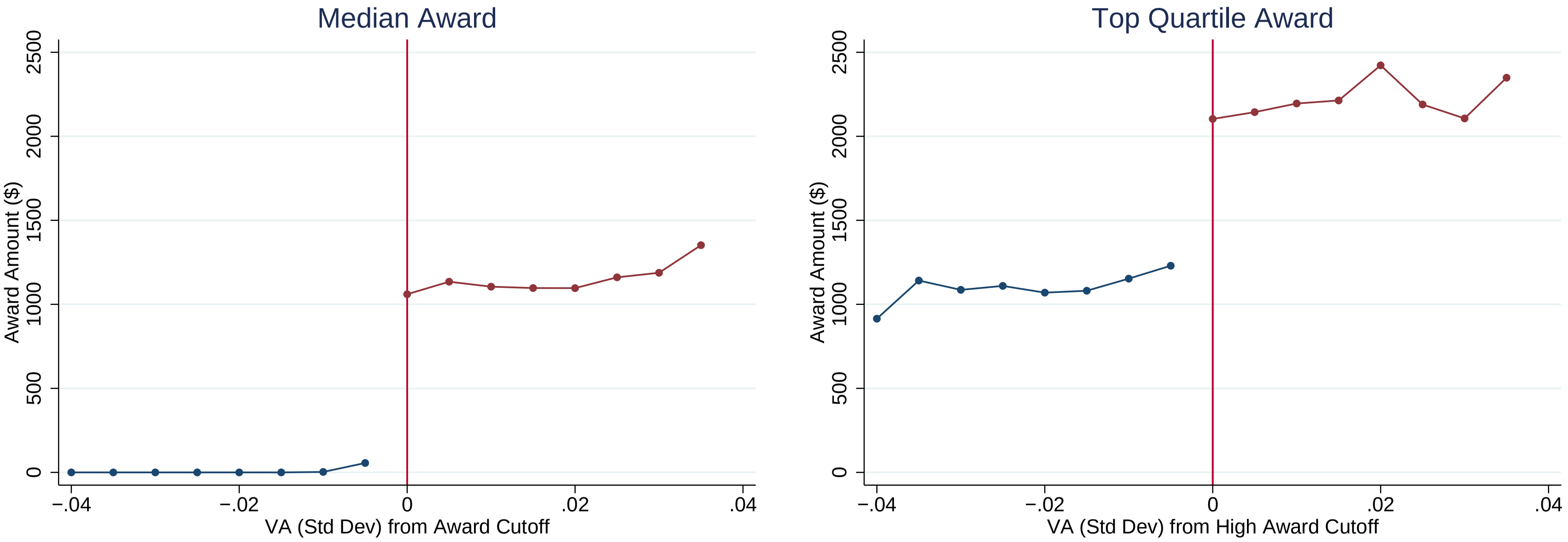
[39] Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper and Brian M. Stecher, 2010. "Teacher Pay For Performance: Experimental Evidence from the Project on Incentives in Teaching." National Center on Performance Incentives: http://www.performanceincentives.org/data/files/pages/POINT%20REPORT_9.21.1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

[40] Sojourner, Aaron, Elton Mykerezi, and Kristine West. 2014. "Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-Comp in Minnesota." *Journal of Human Resources* 49(4): 945-981.

[41] Stone, Dianna L. and Eugene F. Stone. 1985. "The Effects of Feedback Consistency and Feedback Favorability on Self-Perceived Task Competence and Perceived Feedback Accuracy." *Organizational Behavior and Human Decision Processes* 36(2): 167–85.

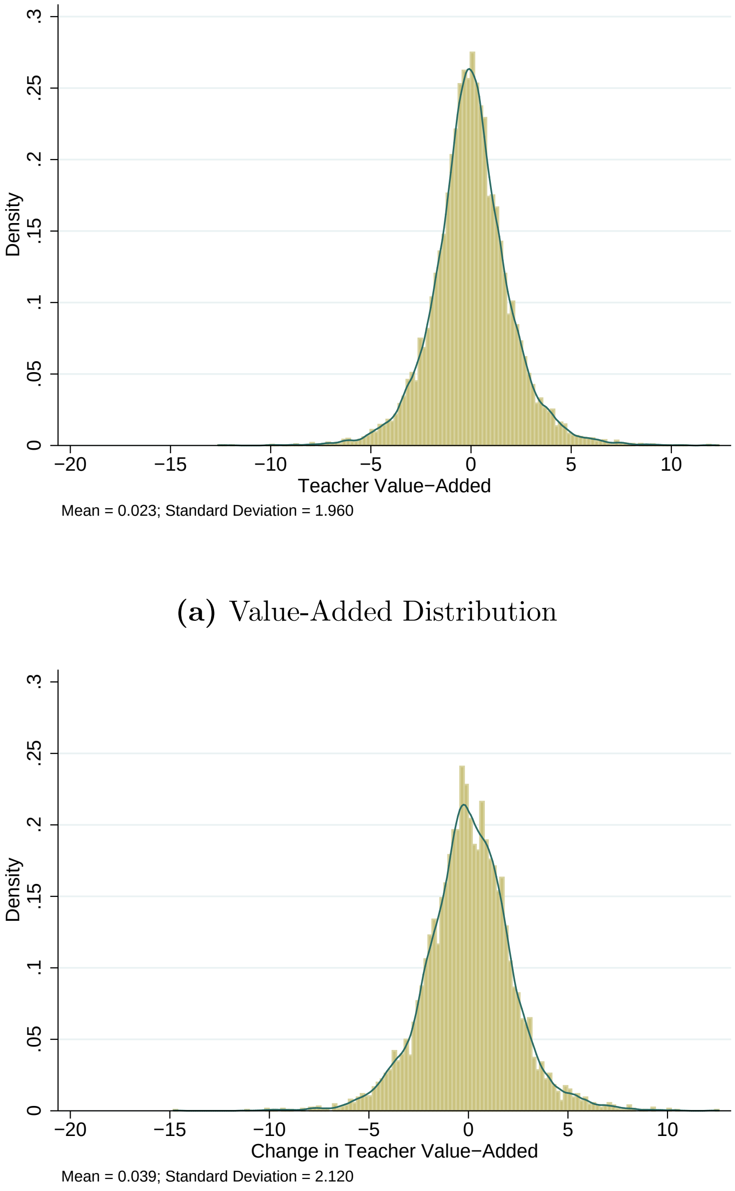
[42] Wright, S. Paul, William L. Sanders and June C. Rivers, 2006. "Measurement of Academic Growth of Individual students toward Variable and Meaningful Academic Standards." In *Longitudinal and Value Added Models of Student Performance*, R. W. Lissitz, ed.: 385–406.

