Journal of Public Economics 172 (2019) 84–110

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

The effect of teacher ratings on teacher performance ☆

Nolan G. Pope

The University of Maryland, 3115H Tydings Hall, 7343 Preinkert Dr., College Park, MD 20742, United States of America

ARTICLE INFO

Article history:
Received 7 March 2015
Received in revised form 24 December 2018
Accepted 1 January 2019
Available online xxxx

Keywords:
Education
Teacher
School
Ratings
Value-added
Tests
Evaluations
Los Angeles
Public

ABSTRACT

In August 2010, the Los Angeles Times publicly released value-added ratings for teachers and elementary schools in Los Angeles. Exploiting the release of these ratings as a natural experiment and using the timing of their release to account for regression to the mean, I find that low-rated teachers saw increases in their students' math and English test scores. High-rated teachers saw little to no change in their students' tests with the release of the ratings. These differential responses from low- and high-rated teachers suggest possible test score gains from the release of teacher ratings. School ratings had no additional impact on student test scores. I find no evidence that the release of the ratings affected classroom composition or teacher turnover.

1. Introduction

Firms, educators, and policymakers have long been interested in how to improve employees' productivity (Black and Lynch, 2001; Ichniowski et al., 1997). Performance evaluations have been widely used in this endeavor (Barankay, 2014). With more rigid compensation structures for public employees, publicly released evaluations and performance ratings may be a useful policy tool for influencing public employees. These publicly released evaluations or performance ratings may be able to use social and peer pressure to help improve the performance of public employees. In recent years, value-added scores have been increasingly used to measure and evaluate teacher performance. Teacher productivity, as measured by value-added scores, has been shown to have substantial, long-term effects on student outcomes (Chetty et al., 2014b). However, how teacher productivity changes in response to evaluations and feedback that use value-added scores is still unknown. Additionally, with the increased technological accessibility and demand for school and teacher accountability, performance ratings of schools and teachers are increasingly common and are often publicly available. As value-added ratings of teachers become increasingly common and public, information on how these ratings affect teachers is needed.

In August 2010, and again in May 2011, the Los Angeles Times publicly released value-added ratings of third- to fifth-grade teachers and elementary schools in the Los Angeles Unified School District (LAUSD), the nation's second largest school district. Following the release of the Times teacher ratings, intense national debate developed regarding the validity and proper use of value-added measures. In addition, during the school year 2010–2011, the LAUSD publicly released school value-added ratings and privately released teacher value-added ratings. A year following the Times teacher ratings release, the New York City Department of Education, the nation's largest school district, followed suit and publicly released teacher value-added ratings. Little of the public debate about releasing value-added ratings had empirical backing, and focused on individuals' opinions of teacher ratings. On one end of the debate, the managing editor of the Wall Street Journal said, "Public education is paid for by the public, used by the public and of crucial public concern, so this data [value-added ratings] should be made public" (Fleisher, 2012). However, others strongly opposed releasing the ratings due to the belief that such ratings have a detrimental effect on teachers. For example, Randi Weingarten, president of the American Federation of Teachers, said that such ratings amount "to a public flogging of teachers based on faulty data" (Banchero, 2012).

In this paper, I study how teachers in the LAUSD responded to learning their individual and school value-added scores, how that

☆ I would like to thank Gary Becker, Steven Davis, Erik Hurst, Sarah Komisarow, Steven Levitt, Mary Li, Derek Neal, Nathan Petek, and Stephen Raudenbush for helpful comments and discussion.
E-mail address: pope@econ.umd.edu (N. Pope).

https://doi.org/10.1016/j.jpubeco.2019.01.001
0047-2727/

response varied for low- and high-rated teachers, and the extent to which the ratings influence classroom composition and teacher turnover. I use administrative student-level panel data for over 5000 teachers and 600,000 third- through fifth-grade students from the LAUSD to estimate how a teacher's performance changes when the teacher is informed publicly that he or she is a low- or high-rated teacher. I analyze how the relationship between a standard-deviation-higher-rated teacher and students' test scores changes over time and particularly how this relationship changes with the release of the Times ratings. I use estimates from the years preceding the release of the Times ratings to provide falsification tests.

An important concern with this methodology is regression to the mean due to positive or negative shocks to teachers' student test scores during the years used to create the ratings (2002–2003 to 2008–2009). The school year 2008–2009 was the last year of data used in the Times ratings, and the Times ratings were not released until the beginning of the school year 2010–2011. Therefore, if regression to the mean is quantitatively large, I should detect evidence of it during the school year 2009–2010. However, due to both the large number of years and students in the data and the Bayesian shrinkage methods used to create the Times ratings, I find little to no evidence of regression to the mean.

I find that when teachers are informed of their Times ratings, the performance of low-rated teachers increases, whereas I find little change for high-rated teachers. After the release of the Times ratings, the benefit of having a standard-deviation-higher-rated teacher (or the cost of having a standard-deviation-lower-rated teacher) on math test scores decreases by 20 %. Similarly, the benefit of having a standard-deviation-higher-rated teacher (or the cost of having a standard-deviation-lower-rated teacher) on English test scores decreases by 23 %. These effects persist for at least two years after the ratings release. These changes close the gap between the test scores of students in high-rated and low-rated teachers' classrooms, and compresses the performance distribution of teachers. This increase in the performance of low-rated teachers is likely due to increased teacher effort and some reallocation of teacher resources from teachers' high-rated subject to their low-rated subject. An analysis of persistence indicates that "teaching to the test" does not drive the results. Besides the effect on teacher performance, I find no evidence that the Times ratings affected teacher turnover or classroom composition. The release of the school ratings had no additional impact on student test scores.

Contrary to the widely held belief that changing teacher performance is very difficult, these results show that a low-intensity intervention can change the performance of low-performing teachers. Of particular policy relevance, informing teachers of their low ratings improves the performance of low-performing teachers. In addition, there appears to be little negative impact of informing high-rated teachers of their ratings. Although the empirical strategy is not conducive to determining the overall impact of the ratings, these results suggest the release of public ratings may help increase teacher performance in a school district. These results may generalize to other industries with strong unions or rigid compensation schemes, but whether these results apply to settings with substantial rewards for excellent performance is unclear. How the results would change if the ratings were not publicly available and only dispersed privately if also unclear. The results also show such evaluations elicit a response when evaluations are at the teacher level. However, the school evaluations had little additional impact when controlling for teacher evaluations.

This paper adds to a long and growing literature of value-added methods. Since Hanushek (1971), valued-added methods have been increasingly used to measure teacher performance. The increased use of teacher value-added methods has largely been due to a lack of other predictors of teacher productivity (Hanushek and Rivkin, 2010). The value-added literature has largely focused on the credibility of value-added methods (Chetty et al., 2014a; Rockoff, 2004) and their use in selective dismissal (Goldhaber and Hansen, 2010; Gordon et al., 2006; Hanushek, 2011) and incentive pay (Fryer, 2013; Goodman and Turner, 2013; Imberman and Lovenheim, 2015; Springer et al., 2010). For example, a few school districts, such as Houston and Denver, have used value-added scores along with incentive pay (Neal, 2011). In addition, Chetty et al. (2014b) find students with higher value-added teachers have higher lifetime incomes, fewer teen births, and a higher likelihood of attending college.

This paper also relates to several recent papers that focus on how information and evaluations influence teachers performance. Rockoff et al. (2012) perform an experimental study in which principals in treatment schools were informed of their teachers' value-added scores. The authors found that treatment schools had higher turnover rates for low-performing teachers and saw small gains in math test scores. Taylor and Tyler (2012) do not use valued-added scores, but find that a yearlong subjective teacher evaluation in which teachers are evaluated by a peer teacher increases the math test scores of evaluated teachers' students for the next five years. Dee and Wyckoff (2015) also find that in the District of Columbia Public Schools, dismissal threats and financial incentives improve teacher performance.[1]

Although valued-added methods are being used widely, little research has looked at how informing teachers (or the entire public) of their value-added scores affects their productivity or student sorting. Although they do no look at productivity, concurrent work by Bergman and Hill (forthcoming) use a regression discontinuity design to look at the effect of the Times ratings on student sorting and teacher attrition. They find positive sorting of higher-achieving student into high-rated teachers after the release of the ratings. This difference in sorting is likely be due to their regression discontinuity estimates being for a specific type of teacher on a particular margin. Specifically, their results rely heavily on a few teachers that are near the publication cutoff. However, this subsample of teachers make up only a small portion of the teachers in my sample (less than 2 %) and therefore are unlike to have a large impact on my estimation. Also, since they estimate the net change between published and unpublished teachers within the same year, their results are from a different estimand. Lastly, my aggregate results may have multiple mechanisms influencing classroom composition that may be offsetting each other. Although I control for observable positive sorting of high-achieving students to high-rated teachers with prior test scores and parents' education, any unobservable positive sorting that is not controlled for by lagged test scores would push in the opposite direction of my results and bias my main effects toward zero. Their results also find that teachers whose ratings are published are significantly less likely to be retained after one year, but this effect dissipates after two years.

The rest of the paper will proceed as follows. Section 2 describes the data used to perform the analysis, and the release of the Times and LAUSD value-added ratings. Section 3 shows the results from a simple analysis using the raw data. Section 4 describes the methodology used, and discusses possible threats to identification, including why regression to the mean does not drive the results. Section 5

[1] More broadly, researchers have theorized how employees respond to different types of performance evaluations and feedback (DeNisi and Kluger, 1996). Engellandt and Riphahn (2011) show employees respond to evaluations with incentive mechanisms by increasing their on-the-job effort. However, much of the research focuses on performance evaluations and feedback not tied to incentives (Deci et al., 1999). Dixit (2002) suggests that when teachers receive information on ways they can improve, they act on this information. Anderson and Rodin (1989) hypothesize that performance evaluations increase the performance of low-performing employees but decreases the performance of high-performing employees. Alternatively, Eisenberger et al. (1990) and Pearce and Porter (1986) hypothesize that evaluations reinforce employees' views of their own productivity and make high-performing employees better and low-performing employees worse.

shows the results for how teachers' performance, classroom composition, and turnover changes with the release of the teacher and school value-added ratings. Section 6 discusses the possible mechanisms for how teachers respond. Section 7 concludes.

2. Data and institutional background

The main data used to analyze how teacher ratings affect teachers' performance are administrative student-level panel data from the Los Angeles Unified School District (LAUSD). The school district consists of over 600,000 students with roughly 70% of the student population being Hispanic. On August 29, 2010, shortly before the beginning of the school year 2010–2011, the Los Angeles Times publicly released teacher ratings for approximately 6,000 third- through fifth-grade teachers in the LAUSD. Buddin (2010) used LAUSD student-level data from school years 2002–2003 to 2008–2009 to estimate these ratings using standard value-added methods. Third- through fifth-grade teachers who taught fewer than 60 students during this time period were omitted from these August 2010 ratings. Fig. 1 outlines the data used and the timing of the ratings release.

There are several features from the structure of the Times ratings that will tend to mute the effect of regression to the mean. Since seven years of student data was used to create the teacher ratings, on average 118 student were used for each teacher's rating. This relatively large sample of students per teacher helps mitigate regression to the mean. In addition, regression to mean will depend on how heavily the data are weighted toward the last year of the sample used to create the ratings (2008–2009). Due to the many years used to create the ratings, only 15.2% of the classrooms used to create the Times ratings where from the last year of the sample (2008–2009). Lastly, Buddin (2010) used Bayesian methods to shrink the teacher value-added scores and correct for measurement error. All of these structural pieces of the Times ratings will tend to mute the impact of regression to the mean.

The Times ratings consist of a math, English, and overall value-added rating. The overall rating is a student-weighted average of the math and English ratings. All three of these ratings were publicly released on a Los Angeles Times webpage solely devoted to these ratings.[2] The ratings placed each teacher into one of five categorical rating labels for math, English, and overall teaching effectiveness. Each categorical rating contained a quintile of the teachers. For all three of these ratings, the five labels were as follows: least effective (bottom quintile of rated teachers), less effective, average effectiveness, more effective, and most effective (top quintile of rated teachers). The ratings were defined over third- through fifth-grade teachers in the school district. On the Los Angeles Times website, one could find a teacher's rating simply by searching a teacher's name or by searching a teacher's school and choosing the name from the list of that school's teachers. For each teacher, the ratings were displayed publicly online as shown by the example in Fig. 2.

To understand how and why teachers responded to the Times ratings, knowing the extent to which parents, students, administrators, and particularly teachers were exposed to the Times ratings is important. Imberman and Lovenheim (2016) find significant evidence that members of the community, and especially teachers, were well informed of the existence of the Times ratings. With a daily circulation of over 600,000,[3] the Los Angeles Times is the largest newspaper in California and the fourth largest in the United States. Over the first 10 months following the release of the Times ratings, the Los Angeles Times published37 articles and editorials about the value-added ratings, including coverage of a teacher who committed suicide shortly after the ratings release. The release also sparked national media attention in outlets such as the New York Times, National Public Radio, the Washington Post, ABC News, CBS News, CNN, and Fox News. Likely due to the widespread media response to the ratings, on the first day of the release, the Los Angeles Times website solely devoted to the ratings received a quarter million views (Song, 2010). Teachers were particularly exposed to the existence of the ratings. Arne Duncan, the US Secretary of Education, spoke out in support of the ratings. The American Federation of Teachers and the LAUSD teachers' union were vocal in their strong opposition to the ratings, and many teacher protests were organized throughout the school district. As part of the release of the ratings, the Los Angeles Times also e-mailed each teacher his or her individual rating and gave each teacher an opportunity to post a personal response to the rating on the Times website. This assured that most teachers saw their individual ratings and knew the Times ratings existed.

On May 8, 2011, the Los Angeles Times updated its teacher ratings using student-level data from the school years 2004–2005 to 2009–2010 (Buddin, 2011). These updated ratings contained third- through fifth-grade teachers regardless of the number of students taught during the time period. For the teachers previously rated in the August 2010 ratings, the correlation between the 2010 and 2011 Times ratings was 0.912. In addition, no teacher moved up or down more than one categorical label, and for each categorical label, more than 80% stayed in the same quintile label between the two ratings.

In addition to the Los Angeles Times ratings, the LAUSD contracted with the Value-Added Research Center to develop value-added scores and reports for many LAUSD teachers, including third- through fifth-grade teachers. The LAUSD denoted these value-added scores as Academic Growth over Time (AGT). For the first time in its history, the LAUSD gave these AGT reports privately to teachers and principals during the late winter and early spring of 2011. Teachers had online access to their own AGT reports starting on April 13, 2011.[4] The Los Angeles Times went to court to have these teacher AGT ratings released publicly as well, but a judge ruled in favor of the LAUSD and these ratings were not released publicly. As such, these teacher AGT ratings are unavailable for use in the analysis in this paper.

Fig. 3 gives an example of the main ratings from the AGT reports the teachers received. Similar to the Los Angeles Times ratings, the AGT reports have both a math and English rating, but not an overall rating. The AGT scores are normalized and centered at 3 with a range from 1 to 5. The AGT reports also have five categorical labels — far below predicted, below predicted, within the range predicted, above predicted, and far above predicted — that correspond with point estimates that have 95% confidence intervals that are entirely below 2, entirely below 3, contain 3, entirely above 3, and entirely above 4, respectively. In addition to the main ratings, AGT scores are reported for different student subgroups that have at least 10 students, such as race, gender, free-lunch status, or students' prior achievement. It is important to note that these AGT scores (and the Times ratings) played no role in determining teacher evaluations or teacher pay.

In addition to teachers, the Los Angeles Times rated 470 elementary schools. These school ratings were released on the same website as the teachers, and were reported in an analogous manner to the teacher ratings. These school ratings used third- through fifth-grade student test scores in the school. On April 13, 2011, through its website, the LAUSD also publicly released AGT scores for all schools in the school district.

The test used to create both the Los Angeles Times and AGT ratings is the California Standards Test (CST). The CST is a high-stakes statewide multiple-choice test given at the end of the school year to all

[2] http://projects.latimes.com/value-added.
[3] Audit Bureau of Circulations.
[4] http://portal.battelleforkids.org/BFK/LAUSD/Home.html?sflang=en.

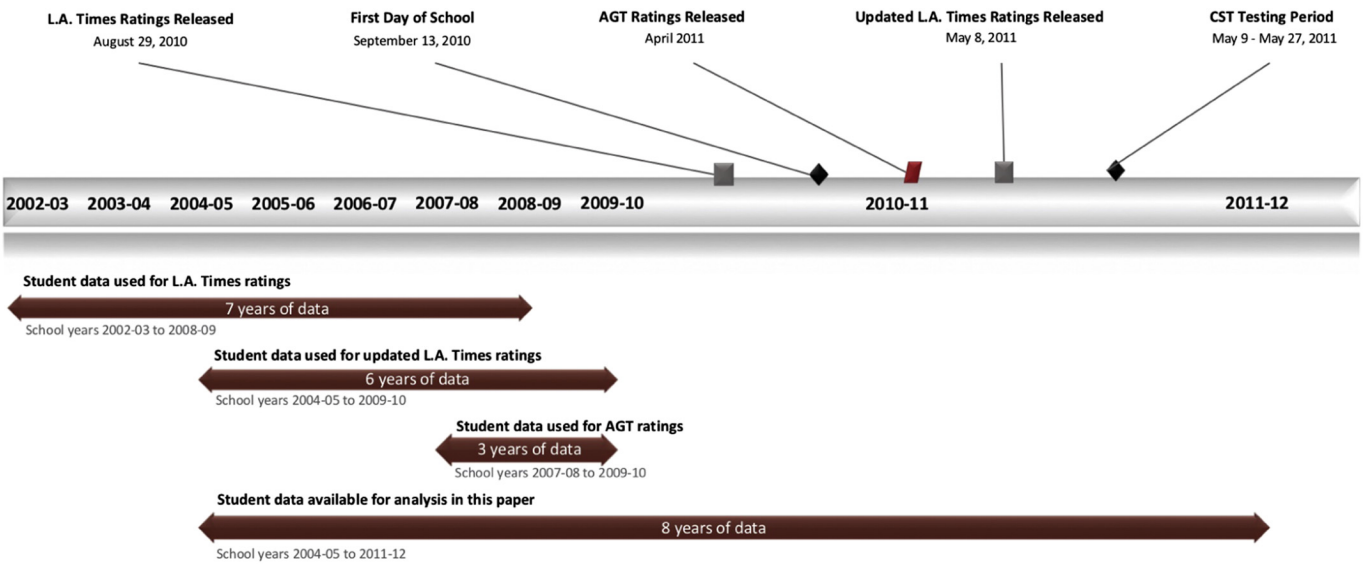

Fig. 1. Timeline of teacher ratings.

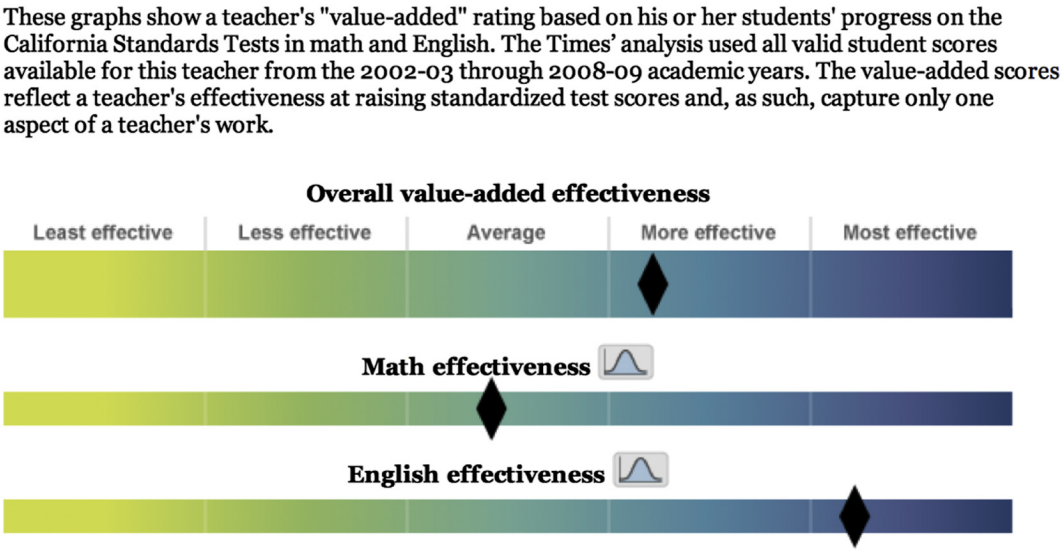

Fig. 2. Example of the Los Angeles Times online teacher ratings.

California students in grades 2–11. In the school year 2010–2011, the test window for elementary schools to administer the CST was May 9–27 and it takes place at a similar time each year. The test contains a math and English portion with 65 and 75 questions, respectively, which are each broken into two 90-minute sections. These CST scores will be heavily used in my analysis. To test what part of the distribution of teachers the Times ratings are affecting, I normalize CST scores to the 2005 CST distribution by grade. This normalization is done, because simply normalizing test scores by grade and year could lead to an artificial compression of performance for some teachers. For example, if the release of the Times ratings increased the performance of low-rated teachers and had no effect on high-rated teachers (i.e. a upward mean shift driven by the bottom of the distribution), then standardizing by year would make it appear as if the low-rated teachers were increasing their performance and the high-rated teachers were decreasing their performance. However, in actuality only the low-rated teachers are improving their performance. Normalizing the test scores to the 2005 distribution by grade avoids this problem.[5]

The administrative student-level data used in this paper include students in the LAUSD from 2004–2005 to 2011–2012. These data contain information on students' math and English CST scores, parents' education level, English Language Learner (ELL) status, grade, teacher, and school. The student-level data are also used to compute classroom size. In addition to these student data, these data contain information on teachers' age, race, gender, education level, and experience. Table 1 shows both the student and teacher summary statistics for third- through fifth-grade students and teachers

[5] I am greatly appreciative of the journal reviewers for this good suggestion and many others.

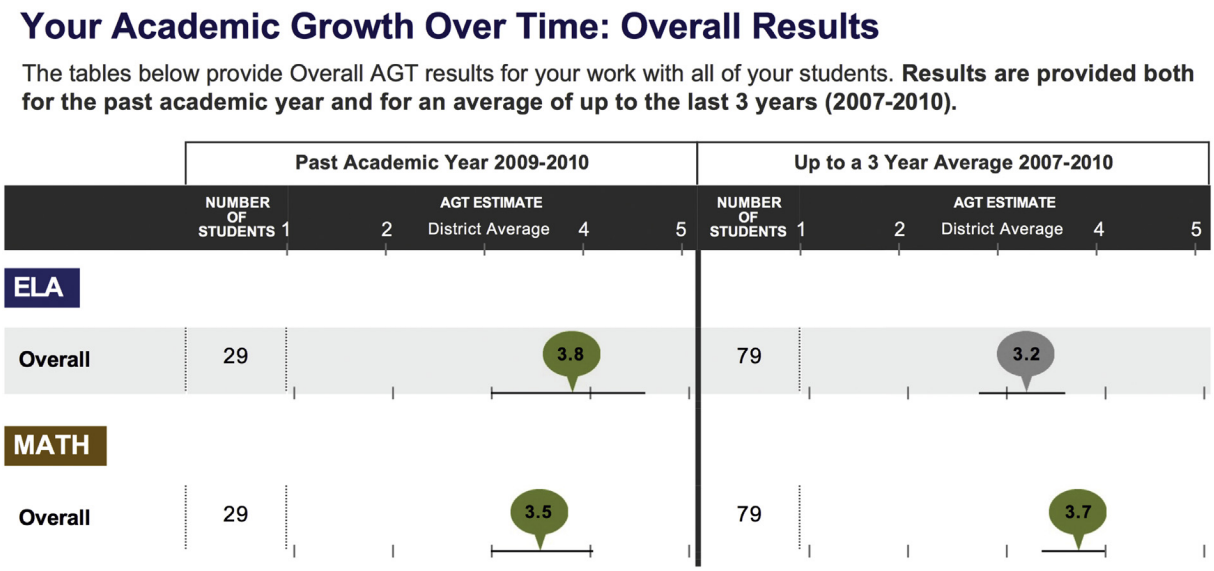

Fig. 3. Academic growth over time report example.

in 2009–2010. The data for the analysis are restricted to only teachers in the LAUSD who received a Times rating. Because teachers must have taught during the school years 2002–2003 to 2008–2009 in order to be included in the Times ratings and in the sample, the teachers in the sample tend to be slightly older and more experienced than all third- through fifth-grade teachers. However, teachers in the analysis sample are only 0.53 years older, have 0.29 more year of experience, are 2.0 percentage points more likely to be male, and 1.6 percentage point more likely to be white than all grades 3–5 LAUSD teachers. The data contain roughly 130,000 students and 5,500 teachers in each year.

3. Descriptive results

In this analysis, I look at how teachers' performance changes when teachers are informed of their value-added ratings. These data allow teachers' student test scores to be tracked over time. Along with the Los Angeles Times ratings, these data allow me to look at how the classroom averages for teachers vary by value-added ratings over time. This analysis compares the change in classroom average test scores for teachers with varying value-added ratings in the year the value-added ratings were released with the years prior to the ratings release.

Table 1
Summary statistics.