Figure 1: Award Amount by Lagged Value-Added Scores



Each dot represents the mean award received for each 0.002 standard deviation bin.

Figure 2: Distributions of Teacher Value-Added



Mean = 0.023; Standard Deviation = 1.960

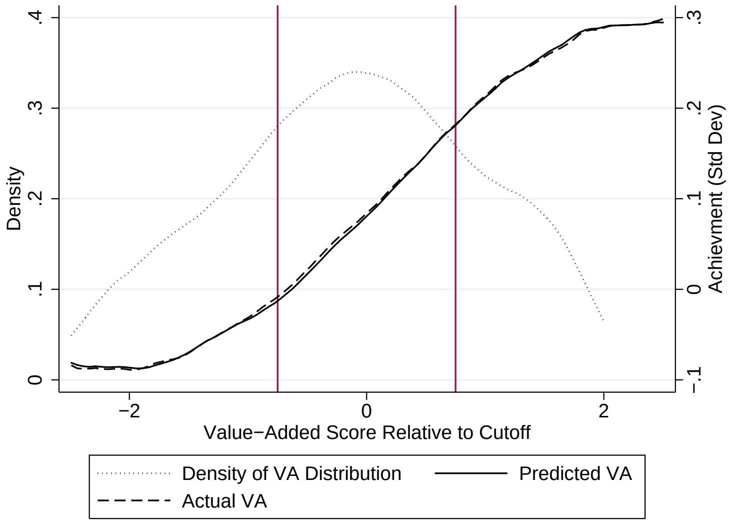
(a) Value-Added Distribution

Mean = 0.039; Standard Deviation = 2.120

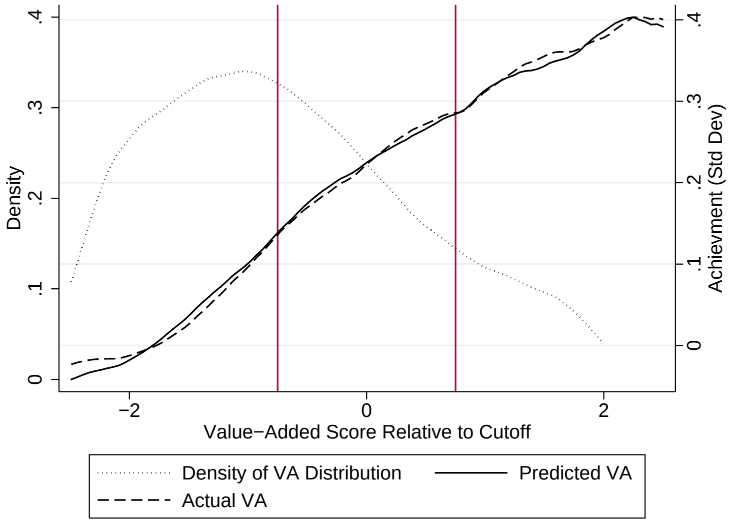
(b) Distribution of Annual Change in Value-Added

Observations are at the teacher-subject-year level. Data is pooled over all subjects. Changes are calculated as VA in year *t-1* minus VA in year *t* within teacher and subject.