Variable	Mean	Standard deviation
Panel A: Students		
Math CST score	380.2	87.1
ELA CST score	347.0	58.1
Parent's educ (years)	12.7	2.7
Female	50.1%	–
ELL	33.0%	–
Number of students	132,137	
Panel B: Teachers		
Class size	24.6	4.6
Age	44.6	10.1
Female	73.8%	–
Master's degree	32.4%	–
Experience	12.9	7.4
Years in district	12.5	6.8
White	39.0%	–
Hispanic	38.3%	–
Black	10.6%	–
Asian	8.9%	–
Other	3.0%	–
Full time	99.3%	–
Number of teachers	5,503	

Note: The summary statistics shown in this table use data from the school year 2009–2010.

Fig. 4 shows a simple version of the analysis using the raw data. In panel (a), the x-axis represents the normalized Times math rating (the normalized Times ratings are the Times value-added scores normalized to have a mean of zero and a standard deviation of one), and the y-axis represents the change in classroom average test scores. Teachers are placed into 1 of 30 equal-sized bins according to their normalized Times math rating. Each point represents the average test scores of students in a teacher's 2010–2011 classroom minus the average test scores of students in the same teacher's 2009–2010 classroom for each bin. The solid line is the linear regression through these points. As this summary figure shows, teachers with low math ratings saw a large increase in their classroom average test scores between 2009–2010 and 2010–2011, whereas teachers with high math ratings saw little to no increase. The change in classroom average test scores decreases monotonically as teachers' math ratings increase. The four dashed lines are the linear regressions for each of the school years 2006–2007 through 2009–2010 and are created analogously to the solid regression line. The figure shows that for all of the falsification years, teachers' ratings are slightly positively correlated with changes in teachers' classroom averages. In the year the teacher ratings were released, they are negatively correlated. A similar pattern exists in panel (b) for English.

Table 2 shows another simple version of the analysis using the Times ratings. The first five columns of Table 2 report the change in classroom averages from 2009–2010 to 2010–2011 for teachers in each of the five value-added quintiles. This is the change in classroom averages from before the ratings were released to after the ratings were released. The last two columns report the difference between the least and most effective teacher quintiles along with its p-value. The first two rows show changes in the classroom average math and English test scores. For math and English, average test scores for the least effective teachers increase by 0.200 and 0.138 standard deviations, respectively. The change in test scores monotonically decreases until math and English test scores increase by only 0.057 and 0.029 standard deviations, respectively, for the most effective

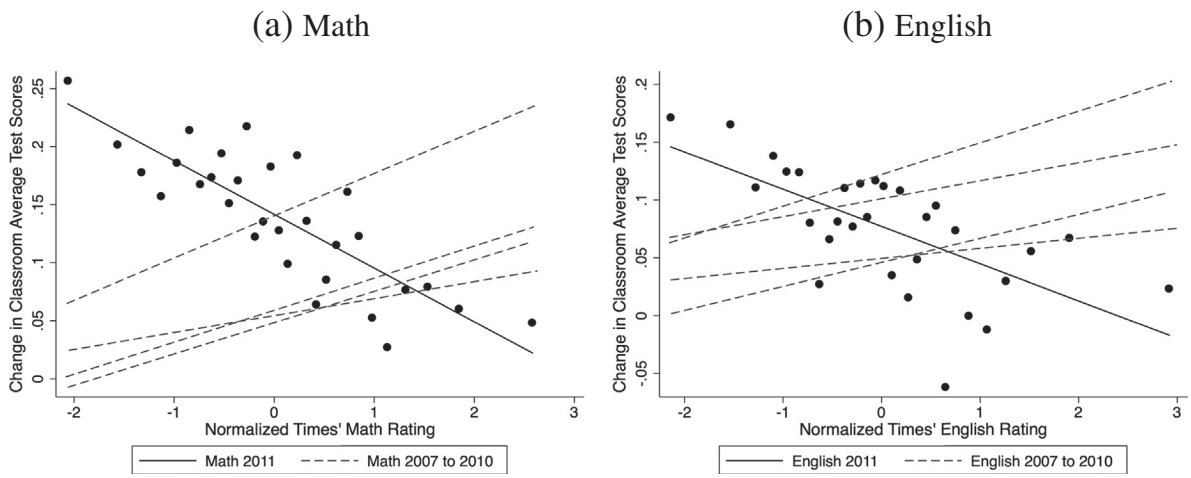

Fig. 4. Change in classroom average test scores by teacher rating. Note: Each point represents the average test scores of students in a teacher's 2010–2011 classroom minus the average test scores of students in the same teacher's 2009–2010 classroom for each bin. The solid line is the linear regression through these points. The four dashed lines are the linear regressions for each of the school years 2006–2007 through 2009–2010 and are created analogously to the solid regression line. The points used to create the four dashed lines are not shown in this figure. Each year contains approximately 130,000 students.

teachers. There is a statistically significant difference between the least and most effective teacher quintiles for math and English of 0.143 and 0.109 standard deviations, respectively, over the year in which the Times ratings and the AGT reports were released. Rows 3 through 7 show little change in other classroom characteristics, such as students' prior math and English achievement, parents' years of education, ELL status, or classroom size. None of these differences between the least and most effective teachers are statistically significant. The small changes in the other classroom characteristics are similar to changes in the previous years.

4. Empirical method

More formally, the model used to look at how the Times ratings affected teachers' performance is as follows:

$$Y_{i,t} = \alpha + \sum_{k=0}^{6} \beta_k Year_{2006+k} * V_j + \sum_{k=1}^{6} \mu_k Year_{2006+k} + \delta P_{i,t-1} + \gamma X_{i,t} + \lambda C_{i,t} + \theta_s + \varepsilon_{i,t} \quad (1)$$

and will be estimated using data on students from all years and from all teachers' classrooms. The main outcome of interest, $Y_{i,t}$, is student i's math or English CST score in year t and is normalized to the 2005 CST score distribution by grade. The variable $Year_{2006+k}$ is an indicator variable equal to one if year t is equal to 2006 + k. The variable V_j is student i's year t teacher's August 2010 Times value-added rating. I will use the August 2010 Times ratings for all of the analysis throughout the paper. The parameters of interest, β_k, are the coefficients on all the interactions between the year binary variables and the Times-rating variable. The coefficient β_k shows how having a standard-deviation-higher-rated teacher is associated with students' test scores in the indicated year. The 2006 year binary variable is the omitted variable. The vector $P_{i,t-1}$ contains student i's previous year's math and English test scores. The vector $X_{i,t}$ contains all available demographic characteristics of student i, including parents' education level, ELL status, and grade fixed effects. The vector $C_{i,t}$ contains the average prior math and English test scores of the other students in student i's classroom in order to control for peer effects. Lastly, θ_s are school fixed effects.

By including school fixed effects, the main results will show changes in test scores within schools. However, because between-school changes in test scores or within-teacher changes in test scores may be just as interesting as within-school changes in test scores, I have included Figs. A.1, A.2, and Table A.1, which report the main results for alternative specifications. The results are similar regardless of which controls are used, including school-by-year fixed effects. I have chosen to include school fixed effects because teachers may be influenced by both their individual ratings and school ratings. By including school fixed effects, I control for the school ratings and therefore isolate the effect of the teachers' individual ratings on teachers' performance. Due to data limitations, the model does not contain some student demographics including race, gender, and free- or

Table 2
Change in classroom averages from 2009–2010 to 2010–2011.

	Teacher effectiveness quintile						
Variables	Least	Less	Average	More	Most	Least–most	p-Value
Math CST	0.200	0.179	0.145	0.105	0.057	0.143	0.000
ELA CST	0.138	0.074	0.091	0.042	0.029	0.109	0.000
Prior math CST	0.019	0.035	0.041	0.035	0.064	−0.044	0.148
Prior ELA CST	0.036	0.049	0.061	0.044	0.086	−0.051	0.132
Parent's educ (years)	0.050	0.073	0.047	0.046	0.117	−0.067	0.133
ELL	0.021	0.024	0.019	0.030	0.023	−0.002	0.858
Class size	−0.072	−0.113	0.130	−0.207	0.000	−0.072	0.737

Note: The first five columns report the change in the classroom mean for the indicated variable between the school years 2009–2010 and 2010–2011 for teachers in each Times-rating quintile. The Times-rating quintiles used are from the August 2010 Times ratings. All math and English Test scores are normalized to the 2005 test score distribution by grade. Each year contains approximately 130,000 students with 25,000 to 30,000 students in each quintile.

reduced-lunch status. This model carries the identification assumptions for attributing student test score growth to teachers described in Chetty et al. (2014a) and Rothstein (2010). To get unbiased results, student assignment to teachers, conditional on observables, should be independent of the error term. The model in this paper includes students' lagged test scores which when include have been shown to eliminate most bias (Chetty et al., 2014a).[6] However, omitted variables could still be causing a similar bias to the estimates in each year. In addition, there may be bias to the estimates if positive sorting of high-achieving students to high-rated teachers occurs once the ratings are released that is not controlled for by students' lagged test scores and other observables. However, bias from positive sorting would be in the opposite direction of the results and would tend to shrink the estimates toward zero.

The main results will look at how the coefficient β_k changes over time and particularly how it changes between the school years 2009–2010 and 2010–2011 when both the Times ratings and the AGT ratings were released. Note the model contains only the August 2010 Times ratings. The model does not contain the AGT rating, because they are unavailable.[7] Because the August 2010 Times ratings (V_j) are so highly correlated with the 2011 Times ratings ($\rho = 0.912$), these variables are essentially the same, and only the August 2010 ratings are included.[8] Changes in the coefficient β_k between the school years 2009–2010 and 2010–2011 could be attributed to the August 2010 Times ratings release, the May 2011 Times ratings release, or if the AGT ratings are correlated with Times ratings, the AGT ratings release. Since these August 2010 and May 2011 Times ratings are so highly correlated, they should be viewed as having essentially the same ratings publicly released twice, once right before the start of the school year 2010–2011 and once right before the CST testing period in the school year 2010–2011. Since the two Times ratings are so highly correlated, it is unclear whether the results are driven by teachers receiving information about their rating early in the school year and changing behavior throughout the year (e.g. better preparation or focusing more on math and English), or due to last minute changes before taking the test from the May 2011 release of the ratings (e.g. test-taking strategies, incentivizing student effort, or even cheating). Due to the teacher AGT ratings data not being available, and the high correlation between the two Times ratings, the model is unable to determine which of these three ratings releases had the most impact. Thus the model will attribute changes to the coefficient β_k to a treatment that includes all three ratings releases.

One concern with using teachers' student tests scores over time along with the Times ratings is regression to the mean. All the analyses performed in this paper use the August 2010 Times ratings. The Los Angeles Times used student-level data from the school years 2002–2003 to 2008–2009 to create the August 2010 Times valued-added ratings. Positive or negative shocks to teachers' student test scores during these seven years might have led to low or high value-added ratings and resulted in regression to the mean in teachers' student test scores. Because the school years 2002–2003 to 2008–2009 are used to create the Times ratings, regression to the mean in teachers' average student test scores should be most prominent between the school years 2008–2009 and 2009–2010.[9] However, the Times ratings and AGT ratings were not released until the school year 2010–2011. Therefore, the school year 2009–2010 can be used to measure the amount of regression to the mean that occurs, and the school year 2010–2011 can be used to measure the effect of the teacher ratings. In Section 5, the estimates show that little to no regression to the mean occurs between 2008–2009 and 2009–2010. The limited amount of regression to the mean may be due to the large number of years used to create the value-added scores or because Bayesian methods were used to shrink the teacher value-added scores and correct for measurement error. In addition, the limited amount of regression to the mean may be because only 15.2% of the classrooms used to create the ratings came from the last year of the sample and because on average 118 students were used for each rating. Due to the years of data the Times used to create the ratings and the timing of the ratings release, regression to the mean is measurable and does not drive the results of this paper.

Another possible threat to the validity of the results from this methodology would be any significant changes in the LAUSD, particularly changes to the CST, concurrent with the release of the Times ratings. Value-added scores can change when school districts switch tests or when major alterations are made to the test being used (Lockwood et al., 2007; Papay, 2011). Qualitatively little change occurred in the CST over the time span studied. From the school years 2004–2005 to 2011–2012, the number of questions, the type of questions, and the amount of time allocated to administer the test remained the same for both the math and English CST. In addition, over this time period all raw and scale scores were determined using identical procedures. However, one notable change did occur. Between the school years 2009–2010 and 2010–2011, the fourth-grade English CST added a writing portion consisting of one essay administered over an additional 75-minute period. This writing portion was graded by one reader, was given a score of 0, 2, 4, 6, or 8, and was combined with the number of correctly answered multiple-choice questions to determine the raw English CST scores of fourth graders. This change did not affect third- or fifth-grade students' English test or any student's math test.[10]

In addition to the qualitative evidence of CST test stability over time, I also test whether the CST testing regime was stable over time, by estimating the following specification separately for each test subject, grade, and year subgroup:

$$Y_{i,t} = \alpha + \beta Y_{i,t-1} + \gamma_c + \varepsilon_{i,t} \tag{2}$$

where $Y_{i,t}$ is student i's test score in year t, $Y_{i,t-1}$ is student i's prior-year test score, and γ_c is a classroom fixed effect. The coefficient β shows how much the students' prior-year test scores predict their current-year test scores conditional on all factors that vary at the classroom level. If large changes to the CST occur between year $t-1$ and year t, then there would likely be a large dip in the coefficient β in year t that should rebound in year $t+1$. Table A.2 reports these coefficients. Each row reports the coefficients for a given test subject

[6] In addition, when estimating the effect on students in 2012 the results are robust to controlling for students' 2010 test scores instead of their lagged 2011 test scores.