Figure 3: Donut Prediction Estimates: Actual vs. Predicted Achievement



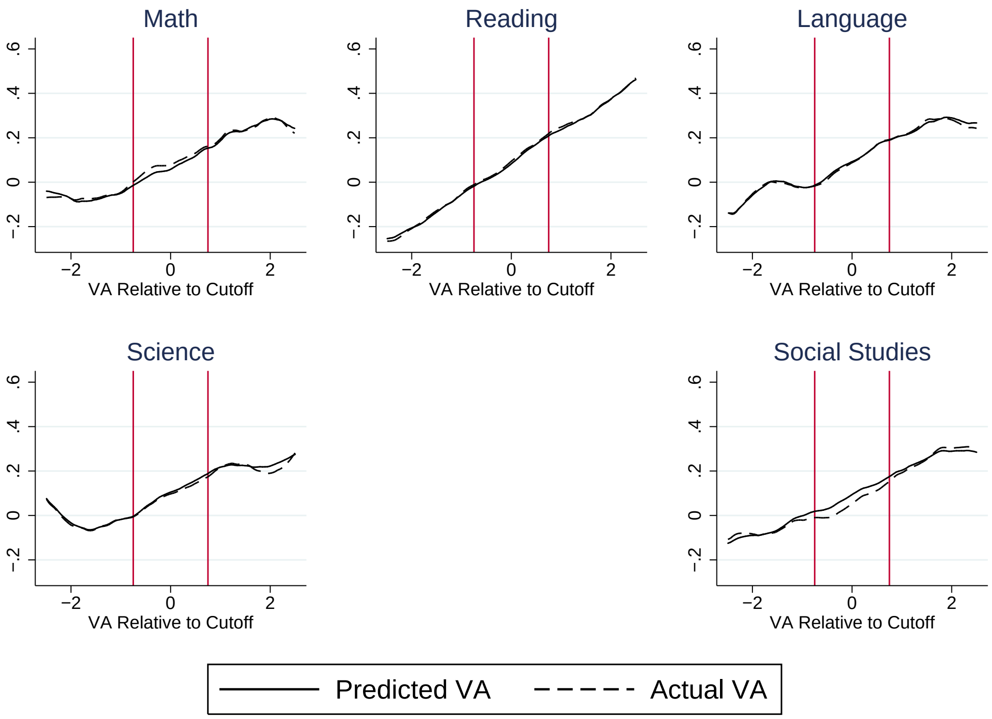
(a) Low Award Cutoff



(b) High Award Cutoff

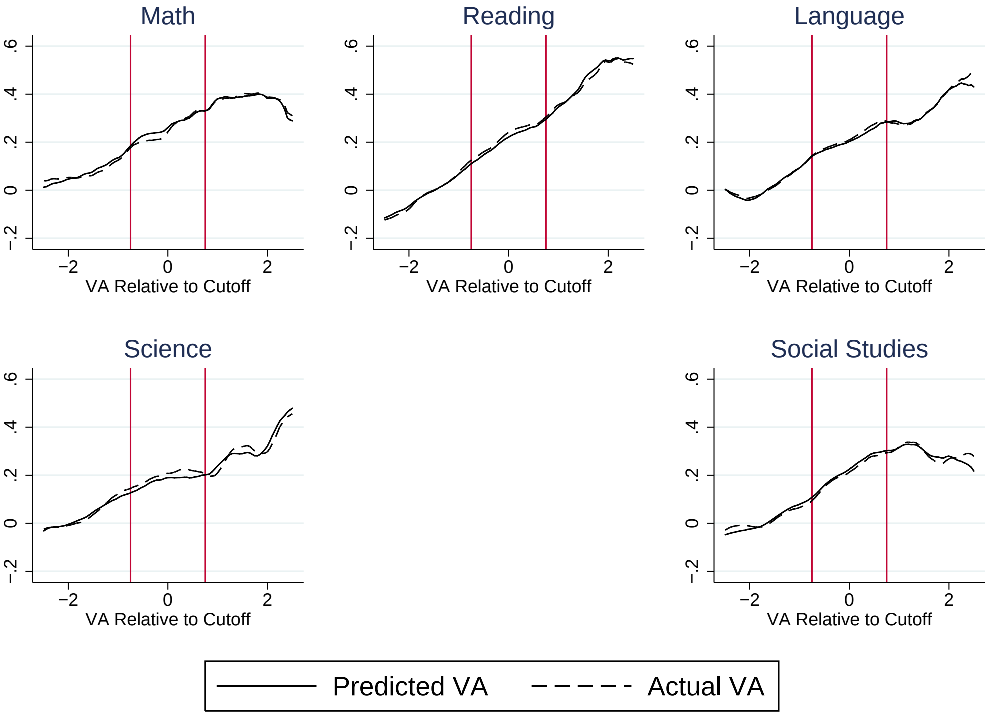
Plots show the actual kernel estimates of student achievement in the year following receipt of a value-added score and predicted values estimated using only observations outside 0.75 VA points from the cutoff. The estimation model includes a quintic in prior value-added, the student's lagged achievement, student ethnicity, gender, grade level, economic disadvantage the year of the tournament and indicators for the specific tournament. All subjects are included in the model. Test scores are standardized within grade, subject and year and the outcome is for the test score used in determining the teacher's award.

Figure 4: Donut Prediction Estimates for Low Cutoff by Subject: Actual vs. Predicted Achievement



Plots show the actual kernel estimates of student achievement in the year following receipt of a value-added score and predicted values estimated using only observations outside 0.75 VA points from the cutoff. The estimation model includes a polynomial in prior value-added, the student's lagged achievement, student ethnicity, gender, grade level, economic disadvantage the year of the tournament and indicators for the specific tournament. The polynomial order is determined by minimizing the Bayesian Information Criterion of between one and six order polynomials. The order for each subject is provided in Table 5. Test scores are standardized within grade, subject and year and the outcome is for the test score used in determining the teacher's award.

Figure 5: Donut Prediction Estimates for High Cutoff by Subject: Actual vs. Predicted Achievement



Plots show the actual kernel estimates of student achievement in the year following receipt of a value-added score and predicted values estimated using only observations outside 0.75 VA points from the cutoff. The estimation model includes a polynomial in prior value-added, the student's lagged achievement, student ethnicity, gender, grade level, economic disadvantage the year of the tournament and indicators for the specific tournament. The polynomial order is determined by minimizing the Bayesian Information Criterion of between one and six order polynomials. The order for each subject is provided in Table 5. Test scores are standardized within grade, subject and year and the outcome is for the test score used in determining the teacher's award.

Table 1: Program Details of the ASPIRE Individual Incentive Pay Program for 3^{rd} Through 8^{th} Grade Teachers

Program Description - Separate award for each subject. Determined by teacher-specific value-added. Must have value-added > 0 to receive award. Self-contained teachers in grades 3 - 6 are compared to self-contained teachers in all schools who teach the same subject and grade level. Departmentalized teachers in grades 3 - 5 are compared to all departmentalized teachers in elementary schools who teach the same subject. Departmentalized teachers in grades 6 - 8 are compared to all departmentalized teachers in middle schools who teach the same subject. Exception is middle school teachers in 2009-10 who only teach one grade level. These teachers are compared only to teachers who teach the same grade level.