[7] It should be noted that although the teacher AGT ratings are unavailable, the school AGT ratings are publicly available. The correlation between the AGT school ratings and the August 2010 Times school ratings is 0.15 and the correlation between the AGT school ratings and the May 2011 Times school ratings is 0.39 (Imberman and Lovenheim, 2016). Although I cannot calculate the actual correlation between the teacher AGT ratings and the Times teacher ratings, I have used the technical report for the AGT ratings and have used my data to create ratings as similar as possible to the AGT ratings. While creating these simulated AGT ratings I was unable to control for race, ethnicity, free and reduce price lunch status, gender, and these same controls at the classroom level for a student's peers. The correlation between the simulated AGT ratings that I created and the August 2010 Times ratings is 0.80 and 0.79 for math and English, respectively. When I estimate an alternative specification for the main results that additionally controls for the simulated AGT ratings, I obtain very similar drops in the main estimate when the ratings are released. The results from this alternative specification can be seen in Fig. A.3.

[8] When I estimate an alternative specification for the main results that additionally controls for the difference between the 2010 and 2011 Times ratings, I obtain very similar results. The results from this alternative specification can be seen in Fig. A.4.

[9] Under reasonable assumptions about the shocks to teachers' student test scores, regression to the mean should be largest between the school years 2008–2009 and 2009–2010 and decline with time.

[10] I perform the analysis for the main results on a sample that excludes teachers in the fourth grade. These results are reported in Fig. A.5 and very similar to the main results when using the full sample.

and grade over time. For each test subject and grade, the coefficient on students' prior-year test scores appears to be relatively stable. Particularly, there appears to be no clear decrease in the coefficients in 2011 when the ratings were released that then rebounds in 2012.

5. Results

In this section I present the results from the above empirical methods. I first report the results for the main research question which shows how teachers' student test scores change when teachers are informed of their low or high individual Times rating. I then move from teacher rating to looking at the impact of school ratings. Using an analogous methodology as with the teacher ratings, I show how the school Times ratings affected teachers' performance conditional on the teachers' individual Times ratings. In addition to the effect of teacher and school ratings on student performance, I also look at how the ratings impact non-performance outcomes. For the first of these non-performance outcomes, I describe how classroom composition changes for teachers with a low or high Times rating. Lastly, I look at how teacher turnover differs for low- and high-rated teachers after the ratings release.

5.1. Teacher ratings

To start, I show how the relationship between teachers' ratings and teachers' performance evolved over time. I estimate Eq. (1) using data on students from all years and from all teachers' classrooms. Fig. 5 plots the coefficients on the year-rating interaction terms from 2005–2006 to 2011–2012. Each estimate shows its 95% confidence interval using standard errors clustered at the teacher level. Each point in Fig. 5 represents the relationship between a teacher with a standard-deviation-higher Times rating and student test scores in the indicated year.[11] The change in this relationship can be seen over time. The vertical dashed line represents the release of the Times ratings. It should be noted that this estimation is related to, but differs, from the analysis shown in Fig. 4. Fig. 4 shows the change in student test scores in teachers' classrooms from year to year without any controls, whereas, the results in Fig. 5 uses the controls shown in Eq. (1) and report the relationship between having a higher-rated teacher in a given year and students tests (not the change from year to year).

Panel (a) shows that in the school year 2005–2006 students who had a teacher with a standard-deviation-higher Times rating scored 0.248 standard deviations higher on their math test scores. For English, having a standard-deviation-higher-rated teacher increases English tests scores by 0.165 in 2005–2006. From 2005–2006 to 2009–2010, the estimate for math rises to 0.306 and the estimate for English rises to 0.193. As can be seen in Fig. 6 this rise in the estimates is driven by high-rated teachers. Since student data from 2002–2003 to 2008–2009 were used to calculate the Times value-added ratings, regression to the mean will be most prominent between the school years 2008–2009 and 2009–2010. If regression to the mean were to occur, the estimated coefficient would drop sharply between 2009 and 2010 in each of the panels. However, little to no regression to the mean appears to occur between the school years 2008–2009 and 2009–2010. The lack of regression to the mean between these two school years implies that the random element in student test scores that may be affecting the valued-added ratings is relatively small. An additional benefit of having the teacher ratings released in August 2010 is that teachers were already committed to teaching for the school year 2010–2011. This fact allows for little selective attrition in the school year 2010–2011 due to the information from the teacher ratings. Although selective attrition may affect the school year 2011–2012, it should have little influence on the school year 2010–2011. Teacher turnover is discussed in more detail in Section 5.4.

In panel (a), the year 2011 shows the relationship between having a teacher with a standard-deviation-higher Times ratings and students' math test scores in the year in which both the Times ratings and the AGT reports were released. In the year both value-added ratings were released, there was a decrease in the relationship between a standard-deviation-higher-rated teacher and both math and English test scores. Once teachers were informed of the Times ratings, the estimate for math falls from 0.306 to 0.245. This decrease is a 20% decline in the benefit of having a higher-rated teacher. The estimate for English falls from 0.193 to 0.148, or a 23% decline. These results indicate that after the release of the Times ratings, there was a weaker relationships between a teacher's rating and their students' test scores than before the ratings release. This indicates that the release of the teacher ratings corresponded to a compression in the teacher-performance distribution, because now moving a standard-deviation in the teacher-performance distribution has less of an impact on student test scores.

In addition to the change seen in 2010–2011 in which the gap between the top- and bottom-rated teachers is reduced and the distribution of teacher performance is compressed, this compression of the teacher-performance distribution persists for at least two years. The year 2012 shows the change that occurred between the school years 2009–2010 and 2010–2011 with the release of the value-added ratings does not disappear the following year, but continues. In panel (a), the year 2012 shows students who had a teacher with a standard-deviation-higher Times rating scored 0.223 standard deviations higher on their math test, which is a 27% decrease from the school year 2009–2010. For English, the effect size in the year 2012 is 0.122 standard deviations, which is a 37% decrease from the school year 2009–2010. It is possible that there was an even larger decrease by 2012 because the Times released ratings in both August 2010 and May 2011. During the 2010–2011 school year teachers would have been informed of their ratings twice (once 9 months before the test and once right before the test). By 2012 teachers would have not only been remind twice but would also have had more time to potentially respond. In addition, if the AGT impacted the teachers, by 2012 they would have received the AGT reports twice instead of just once. Overall these results show that once the Times ratings were released, the distribution of teacher performance was compressed over the next two years.

Fig. 5 indicates the release of the Times and AGT ratings corresponded to a compression in the teacher-performance distribution. However, Fig. 5 does not show where in the teacher performance distribution this compression occurs. Fig. 6 helps demonstrate where this compression occurs. Fig. 6 shows the results from estimating Eq. (1) with the teachers' Times-rating variable (V_j) replaced with binary variables for each of the five Times-rating quintiles.[12] The coefficient on the interaction for the year 2006 and the least effective quintile is normalized to zero. The level differences in quintile lines shows the difference in teacher performance between teachers in different quintiles. The compression of the teacher-performance distribution between 2010 and 2011 appears to be due primarily to an increase in performance by low-rated teachers, with little change in performance by higher-rated teachers. This pattern is similar for both math and English. However, if the upward trend for the higher-rated teachers between 2006 and 2010 would have continued, then the lack of change in performance for higher-rated teachers may imply a reduction in performance for higher-rated teachers. In 2012,

[11] Figs. A.6–A.8 show an alternative depiction of these slopes in each year by plotting the relationship between student test scores and the Times rating for each year.

[12] Fig. A.9 reports analogous results using a leave-year-out value-added score instead of the Times ratings.

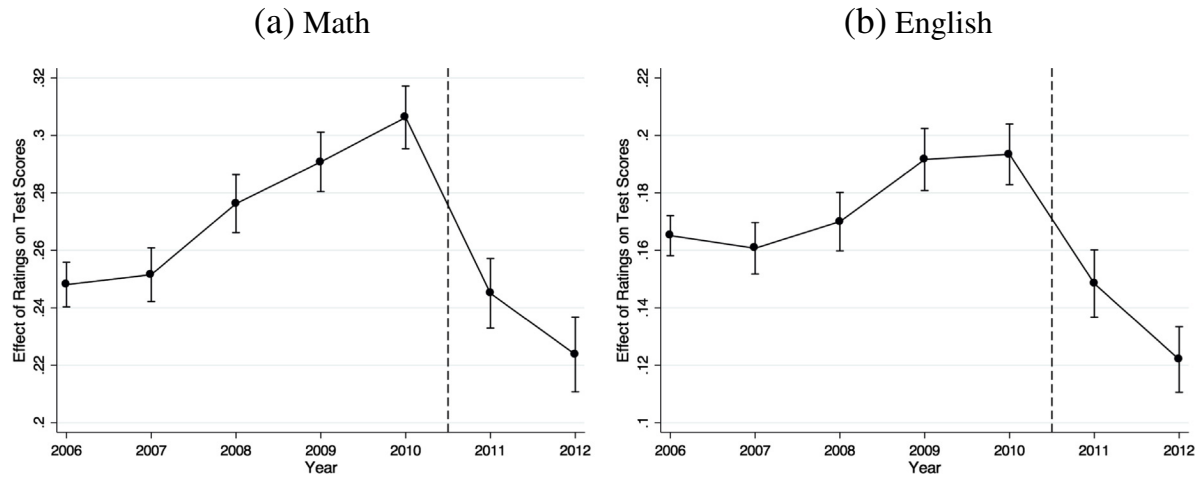

Fig. 5. Effect of teacher ratings on performance by year. Note: This figure plots the estimated coefficients on the year-rating interaction terms from 2005–2006 to 2011–2012 from Eq. (1). This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects. Each point in this figure represents how much a teacher with a standard-deviation-higher value-added rating in the indicated year increases student test scores. The vertical dashed line represents the release of the Times ratings. Test scores are normalized to the 2005 test score distribution by grade. Each year contains approximately 100,000 students. All 95% confidence intervals use standard errors clustered at the teacher level. For the exact point estimates and standard errors see Table A.4.

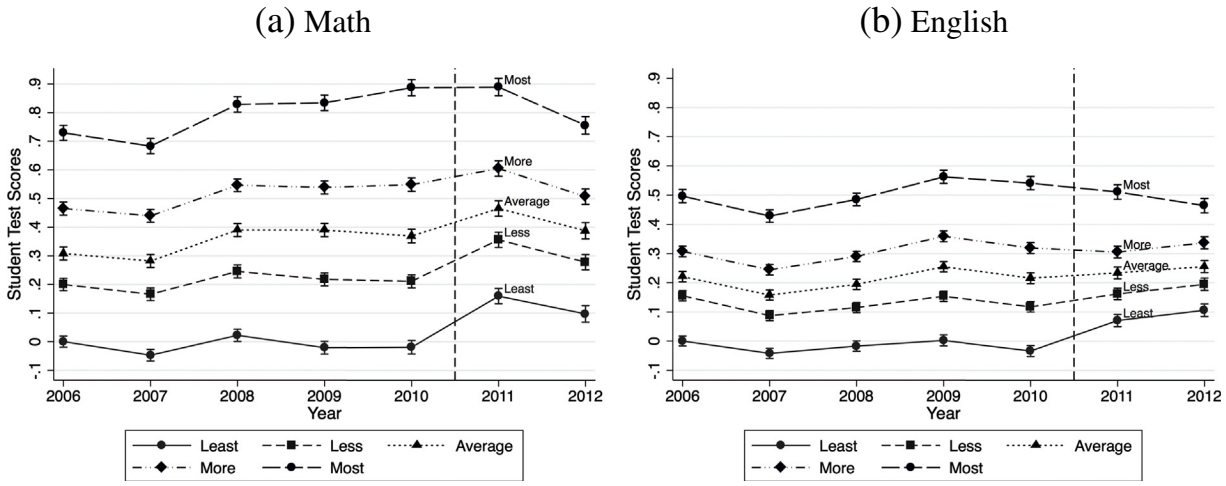

Fig. 6. Effect of teacher ratings by quintile. Note: This figure plots the estimated coefficients on the year-quintile interaction terms from Eq. (1) when V_j is replaced with a vector of binary variables for each Times ratings quintile. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects. Each point represents how much a teacher in the indicated quintile and year increases student test scores compared to a bottom-quintile teacher in 2006. Test scores are normalized to the 2005 test score distribution by grade. The vertical dashed line represents the release of the Times and AGT ratings. Each year contains approximately 100,000 students with 20,000 students in each quintile. All 95% confidence intervals use standard errors clustered at the teacher level. For the exact point estimates and standard errors see Tables A.5 and A.6.

some additional compression occurs in the teacher-performance distribution. These results suggest that the release of teacher ratings could possibly help improve student test scores by increasing the performance of low-rated teachers while having a minimal impact on high-rated teachers.