Year	Per-Subject Award For Being in Top 50%	Per-Subject Award For Being in Top 25%	Max Award (with 10% Attendance Bonus)
2006-2007	$\frac{\$2500}{\#of Subjects Taught}$	$\frac{\$5000}{\#of Subjects Taught}$	\$5500
2007-2008	$\frac{\$2500}{\#of Subjects Taught}$	$\frac{\$5000}{\#of Subjects Taught}$	\$5500
2008-2009	$\frac{\$3500}{\#of Subjects Taught}$	$\frac{\$7000}{\#of Subjects Taught}$	\$7700
2009-2010	$\frac{\$3500}{\#of Subjects Taught}$	$\frac{\$7000}{\#of Subjects Taught}$	\$7700

Table 2: Descriptive Statistics of Teacher and Student Demographic Variables

	Below Median Award	Between Median and Top Award	Above Top Quartile Award
Student Characteristics			
Female	0.50	0.50	0.51
	(0.50)	(0.50)	(0.50)
Economic Disadvantage	0.83	0.79	0.75
	(0.37)	(0.41)	(0.43)
Black	0.29	0.28	0.26
	(0.45)	(0.45)	(0.44)
Hispanic	0.61	0.58	0.58
	(0.49)	(0.49)	(0.49)
White	0.07	0.10	0.12
	(0.25)	(0.29)	(0.32)
Gifted	0.15	0.18	0.22
	(0.36)	(0.39)	(0.41)
LEP	0.27	0.24	0.24
	(0.45)	(0.43)	(0.43)
Achievement (Std Devs)	-0.04	0.15	0.33
	(0.88)	(0.88)	(0.87)
Observations	292,765	146,191	153,108
Teacher Characteristics			
Female	0.78	0.82	0.81
	(0.42)	(0.38)	(0.39)
Grad Degree	0.27	0.26	0.28
	(0.44)	(0.44)	(0.45)
Experience	8.56	8.89	8.69
	(8.69)	(9.06)	(8.39)
Value-Added Score	-1.39	0.56	2.43
	(1.29)	(0.37)	(1.36)
Observations	7,856	3,787	3,777
Δ VA	0.91	-0.22	-1.26
	(2.02)	(1.84)	(2.10)
Observations	3,018	1,585	1,708

Source: HISD administrative data from 2006-2009. "LEP" denotes limited English proficiency. Standard deviations are shown in parentheses. Student characteristics are based on student-subject-year level observations. Teacher characteristics use teacher-subject-year level observations.

Table 3: Regression Discontinuity Estimates of the Effect of Winning an Award on Subsequent Value-Added - Pooled Subjects

	Median Award			Top Quartile Award		
	(1)	(2)	(3)	(4)	(5)	(6)
0.50 Bandwidth						
Win Award	-0.0452	-0.0268*	-0.0166	0.0165	0.0108	0.0137
	(0.0285)	(0.0147)	(0.0124)	(0.0312)	(0.0162)	(0.0150)
Observations		154,703			108,042	
0.75 Bandwidth						
Win Award	-0.0247	-0.0161	-0.0089	0.0349	0.0233*	0.0222*
	(0.0232)	(0.0121)	(0.0103)	(0.0264)	(0.0138)	(0.0123)
Observations		224,176			159,553	
1.00 Bandwidth						
Win Award	-0.022	-0.016	-0.0078	0.0171	0.0201	0.0201*
	(0.0203)	(0.0106)	(0.0090)	(0.0240)	(0.0125)	(0.0108)
Observations		281,514			214,097	
Student characteristics	N	N	Y	N	N	Y
Year indicators	N	Y	Y	N	Y	Y
Subject indicators	N	Y	Y	N	Y	Y

Source: HISD administrative data as described in the text. All regressions include a linear spline in prior value-added score. "Student Characteristics" include controls for race, gender, grade level, and tournament-by-subject indicators. Standard errors clustered at the teacher level are in parentheses: *** indicates significance at the 1% level, ** indicates significance at the 5% level, and * indicates significance at the 10% level.

Table 4: The Effect of Being Close to an Award Cutoff on Achievement: Difference Between Actual and Predicted within Donut

		Median Award			Top Quartile Award		
Subject	Donut Size (VA) →	1.0	1.5	2.0	1.0	1.5	2.0
		(1)	(2)	(3)	(4)	(5)	(6)
All Subjects	Mean Difference	-0.005	-0.006	-0.006	-0.001	-0.002	0.002
		(0.005)	(0.008)	(0.010)	(0.005)	(0.006)	(0.005)
	BIC Min Polynomial	3	3	3	2	2	2
	# of Obs in Donut	147,279	213,183	267,515	101,488	150,663	202,730
	Total Observations	592,064	592,064	592,064	592,064	592,064	592,064
Math	Mean Difference	0.006	0.051**	-0.001	-0.021**	-0.014	-0.005
		(0.014)	(0.022)	(0.041)	(0.010)	(0.011)	(0.011)
	BIC Min Polynomial	4	4	4	2	2	2
	# of Obs in Donut	26,234	37,548	47,659	16,440	24,232	33,712
	Total Observations	135,503	135,503	135,503	135,503	135,503	135,503
Reading	Mean Difference	-0.006	0.006	0.004	0.008	0.013	0.020***
		(0.009)	(0.010)	(0.012)	(0.009)	(0.008)	(0.007)
	BIC Min Polynomial	2	2	2	1	1	1
	# of Obs in Donut	32,991	48,259	59,030	24,727	36,890	48,557
	Total Observations	110,701	110,701	110,701	110,701	110,701	110,701
Language	Mean Difference	-0.018*	-0.010	0.007	0.012	0.009	0.007
		(0.009)	(0.013)	(0.011)	(0.009)	(0.009)	(0.009)
	BIC Min Polynomial	3	3	2	1	1	1
	# of Obs in Donut	27,129	39,459	50,238	20,437	30,175	39,422
	Total Observations	97,077	97,077	97,077	97,077	97,077	97,077
Science	Mean Difference	0.022	-0.006	-0.070	0.020**	0.021**	0.018*
		(0.018)	(0.029)	(0.054)	(0.010)	(0.010)	(0.011)
	BIC Min Polynomial	6	6	6	1	1	1
	# of Obs in Donut	35,440	51,147	63,696	20,689	31,936	44,160
	Total Observations	134,064	134,064	134,064	134,064	134,064	134,064
Social Studies	Mean Difference	-0.004	-0.042**	-0.010	-0.001	-0.011	-0.043**
		(0.017)	(0.019)	(0.030)	(0.012)	(0.014)	(0.017)
	BIC Min Polynomial	5	5	5	3	3	5
	# of Obs in Donut	25,485	36,770	46,892	20,437	27,430	36,879
	Total Observations	114,719	114,719	114,719	114,719	114,719	114,719