5.2. School ratings

In addition to the teacher ratings, both the Times ratings and the AGT reports publicly released school value-added ratings for 470 elementary schools. In this section, I examine whether the release of these school ratings had an impact on the overall performance of the teachers at these schools. Just as with the teacher ratings, I will not be able to identify whether the effects found here are due the August 2010 Times ratings, May 2011 Times rating, or the AGT reports. There are several ways in which the school ratings could impacted teachers in addition to their individual ratings. The overall performance of teachers might increase (or decrease) if the school ratings caused principals or a contingent of teachers to make changes at the school (e.g. more collaboration among teachers or additional training) or if teachers have a strong sense of community at their school and are highly motivated (or demotivated) by the collective school rating.

To examine the impact of the release of the school ratings, I use a modified version of Eq. (1). Eq. (1) is estimated with V_j representing the Times school rating, the teacher's individual Times rating included as an independent variable, and no school fixed effects included as controls. This approach will estimate the relationship between school ratings and student test scores conditional on teachers' individual ratings.[13] Therefore, Fig. 7 shows the relationship between the test

[13] Most of the variation in the Times school value-added ratings can be explained by the aggregated teacher value-added ratings. This indicates that the school ratings gave little additional information to teachers, parents, and students that could not be obtained from the teacher ratings. However, most of the variation in the teacher ratings was within schools and not between schools.

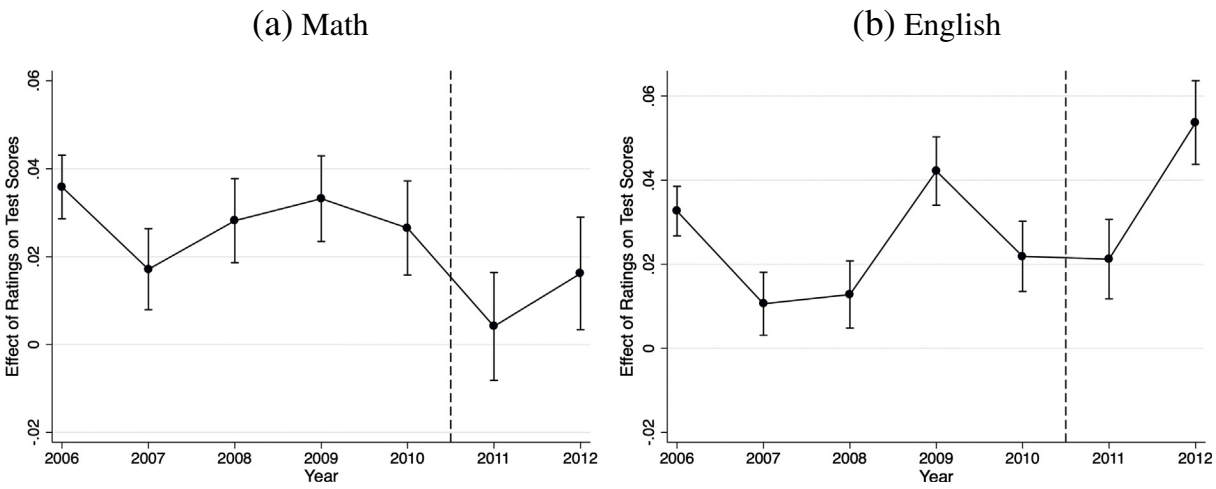

Fig. 7. Effect of school ratings on performance by year. Note: This figure plots the estimated coefficients on the year-rating interaction terms from 2005–2006 to 2011–2012 from Eq. (1) where V_j represents the Times school rating and the teacher's individual Times rating is included as an independent variable. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade and year fixed effects. Each point in Fig. 7 represents how much a school with a standard-deviation-higher value-added rating in the indicated year increases student test scores conditional on teachers' individual Times ratings. The vertical dashed line represents the release of the Times ratings. Test scores are normalized to the 2005 test score distribution by grade. Each year contains approximately 100,000 students. All 95% confidence intervals use standard errors clustered at the teacher level. For the exact point estimates and standard errors see Table A.7.

scores of students with similar-rated teachers yet different-rated schools and their school rating over time. Analogous to Fig. 5 for individual teachers, Fig. 7 plots the coefficients on the year-rating interaction terms from 2005–2006 to 2011–2012.

Fig. 7 shows that conditional on students' characteristics and teachers' ratings, the school ratings had little additional impact on student test scores.[14] More importantly, Fig. 7 shows that student test scores did not appear to change systematically when the Times ratings were release. For English, I find no evidence of change between 2010 and 2011, while there is a statistically significant increase in the estimate for English in 2012. However, this increase is similar to that found between 2008 and 2009. Similarly for math, there is a small drop between 2010 and 2011; however, this drop is similar in size and statistical significance to the drop between 2006 and 2007 and mostly rebounds in 2012. In general, I find no clear evidence that the school ratings affected teachers' performance. However, school ratings might have had an impact on student test scores if no individual ratings were reported.

5.3. Classroom composition

Knowing whether the classroom composition of high- and low-rated teachers changed with the release of the ratings is important. One of the main arguments for releasing teachers' value-added scores publicly rather than to individual teachers privately is that parents have the right to know the quality of teaching their children are receiving and should be allowed to act on this knowledge. The release of the ratings could change classroom composition in many ways, such as by giving additional information to principals to make classroom assignments, giving more negotiation power to high-rated teachers, or by influencing parents to request or lobby for their child to be placed in a high-rated teacher's classroom. Jacob and Lefgren (2007) find evidence that a substantial portion of parents request particular teachers and are able to influence classroom assignment. In addition, Clotfelter et al. (2006) find that non-random sorting is present in many schools. However, schools may be rigid in their classroom assignments, and the release of the Times ratings may not affect classroom composition. Also, the different ways in which classroom composition could change (e.g. giving principals more information, giving high-rated teachers more negotiation power, and informing parents lobbying) may offset each other, and no change in classroom composition will occur in the aggregate. In addition, since the Times ratings were not released until August of 2010, it is likely that most classroom assignments would have already been made for the 2010–2011 school year and therefore potentially limiting the impact on classroom composition in the first year.

Fig. 8 shows how the classroom composition changed for high- and low-rated teachers over time. Each point in Fig. 8 shows the relationship between a standard-deviation-higher rating and a teacher's classroom composition in the given year. No systematic change appears to occur in the relationship between a standard-deviation-higher-rating and teacher composition for the two years after the release of the teacher ratings to the years before. The results show little to no change in the composition of students in low- or high-rated teachers' classrooms after the release of the Times ratings. Panel (a) shows that in 2006, teachers with a standard-deviation-higher rating on average had students with 0.042 standard-deviation-higher prior math test scores. This estimate rose slowly over then next six years. However, importantly, no large change occurred between 2010 and 2011. Panel (b) shows similar results for the average prior English test scores for teachers. Panel (c) of Fig. 8 shows the relationship between a standard-deviation-higher rating and the average years of parents' education. Similar to students' prior test scores, no systematic difference for teachers appears to exist for the two years after the release of the teacher ratings compared to the years before. For all four variables shown in Fig. 8, the change between2010 and 2011 was not statistically significant, with p-values ranging from 0.442 to 0.572. The results show little evidence that the ratings had any impact on classroom composition. This finding is in line with Imberman and Lovenheim (2016) who find the Times ratings were not capitalized into housing prices.[15]

[14] Conditional on students' characteristics and teachers' ratings, on average having a standard-deviation-higher-rated school rating increased math test scores by 0.023 standard deviations and English test scores by 0.027 standard deviations.

[15] If schools do not have strict policies on classroom size, parent lobbying may increase the number of children in high-rated teachers' classrooms after the release of the ratings. Panel (d) of Fig. 8 shows the change in the relationship between a standard-deviation-higher rating and the classroom size of a teacher's classroom from year to year. No systematic difference for teachers appears to exist for the two years after the teacher ratings were released compared to the four years before.

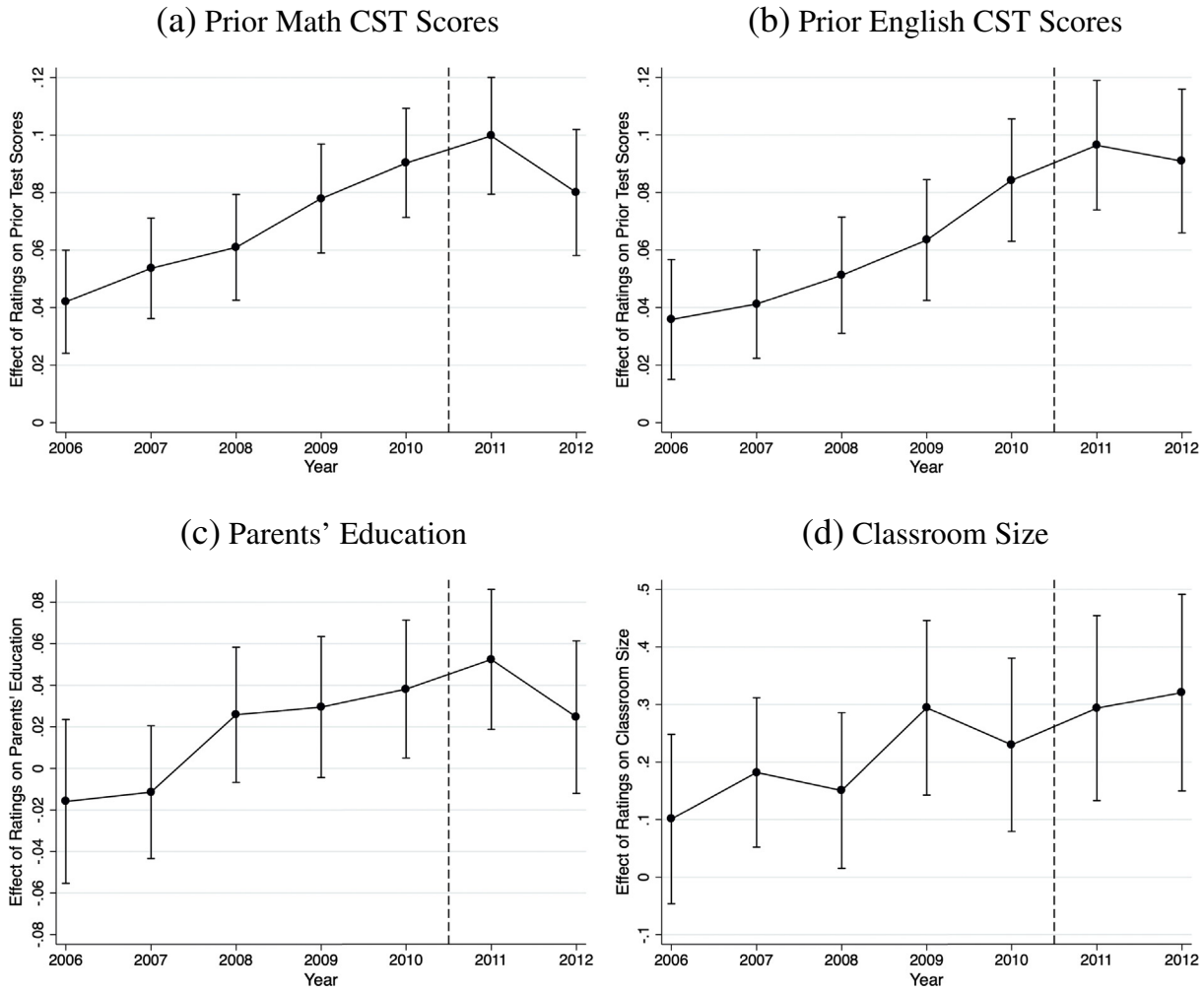

Fig. 8. Change in classroom composition. Note: This figure plots the estimated coefficients on the year-ratings interactions from the following equation: $W_{j,t} = \alpha + \sum_{k=0}^{6} \beta_k Year_{2006+k} * V_j + \sum_{k=1}^{6} \mu_k Year_{2006+k} + \theta_{s,t} + \varepsilon_{i,t}$ where $W_{j,t}$ represents the average of the indicated classroom composition variable for teacher j in year t. Each point represents how the average indicated variable increases for a teacher with a standard-deviation-higher value-added rating. Prior test scores are normalized to the 2005 test score distribution by grade. The vertical dashed line represents the release of the Times and AGT ratings. Each year contains approximately 100,000 students. All 95% confidence intervals use standard errors clustered at the teacher level. For the exact point estimates and standard errors see Table A.8.

5.4. Teacher turnover

Besides affecting teachers' performance, the Times ratings could also affect teachers' decision to leave the school district or switch to a new school or grade. Rockoff et al. (2012) and Sartain and Steinberg (2016) find that new evaluation information causes a higher likelihood of job separation for teachers with low performance. With the release of the Times ratings and the AGT ratings, third- through fifth-grade teachers in the LAUSD were under heightened public scrutiny from students, parents, and other teachers. Although the ratings played no role in teacher evaluations or pay, this additional scrutiny may have caused low-performing teachers to leave voluntarily the school district more than high-rated teachers. To test whether the release of the Times ratings affected teacher turnover, I estimate the following specification separately for each year:

$$E_{j,t} = \alpha + \beta V_j + \theta_{s,t-1} + \varepsilon_{i,t} \quad (3)$$

where $E_{j,t}$ is an indicator for whether teacher j left the school district between year $t-1$ and year t, V_j is teacher j's Times overall rating, and $\theta_{s,t-1}$ is a school fixed effect. Note that since Eq. (3) is estimated separately by year, these school fixed effects should be thought of as school-by-year fixed effects. For this analysis, it is important to consider who teachers are comparing themselves to when making the decision to either leave the school district or switch to a new school or grade. Whether teachers are comparing themselves to other teachers in their school or to other teachers in the school district will determine whether the analysis should use the within or across school variation. I will be using a school fixed effect to keep the specification analogous to those used earlier; however, both specification have value. I report the results without a school fixed effect in Fig. A.10, and find that the result are very similar regardless of whether or not a school fixed effects is included.

The coefficient on the teachers' Times ratings for each year along with their 95% confidence intervals are plotted in panel (a) of Fig. 9.[16] As panel (a) shows, having a standard-deviation-higher rating has no statistically significant association with leaving the school district from 2006 to 2009. This is likely due to the fact that rated teachers must have taught some years in the LAUSD between 2002–2003 and 2008–2009, and therefore rated teachers were unlikely to leave the LAUSD before 2010. However, after 2009, having a standard-deviation-higher

[16] Analogous results by rating quintile can be seen in Fig. A.11.

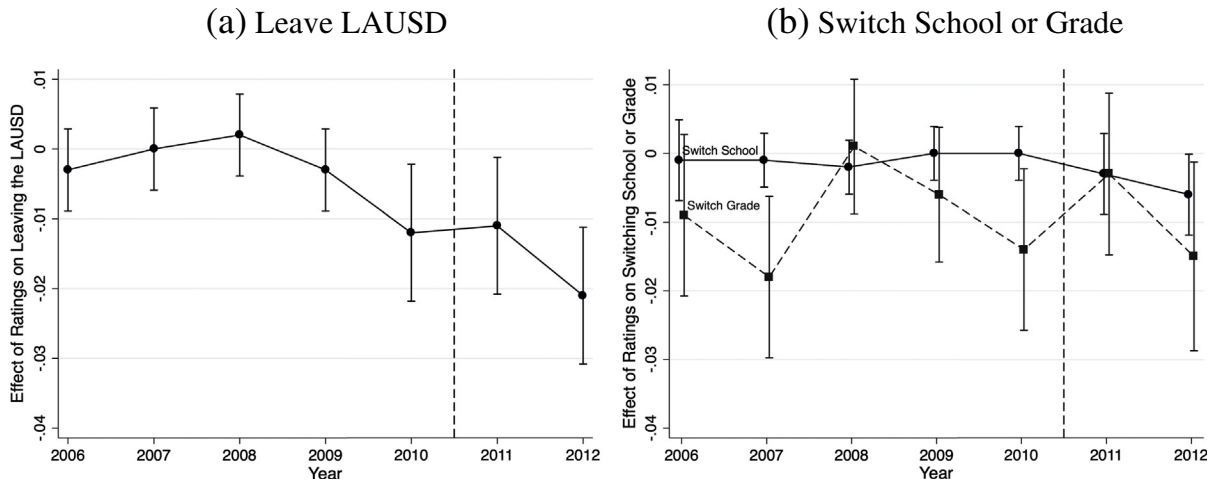

Fig. 9. Effect of teacher ratings on teacher turnover. Note: Each point represents the coefficient on the Times overall rating when the indicated binary variable is regressed on the teachers' Times ratings and a school fixed effect (see Eq. (3)). For each year, the sample size of teachers is approximately 5,000. The vertical dashed line represents the release of the Times and AGT ratings. All 95% confidence intervals use robust standard errors. For the exact point estimates and standard errors see Table A.9.

rating has a statistically significant negative association with leaving the school district. In 2012, having a standard-deviation-higher rating is associated with being 2.1 percentage points less likely to leave the school district. The large teacher layoff in the summer between2009 and 2010 could also have influenced the drop between 2009 and 2010. Panel (b) shows analogous results using whether teachers switched schools (but stayed in the school district) or switched grades (but stayed in the same school) as the outcome variable. The release of the Times ratings does not appear to have a systematic impact on the likelihood of switching schools or grade. In addition, Fig. A.12 looks at (1) whether high-rated teachers are differentially likely to move within the school to grades or subjects that were not rated by the LA Times and (2) for those teachers who moved to a new school, whether high-rated teachers were more like to move to schools with higher performing students. I find no evidence that either of these types of teacher movement are occurring.

In addition to testing whether the release of the Times ratings had an impact on teacher turnover, these results helps determine whether teacher turnover is one of the possible mechanisms for how the Times ratings affect teacher performance found in Section 5.1. If the Times ratings affect teacher turnover in particular ways, teacher turnover could cause for the changes in teacher performance. For example, Jackson (2013) finds that teachers improve their performance when they switch schools, and Ost (2014) finds that teachers' performance declines in the short run when they switch grades. Therefore, if the release of the ratings caused low-rated teachers to switch schools more and high-rated teachers to switch grades more, this could cause the results found in Figs. 5 and 6. As Fig. 9 shows, the release of the Times ratings did not differentially affect low- and high-rated teachers' likelihood of switching schools or grades. Therefore, the main results in Section 5.1 are not due to low-rated teachers switching schools more and high-rated teachers switching grades more.

6. Mechanisms and robustness

There are several possible reasons why low-rated teachers increase their students' test scores when informed of their rating. First, when teachers learn of their math and English ratings, they may reallocate time and energy from their higher-rated subject to their lower-rated subject. They may also reallocate time and energy between math and English and other subjects depending on their rating. Second, based on their ratings, teachers may adjust how much they "teach to the test" by spending more or less time and effort on the specific content and style of questions the CST will cover. Lastly, when parents learn of the Times rating of their child's teacher, they may adjust their at-home support of their child's math and English learning.

6.1. Reallocation between subjects

The first possible mechanism is that teachers may reallocate time and energy from their higher-rated subject to their lower-rated subject. For example, if a teacher received a high math rating and a low English rating, he or she may reallocate classroom time from math to English, causing increases in students' English scores and possible declines in students' math scores. This response would cause the coefficient on teachers' Times rating to fall after the release of the ratings.

To test whether teachers are reallocating time from their high- to low-rated subject, I estimate the following specification separately for students' math and English CST scores in each year:

$$Y_{i,t} = \alpha + \beta(V_{Other} - V_{Same}) + \eta V_{Same} + \delta P_{i,t-1} + \gamma X_{i,t} + C_{i,t} + \theta_{s,t} + \varepsilon_{i,t} \quad (4)$$

where $Y_{i,t}$, is student i's math or English CST score in year t, V_{Other} is the normalized Times rating of student i's year t teacher in the opposite subject as the outcome variable, and V_{Same} is the normalized Times rating of student i's year t teacher in the same subject as the outcome variable. Table 3 reports the estimated coefficients on the subject-difference in teacher ratings and the same-subject rating for each subject and year. The coefficient of interest, β, reports how having a standard-deviation-higher rating in the other subject compared to the same subject affects the same-subject test scores conditional on the same-subject rating and other controls. The following results are very similar when an interaction term between the same-subject rating and the subject-difference in teacher ratings is included in the specification. If reallocation were occurring, the coefficient should be near zero from 2006 to 2010, should increase between 2010 and 2011, and should retain this increase in 2012. For example, if a teacher had a higher English than math rating and therefore reallocated time from English to math when the ratings were released, then his or her students' math scores would improve, causing an increase in the coefficient on the subject-difference in teacher ratings. The results for math in panel A show the coefficient on the subject-difference in teacher ratings is near zero from 2006 to 2012 with no statistically

Table 3
Substitution between subjects.

	2006	2007	2008	2009	2010	2011	2012
Panel A: Math							
English–math rating	0.003	0.015	0.012	0.004	−0.005	0.004	−0.017
	[0.005]	[0.005]	[0.005]	[0.006]	[0.006]	[0.007]	[0.008]
Math rating	0.242	0.253	0.274	0.289	0.302	0.250	0.221
	[0.004]	[0.004]	[0.004]	[0.004]	[0.004]	[0.005]	[0.006]
Observations	88,751	91,052	92,536	95,872	99,807	102,711	94,197
R-squared	0.696	0.688	0.684	0.681	0.670	0.638	0.633
Panel B: English							
Math–English rating	−0.003	−0.001	−0.002	0.007	0.008	0.041	0.052
	[0.005]	[0.004]	[0.004]	[0.005]	[0.005]	[0.006]	[0.006]
English rating	0.161	0.160	0.172	0.190	0.195	0.158	0.135
	[0.003]	[0.003]	[0.003]	[0.003]	[0.003]	[0.004]	[0.004]
Observations	88,667	90,991	92,549	95,851	99,487	102,407	93,753
R-squared	0.739	0.741	0.728	0.727	0.727	0.708	0.715

Note: This table reports the coefficients on the subject-difference in teacher ratings and the same-subject teacher rating when Eq. (4) is estimated for students' math and English test scores. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects. Standard errors are clustered at the teacher level.

significant change between 2010 and 2011 (p-value = 0.329). This stability over time in the coefficient implies little reallocation of time and energy is occurring from English to math. In addition to seeing little reallocation between subjects, the main coefficient on the teacher ratings can be seen in the second row of each panel, similar to Fig. 5.

Although the math results provide no evidence of reallocation, the results for English in panel B provide some evidence of reallocation. The coefficient on the subject-difference in teacher ratings is near zero from 2006 to 2010. Between 2010 and 2011, a statistically significant 0.033 increase occurs in the coefficient (p-value = 0.000) that remains at least through 2012. This increase implies that some of the effect of the Times ratings on students' English test scores is attributable to the reallocation of time and energy from math to English. Since teachers' math and English ratings are highly correlated ($\rho = 0.769$), on average, the difference between a teacher's two ratings is only 0.55 standard deviations. This finding implies that even if the 0.033 increase in the coefficient were applied to both the reallocation from math to English and English to math, reallocation could only explain 20% of the change in teacher performance in English and 14% in math. The results also show an asymmetry in the reallocation effect. The impact of the subject-difference in teacher ratings increases for English when the ratings are released, but not for math. This asymmetry implies teachers are more willing to reallocate time and energy from math to English than they are from English to math. Although the reason for this asymmetry is not clear, it might result from teachers having a stronger aversion to teaching math than English, making reallocation from math to English easier than from English to math. In addition, teachers may find incorporating English learning into math lessons (e.g. by using word problems) easier than incorporating math learning into English lessons. Although the results imply that only a small portion of the effect is coming from reallocation between math and English, reallocation between other subjects and math and English may be occurring. Due to data limitations, I am unable to test for reallocation between other subjects and math and English.

6.2. Teaching to the test

A second possible mechanism for the main results may be due to teachers adjusting how much they "teach to the test" depending on their rating. Here I define "teaching to the test" as teaching methods that help kids perform well on the current test but do not benefit their future test-taking performance. For example, teachers who receive a low rating may increase the amount of time and effort spent on the specific content and style of questions in that year's CST, whereas teachers who receive a high rating may not change their amount of time or effort on these type of items. One way to investigate this mechanism is by examining how persist the benefit of having a standard deviation higher value-added teacher is over time and particularly test whether this persistence changes after the release of the Times ratings. If the test score gains for students in low-rated teachers' classrooms following the release of the Times ratings are due to increased "teaching to the test," then following the release of the Times ratings there should be an increase in the persistence of the beneficial effect of having a standard-deviation-higher value-added teacher. For example, if low-rated teachers increased their "teaching to the test," the effect of having a standard-deviation-higher-rated teacher on test scores in that year will go down. However since teaching to the test only affects the current-year tests scores, the effect of having a standard-deviation-higher-rated teacher on next year's test scores should not change or even possibly go up. Therefore, the persistence of the beneficial effect of having a standard-deviation-higher value-added teacher would increase after the release of the Times ratings. If the test score gains for students in low-rated teachers' classrooms are not due to increased "teaching to the test," there should be no gain in this persistence following the release of the Times ratings.