Source: Estimates of equation (13) using HISD administrative data as described in the text. All estimates include controls for student race, gender, grade level, economic disadvantage, year, and tournament indicators. "All-subjects" model also includes tournament-by-subject indicators. The BIC Min Polynomial is the order for the polynomial in value-added that minimizes the Bayesian Information Criterion up to a maximum of 6. The donut size is the width of the range of value-added around the award cutoff that is not used to estimate the empirical model. Predicted values for observations in that range are then calculated using this estimation and the average difference in the donut between actual and predicted achievement provides the impact estimate. Bootstrap standard errors clustered at the teacher-year level (500 reps) are in parentheses: *** indicates significance at the 1% level, ** indicates significance at the 5% level, and * indicates significance at the 10% level.

Table 5: The Effect of Being Close to an Award Cutoff on Achievement: Heterogeneity

	Sample	Median Award	Top Quartile Award	Sample	Median Award	Top Quartile Award
Mean Difference	Teacher Exp	-0.020**	0.004	Year of Tournament	-0.012	0.000
	0 - 5 Yrs	(0.009)	(0.010)	2007	(0.012)	(0.007)
BIC Min Polynomial		3	6		6	1
Total Observations		298,114	298,114		255,987	255,987
Mean Difference	Teacher Exp	0.002	-0.019	Year of Tournament	-0.001	-0.006
	6 - 12 Yrs	(0.012)	(0.011)	2008	(0.013)	(0.010)
BIC Min Polynomial		6	2		6	3
Total Observations		143,261	143,261		223,356	223,356
Mean Difference	Teacher Exp	-0.005	0.004	Year of Tournament	-0.004	0.008
	13+ Yrs	(0.010)	(0.009)	2009	(0.009)	(0.008)
BIC Min Polynomial		2	1		1	1
Total Observations		150,689	150,689		110,591	110,591
Mean Difference	Teacher Gender	0.006	0.018	Teachers' Student Count	0.000	0.007
	Male	(0.012)	(0.013)	Bottom Tercile	(0.009)	(0.011)
BIC Min Polynomial		1	1		1	1
Total Observations		139,144	139,144		122,013	122,013
Mean Difference	Teacher Gender	-0.002	-0.006	Teachers' Student Count	-0.003	0.010
	Female	(0.009)	(0.007)	Middle Tercile	(0.008)	(0.007)
BIC Min Polynomial		3	3		2	1
Total Observations		452,920	452,920		189,306	189,306
Mean Difference				Teachers' Student Count	0.001	-0.005
				Top Tercile	(0.011)	(0.008)
BIC Min Polynomial					5	2
Total Observations					280,745	280,745

Source: Estimates of equation (13) using HISD administrative data as described in the text. All estimates include controls for student race, gender, grade level, economic disadvantage, year, and tournament-by-subject indicators. The BIC Min Polynomial is the order for the polynomial in value-added that minimizes the Bayesian Information Criterion up to a maximum of 6. The donut size for all models is 1.5 value-added points and refers to the width of the range of value-added around the award cutoff that is not used to estimate the empirical model. Predicted values for observations in that range are then calculated using this estimation and the average difference in the donut between actual and predicted achievement provides the impact estimate. Bootstrap standard errors clustered at the teacher-year level (500 reps) are in parentheses: *** indicates significance at the 1% level, ** indicates significance at the 5% level, and * indicates significance at the 10% level.

Table 6: Distribution of Value-added Quintile Changes

Year t Value-Added Quintile	Year $t+1$ Value-Added Quintile 1	 2	 3	 4	 5	 Sum
1	0.073	0.042	0.028	0.020	0.017	0.180
2	0.042	0.042	0.045	0.037	0.025	0.191
3	0.031	0.045	0.045	0.045	0.031	0.198
4	0.030	0.041	0.046	0.050	0.049	0.217
5	0.019	0.025	0.037	0.050	0.083	0.214
Sum	0.196	0.195	0.201	0.203	0.206	1.000

Source: HISD administrative data as described in the text. The tabulations in each cell show the proportion of teachers in a given quintile in year t who are in the given quintile in year $t+1$.

Online Appendix: Not for Publication

Figure A-1: Densities of Lagged Value-Added Distributions

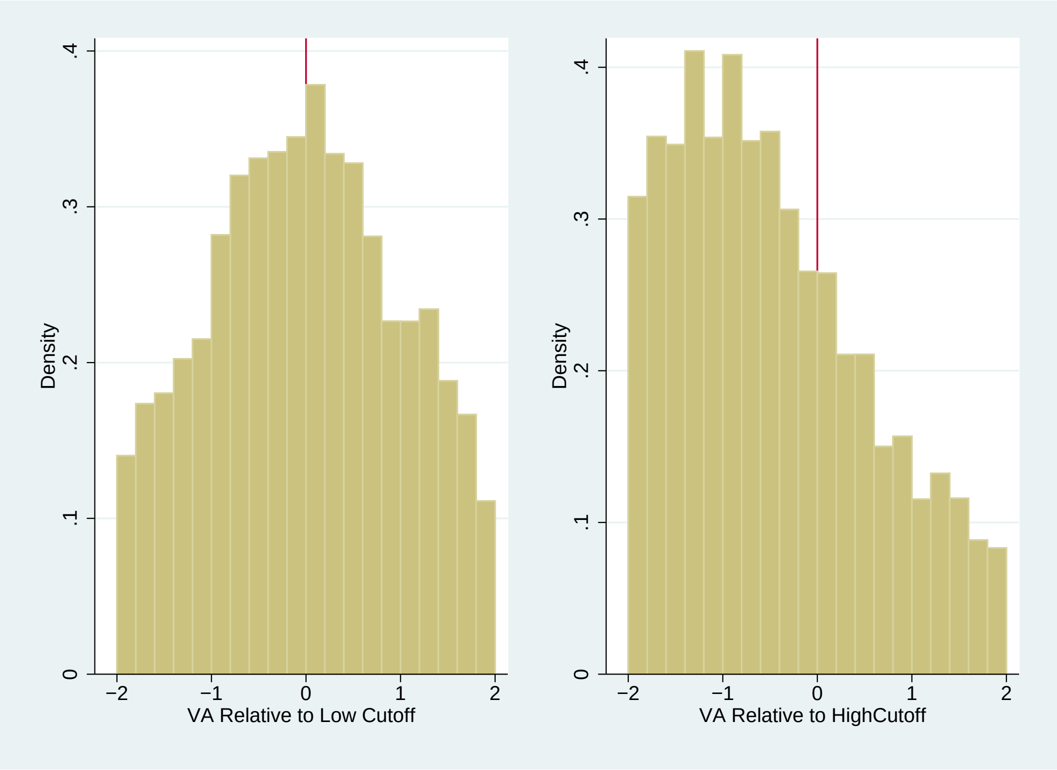
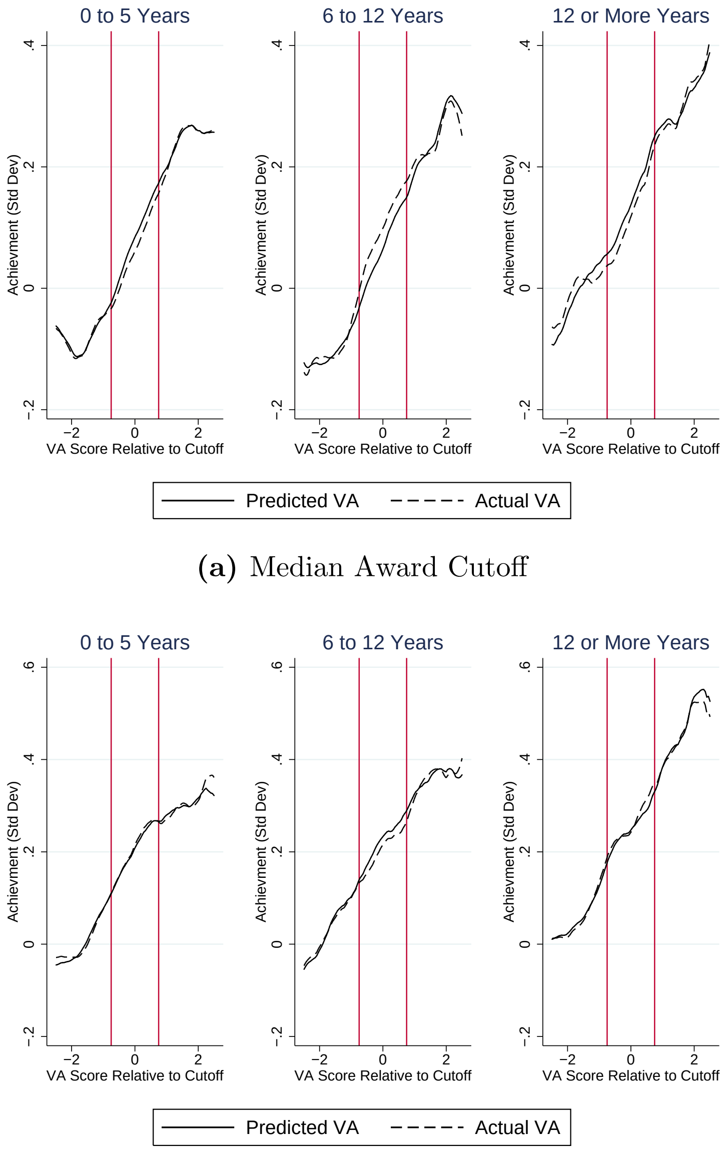


Figure A-2: Donut Prediction Estimates by Teacher Experience: Actual vs. Predicted Achievement