To examine this potential mechanism, I follow Jacob et al. (2010) and use a two step procedure to estimate the persistence of the effect of having a standard-deviation-higher value-added teacher. In the first step I estimate a leave-year-out value-added score for each teacher in each year. This measure is a value-added score for a teacher in a given year that is estimated using students taught by the teacher in all years expect the given year (i.e. leaving out the students in the given year).[17] In the second step, I used these leave-year-out value-added scores to estimate the persistence of having a standard-deviation-higher value-added teacher over time. To do this I regression students' $t + 1$ test score on their year t teacher's leave-year-out value-added score along with the same controls as shown in Eq. (1) separately for each year. The coefficient on the leave-year-out value-added score reports how much of the benefit of having a standard-deviation-higher value-added teacher persists into the year after being in that teacher's classroom. For the different years and subjects, I obtain estimates of persistence that range from 0.19 to 0.32. These estimates

[17] Instead of using leave-year-out value-added scores, I have also performed this analysis using the Times value-added ratings. However, since the Times ratings use all students, using the Times ratings in the second stage could lead to bias. Despite this potential bias, when I use the Times ratings and a similar methodology, I find quite similar results. These results can be seen in Fig. A.13.

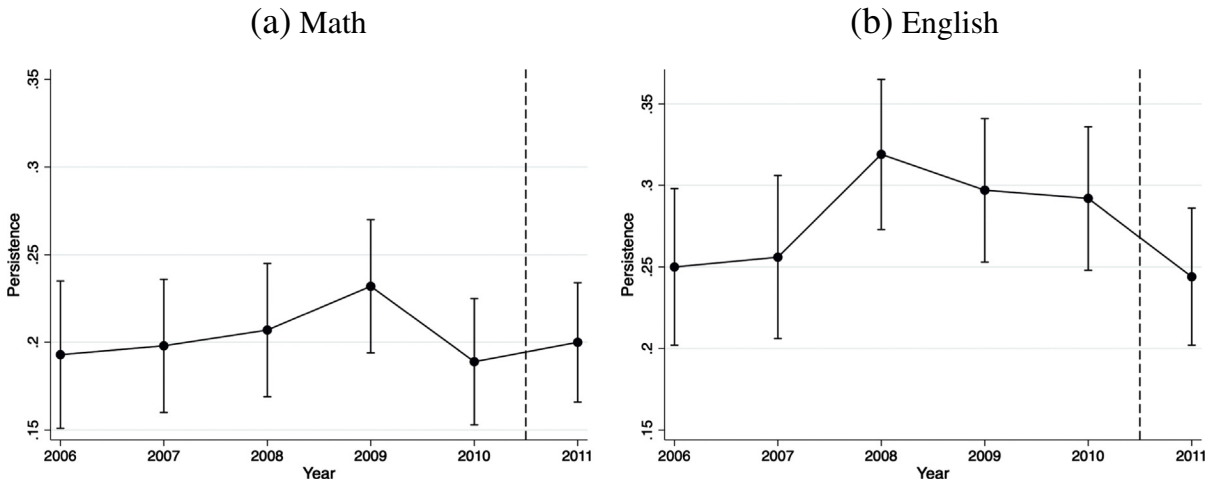

Fig. 10. Persistence of teacher effects over time. Note: This figure reports the persistence of effects on students' math and English tests in the following year from having a standard-deviation-higher leave-year-out value-added teacher. For example, the first estimate for math shows that 19.3% of the positive effect of having a standard-deviation-higher-rated teacher in 2006 persisted one year later in 2007. Each year contains approximately 90,000 students. Standard errors are clustered at the teacher level. The estimates for this figure can be seen in Table A.10. A similar estimation using the Times ratings instead of the leave-year-out value-added measure can be seen in Fig. A.13. For the exact point estimates and standard errors see Table A.10.

are similar to the point estimates found by Jacob, Lefgren, and Sims of 0.20 and 0.27 for English and math, respectively.

Fig. 10 reports the estimates of persistence for both math and English from 2006 to 2011. Each point reports how much of the positive effect of having a standard-deviation-higher value-added teacher persists into the year after having that teacher. For example, the estimate for math in 2006 shows that 19.3% of the positive effect of having a standard-deviation-higher teacher in 2006 persisted one year later in 2007. If the performance gains that occurred for low-rated teachers once the ratings were released are cause by increased "teaching to the test," then there should be an increase in persistence between 2010 and 2011. However, as Fig. 10 shows, the persistence estimates do not meaningfully increase between 2010 and 2011, implying that increased "teaching to the test" by low-rated teachers is not occurring. The change from 2010 to 2011 for both math (p-value = 0.66) and English (p-value = 0.11) is statistically insignificant. These results imply "teaching to the test" does not drive the main results.

6.3. At-home support

The most straightforward way parents may respond to the Times ratings is by moving their child into a high-rated teacher's classroom. However, either due to the rigidness of the schools or parents not knowing or caring about the Times ratings, I found little change in classroom composition following the release of the ratings as shown in Section 5.3. Despite parents not systematically switching teachers, another possible way parents may respond is by adjusting their at-home support of their child's math and English learning (Pop-Eleches and Urquiola, 2013). For example, if parents learn their child is in a low-rated teacher's classroom, they may increase the time they spend with their child learning math or English at home. Without any data available for what is happening in students' homes, I can only indirectly test this explanation. In 2013, college graduates in the Los Angeles Designated Market Area were 2.01 times more likely to have visited the Los Angeles Times website in the last month than were non-college graduates (on a base of 17.8 %).[18] Although this report does not directly inform whether high- or low-educated parents viewed the Times' ratings, and information about the Times' ratings could have spread through other channels besides individually viewing the Times' website, highly educated parents might have been more informed of the Times ratings due to their higher likelihood of using the Times' website. By potentially having more exposure to the Times ratings, highly educated parents might have changed the amount of at-home support they provided their children more than less educated parents. Note that due to the many components of home production and the many variables potentially correlated with parental education, one should view the results cautiously.

If the increase in the amount of at-home support was more for high-educated than low-educated parents, then the release of the Times ratings would likely have a larger impact on students of highly educated parents than those of less educated parents. I perform a heterogeneity test to see if the relationship between teacher ratings and student test scores differs for students with parents that had a high school degree or less and for students with parents that had more than a high school degree. I find no statistically significant difference for students of high- or low-educated parents in both math (p-value = 0.173) and English (p-value = 0.421) (see Table A.3). Note this test is likely to be a weak test for at-home support and therefore is at best supportive, and by no means conclusive. However, this test is also useful as a heterogeneity test, and shows no large differences across students with parents with differing education levels.

7. Conclusion

With the increasing use of value-added scores and a greater push for teacher and school accountability, policymakers will need to make decisions on how and which individuals should receive information on value-added scores. In a court case in New York City[19] regarding the release of teachers' value-added ratings, a four-judge panel stated that, "The reports [value-added ratings] concern information of a type that is of compelling interest to the public, namely, the proficiency of public employees in the performance of their job duties."[20] While these value-added scores may be of compelling

[18] Scarborough L.A. Times Custom Recontact Study 2013 Release 1.

[19] Similarly, in the case between the LA Times and the LAUSD, a judge stated that "The public has an interest in disclosure of the scores because they reflect on both student achievement and teacher performance, as well as on LAUSD's choices in allocating time and resources." http://articles.latimes.com/2013/aug/01/local/la-me-ln-teachers-ratings--20130801.

[20] http://articles.latimes.com/2011/aug/26/local/la-me-nyc-teachers-20110826.

interest to the public, to make policies that best support student learning it is important to understand how teachers respond to the public release of value-added scores at both the individual and group level. By using a natural experiment in which the Los Angeles Times publicly released teacher and school value-added ratings for the Los Angeles Unified School District, I am able to look at how teachers' performance changed when teachers and the public were informed of their value-added ratings. Teachers who were informed they had a low rating saw an increase in their average student test scores. This change narrowed the gap between high- and low-performing teachers and compressed the teacher-performance distribution. When schools and teachers additionally learned of their schools' ratings, student test scores did not change. Although the ratings impacted teachers' performance, I find no evidence that the ratings affected the classroom composition or turnover of high- and low-rated teachers.

There are several ways to consider the magnitude of the effect of the Times ratings release on teachers' student tests scores. The impact of the ratings release on a standard-deviation-lower-rated teacher is the same as improving the teacher match quality by approximately two thirds of a standard deviation (Jackson, 2013). Similarly, the impact of the ratings release on a standard-deviation-lower-rated teacher is similar to the impact of having a 1.2 standard-deviation increase in the average value-added score of peer teachers (Jackson and Bruegmann, 2009). Lastly, this effect size is similar to roughly half of the benefit of a teacher being evaluated (Taylor and Tyler, 2012).

In addition, I rule out multiple possible mechanisms, particularly an increase in teachers "teaching to the test." A remaining possible reason for the change in teacher performance is that low-rated teachers may increase the amount of time and effort they put into teaching math and English instead of other activities such as helping colleagues, teaching other subjects, or leisure time. This change in effort may be due to fear of job loss, a desire for increased respect in the workplace, or altruistic motives.[21] Lack of data, such as the amount of time teachers spend at school or preparing lesson plans, does not allow for analysis to support or refute this plausible mechanism for the differing effect of ratings for low- and high-rated teachers.[22]

These results show a policy that changes the performance of low-performing teachers with a limited impact on high-rated teachers. Although these data and this empirical method are not conducive to determining the overall impact of the ratings, my results suggest the public release of teacher ratings could raise the performance of low-rated teachers without being detrimental to high-rated teachers. This would suggest that teacher ratings may be a useful tool for school districts in improving student performance. Future work is needed to determine if the same benefits can be obtained by privately releasing ratings to teachers without some of the potential social or other unseen costs that may arise from publicly releasing ratings.

Appendix A

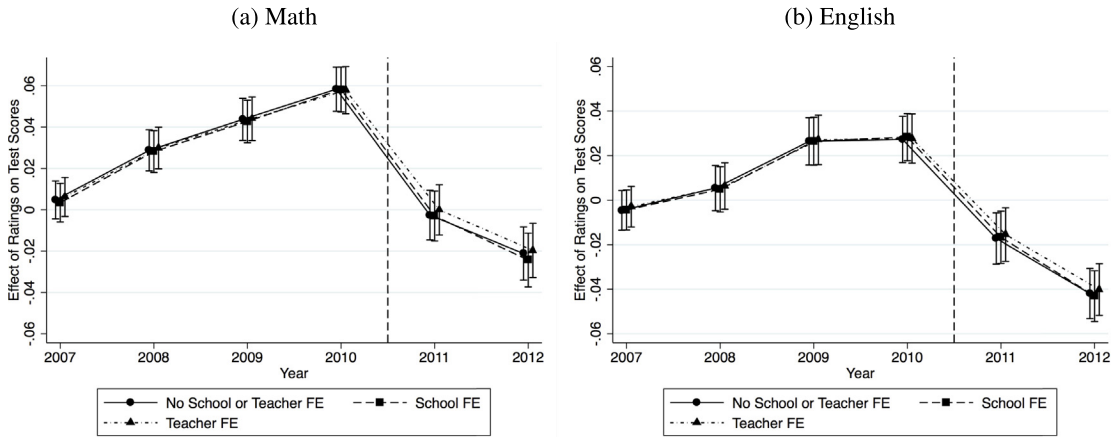

Fig. A.1. Effect of teacher ratings by specification. Note: This figure plots the estimated coefficients on the year-rating interaction terms from 2006–2007 to 2011–2012 from Eq. (1) with school fixed effects, without school fixed effects, and with teacher fixed effects. All models includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. All the models also includes grade and year fixed effects. Each point in Fig. A.1 represents how much a teacher with a standard-deviation-higher value-added rating in the indicated year increases student test scores. The 2006 year is not shown because it is the omitted year when teacher fixed effects are used. The vertical dashed line represents the release of the Times ratings. Test scores are normalized to the 2005 test score distribution by grade. Each year contains approximately 100,000 students. All 95% confidence intervals use standard errors clustered at the teacher level.

[21] Kuhnen and Tymula (2012) support the finding of an increase in performance by low-rated teachers due to effort, and find that when rankings are introduced in a lab experiment, individuals who performed below their expectations subsequently increased their effort and output. The authors suggest that the increase in performance for low performers when rankings are released is likely due to the fight for dominance in the rank hierarchy.

[22] I have also attempted to obtain teacher-absence data from the LAUSD to look at effort effects for teachers (Jacob, 2013), but the LAUSD does not have these data.