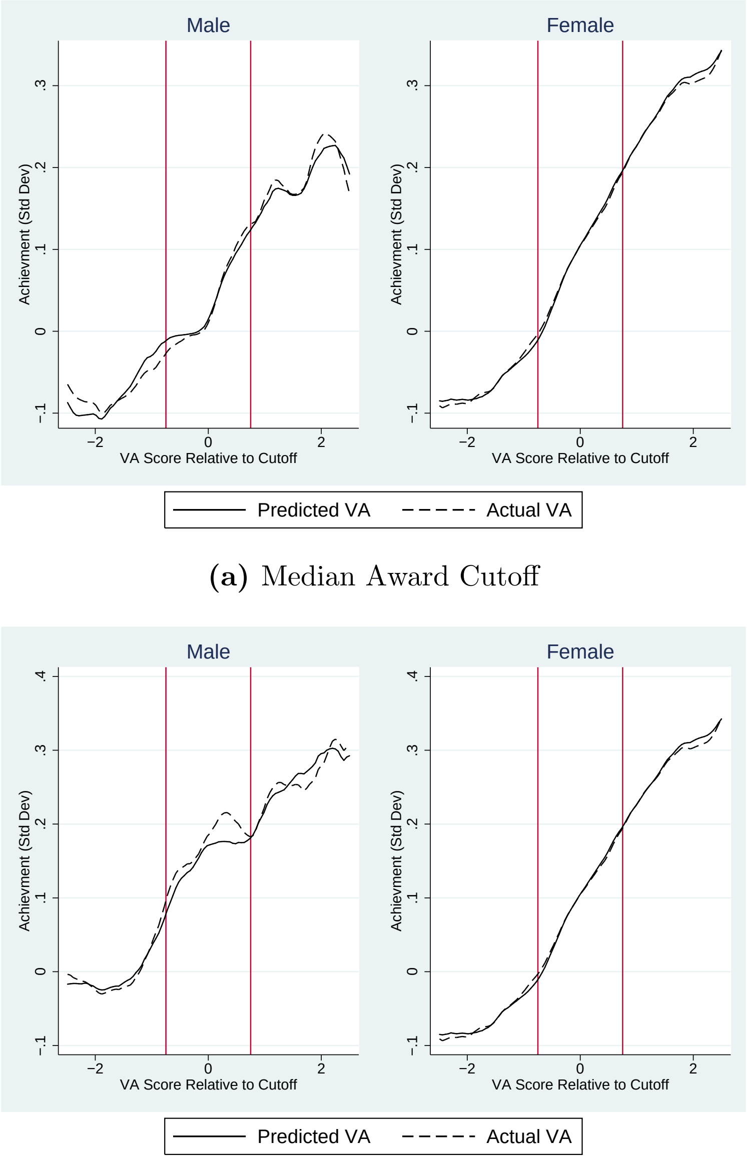


(a) Median Award Cutoff

(b) Top Quartile Award Cutoff

Plots show the actual kernel estimates of student achievement in the year following receipt of a value-added score and predicted values estimated using only observations outside 0.75 VA points from the cutoff. The estimation model includes a polynomial in prior value-added, the student's lagged achievement, student ethnicity, gender, grade level, economic disadvantage the year of the tournament and indicators for the specific tournament. The polynomial order is determined by minimizing the Bayesian Information Criterion of between one and six order polynomials. The order for each sub-group is provided in Table A-X. Test scores are standardized within grade, subject and year and the outcome is for the test score used in determining the teacher's award.

Figure A-3: Donut Prediction Estimates by Teacher Gender: Actual vs. Predicted Achievement

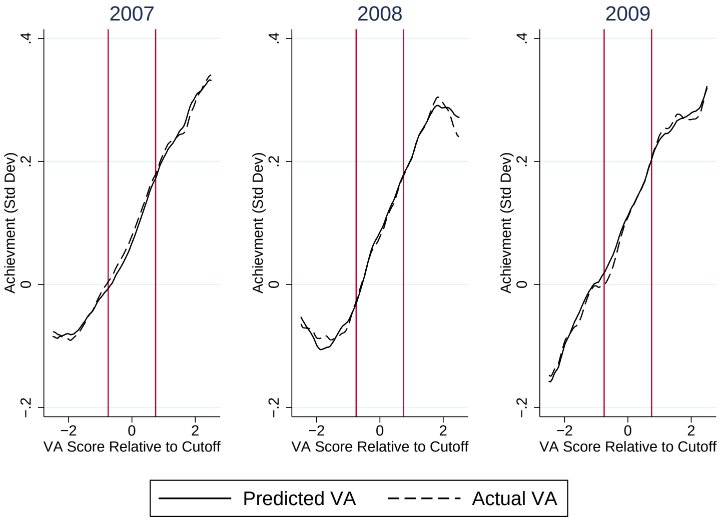


(a) Median Award Cutoff

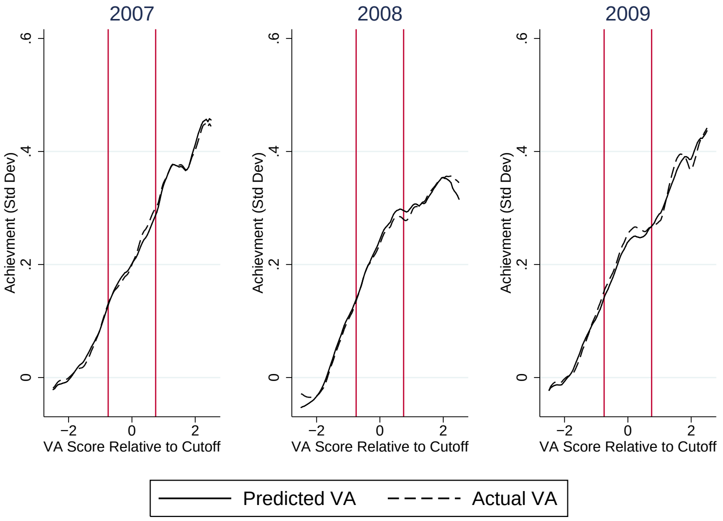
(b) Top Quartile Award Cutoff

Plots show the actual kernel estimates of student achievement in the year following receipt of a value-added score and predicted values estimated using only observations outside 0.75 VA points from the cutoff. The estimation model includes a polynomial in prior value-added, the student's lagged achievement, student ethnicity, gender, grade level, economic disadvantage the year of the tournament and indicators for the specific tournament. The polynomial order is determined by minimizing the Bayesian Information Criterion of between one and six order polynomials. The order for each sub-group is provided in Table 5. Test scores are standardized within grade, subject and year and the outcome is for the test score used in determining the teacher's award.

Figure A-4: Donut Prediction Estimates by Year of Tournament: Actual vs. Predicted Achievement



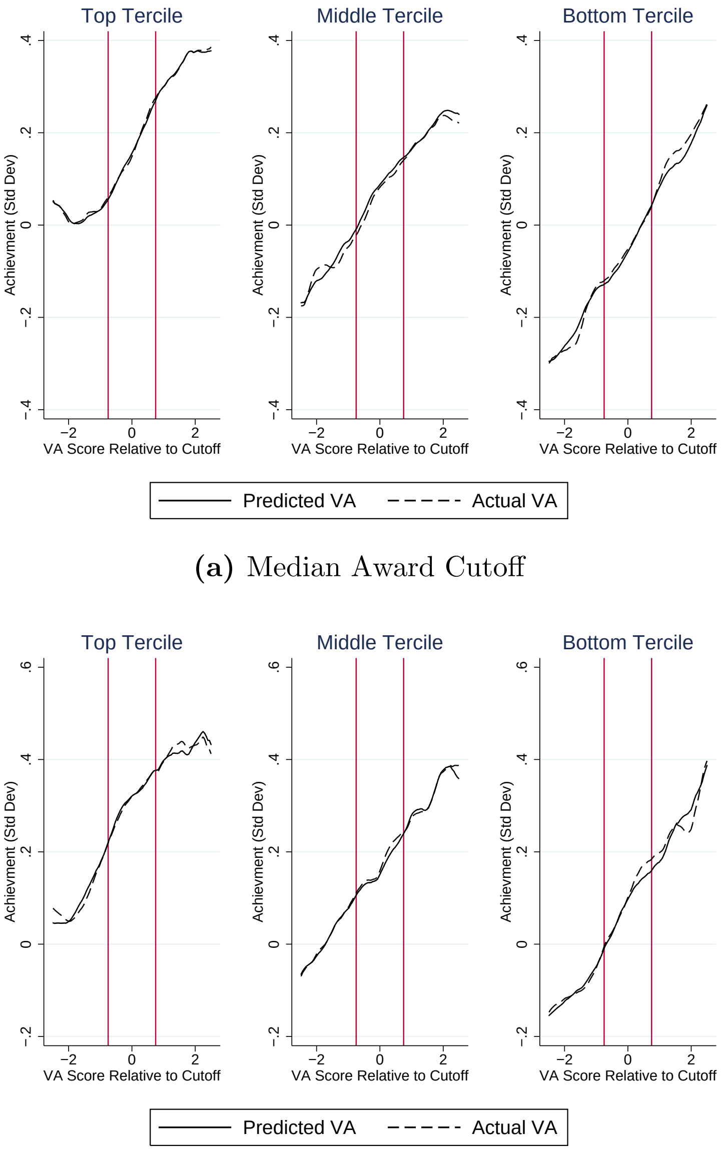
(a) Median Award Cutoff



(b) Top Quartile Award Cutoff

Plots show the actual kernel estimates of student achievement in the year following receipt of a value-added score and predicted values estimated using only observations outside 0.75 VA points from the cutoff. The estimation model includes a polynomial in prior value-added, the student's lagged achievement, student ethnicity, gender, grade level, economic disadvantage the year of the tournament and indicators for the specific tournament. The polynomial order is determined by minimizing the Bayesian Information Criterion of between one and six order polynomials. The order for each sub-group is provided in Table 5. Test scores are standardized within grade, subject and year and the outcome is for the test score used in determining the teacher's award.

Figure A-5: Donut Prediction Estimates by Number of Students Taught in Subject: Actual vs. Predicted Achievement



(a) Median Award Cutoff

(b) Top Quartile Award Cutoff

Plots show the actual kernel estimates of student achievement in the year following receipt of a value-added score and predicted values estimated using only observations outside 0.75 VA points from the cutoff. The estimation model includes a polynomial in prior value-added, the student's lagged achievement, student ethnicity, gender, grade level, economic disadvantage the year of the tournament and indicators for the specific tournament. The polynomial order is determined by minimizing the Bayesian Information Criterion of between one and six order polynomials. The order for each sub-group is provided in Table 5. Test scores are standardized within grade, subject and year and the outcome is for the test score used in determining the teacher's award.

Table A-1: Regression Discontinuity Estimates of Teacher and Student Characteristics at Award Cutoffs - 0.75 Bandwidth

Ind. Var.: Indicator For Whether Win Award		
Dependent Variable	Median Award	Top Quartile Award
Teacher Characteristics		
Female	-0.003 (0.029)	-0.084** (0.034)
Has Grad Degee	-0.017 (0.028)	0.024 (0.034)
Experience	-0.393 (0.609)	-1.188 (0.734)
Student Characteristics		
Female	0.003 (0.004)	0.005 (0.006)
Economically Disadvantaged	0.021 (0.016)	-0.004 (0.021)
Black	0.005 (0.018)	-0.016 (0.019)
Hispanic	0.014 (0.021)	0.016 (0.024)
White	-0.007 (0.010)	-0.002 (0.014)
Gifted	-0.024* (0.014)	-0.008 (0.018)
Limited English Proficiency	0.014 (0.017)	0.018 (0.019)
Lagged Achievement	-0.015 (0.030)	0.025 (0.037)
Observations	224,176	159,553

Source: HISD administrative data as described in the text. All regressions include a linear spline in prior value-added score. Standard errors clustered at the teacher level are in parentheses: ** indicates significance at the 5% level and * indicates significance at the 10% level.

Table A-2: **Regression Discontinuity Estimates of the Effect of Winning an Award on Subsequent Value-Added by Subject**

	Median Award (i)	Top Quartile Award (ii)
Math		
Win Award	0.002	0.046
	(0.027)	(0.033)
Observations	38,055	24,607
Reading		
Win Award	0.010	0.023
	(0.028)	(0.030)
Observations	48,960	37,378
Language		
Win Award	-0.040*	0.002
	(0.021)	(0.022)
Observations	49,087	38,152
Science		
Win Award	0.043	0.015
	(0.030)	(0.032)
Observations	51,283	31,971
Social Studies		
Win Award	0.020	0.038
	(0.028)	(0.033)
Observations	36,791	27,445

Bandwidth = 0.75 value-added points. Source: HISD administrative data as described in the text. All regressions include a linear spline in prior value-added score. Controls include controls for race, gender, grade level, year, tournament indicators. Standard errors clustered at the teacher level are in parentheses: *** indicates significance at the 1% level, ** indicates significance at the 5% level, and * indicates significance at the 10% level.