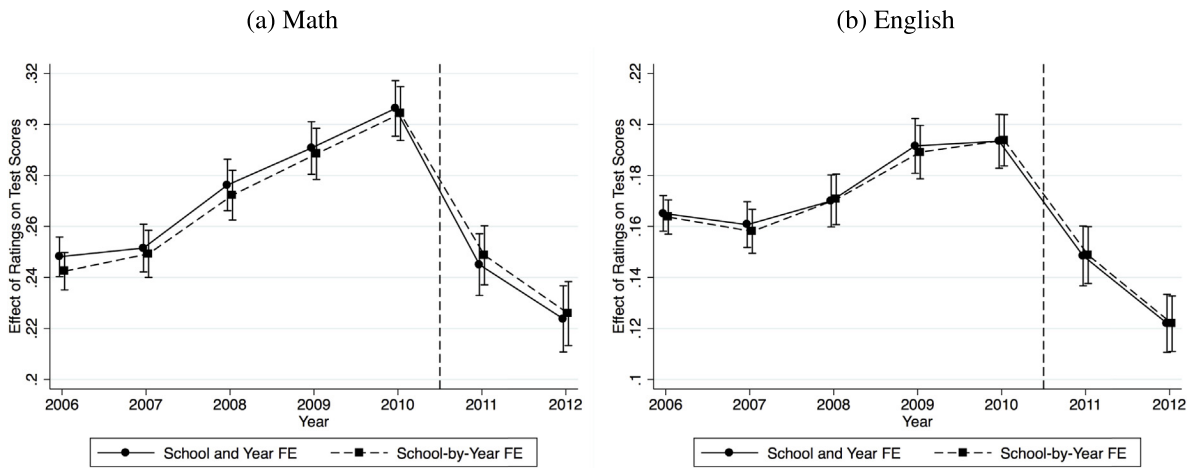

Fig. A.2. School and year fixed effects versus school-by-year fixed effects. Note: This figure plots the estimated coefficients on the year-rating interaction terms from 2005–2006 to 2011–2012 from Eq. (1) with school and year fixed effects and with school-by-year fixed effects. All models include controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. All the models also include grade fixed effects. Each point in this figure represents how much a teacher with a standard-deviation-higher value-added rating in the indicated year increases student test scores. The vertical dashed line represents the release of the Times ratings. Test scores are normalized to the 2005 test score distribution by grade. Each year contains approximately 100,000 students. All 95% confidence intervals use standard errors clustered at the teacher level.

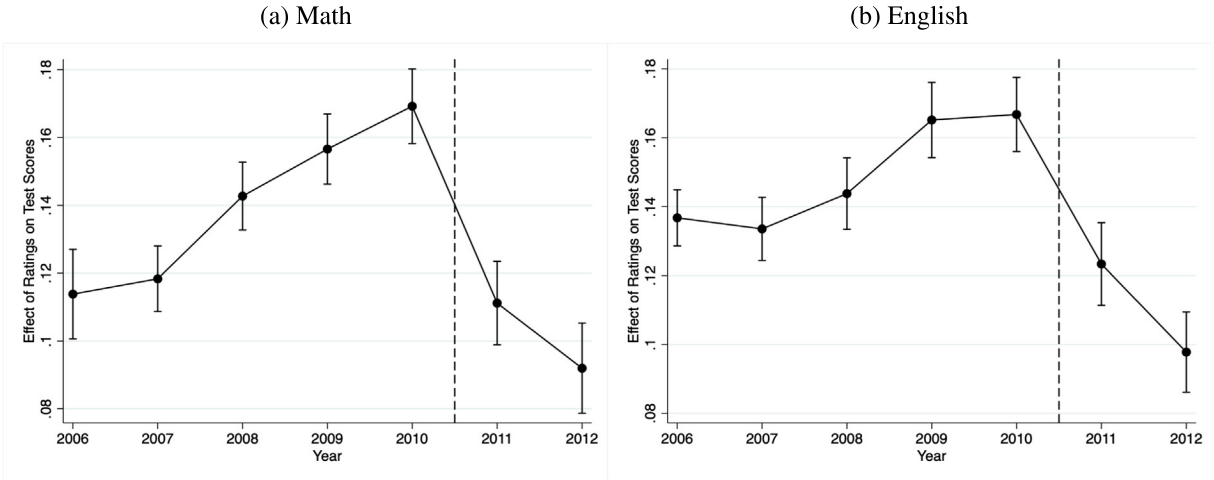

Fig. A.3. Effect of teacher ratings on performance by year controlling for simulated AGT ratings. Note: This figure plots the estimated coefficients on the year-rating interaction terms from 2005–2006 to 2011–2012 from Eq. (1) while additionally controlling for the simulated AGT ratings. This model includes controls for lagged student test scores, parents' education level, ELL status, the classroom average of these variables for other students in the classroom, and the simulated AGT ratings. The model also includes grade, year, and school fixed effects. Each point in this figure represents how much a teacher with a standard-deviation-higher value-added rating in the indicated year increases student test scores. The vertical dashed line represents the release of the Times ratings. Test scores are normalized to the 2005 test score distribution by grade. Each year contains approximately 100,000 students. All 95% confidence intervals use standard errors clustered at the teacher level.

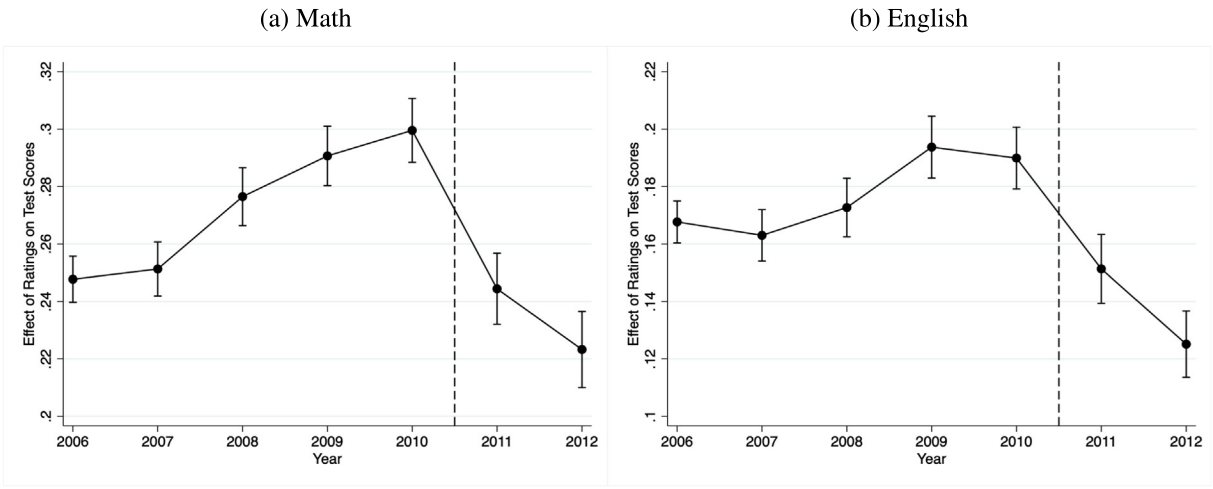

Fig. A.4. Effect of teacher ratings on performance by year controlling for ratings difference. Note: This figure plots the estimated coefficients on the year-rating interaction terms from 2005–2006 to 2011–2012 from Eq. (1) while additionally controlling for the difference between the 2010 and 2011 Times ratings. This model includes controls for lagged student test scores, parents' education level, ELL status, the classroom average of these variables for other students in the classroom, and the difference between the 2010 and 2011 Times ratings. The model also includes grade, year, and school fixed effects. Each point in this figure represents how much a teacher with a standard-deviation-higher value-added rating in the indicated year increases student test scores. The vertical dashed line represents the release of the Times ratings. Test scores are normalized to the 2005 test score distribution by grade. Each year contains approximately 100,000 students. All 95% confidence intervals use standard errors clustered at the teacher level.

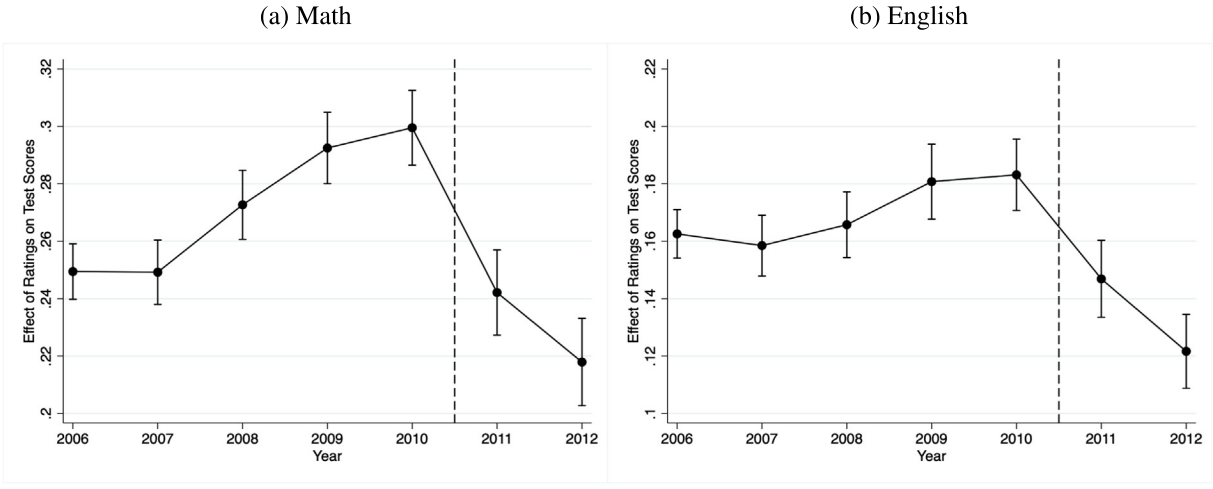

Fig. A.5. Effect of teacher ratings on performance by year excluding 4th grade teachers. Note: This figure plots the estimated coefficients on the year-rating interaction terms from 2005–2006 to 2011–2012 from Eq. (1) when excluding teachers in the 4th grade from the sample. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects. Each point in this figure represents how much a teacher with a standard-deviation-higher value-added rating in the indicated year increases student test scores. The vertical dashed line represents the release of the Times ratings. Test scores are normalized to the 2005 test score distribution by grade. Each year contains approximately 100,000 students. All 95% confidence intervals use standard errors clustered at the teacher level.

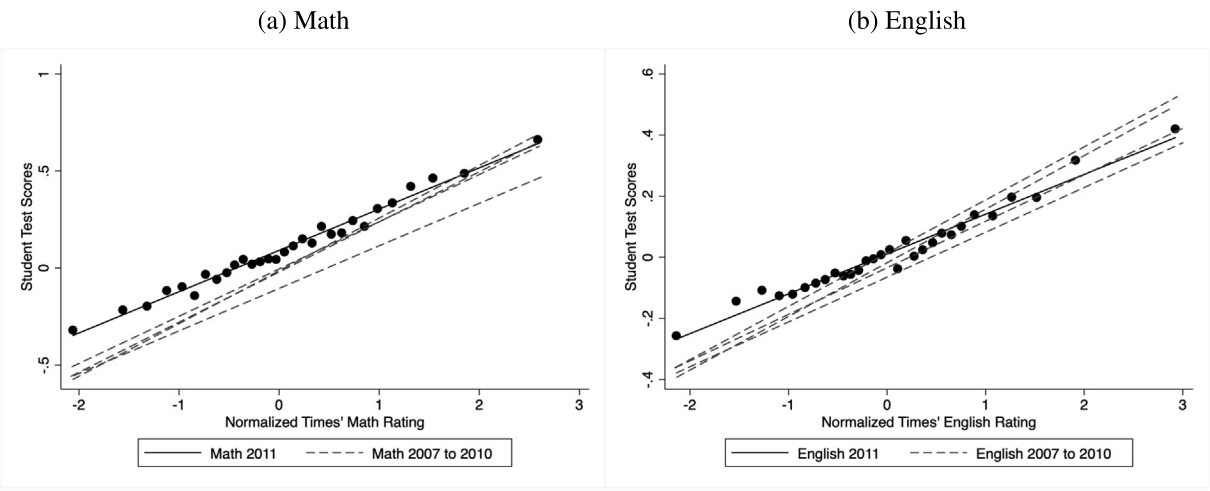

Fig. A.6. Student test scores by teacher rating. Note: Each point represents the average residualized test scores for each bin of students in a teacher's 2010–2011 classroom using the controls from Eq. (1). The solid line is the linear regression through these points. The four dashed lines are the linear regressions for each of the school years 2006–2007 through 2009–2010 and are created analogously to the solid regression line. The slope of these lines are estimated analogously to the points estimates shown in Fig. 5. The points used to create the four dashed lines are not shown in this figure. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects. Each year contains approximately 100,000 students.

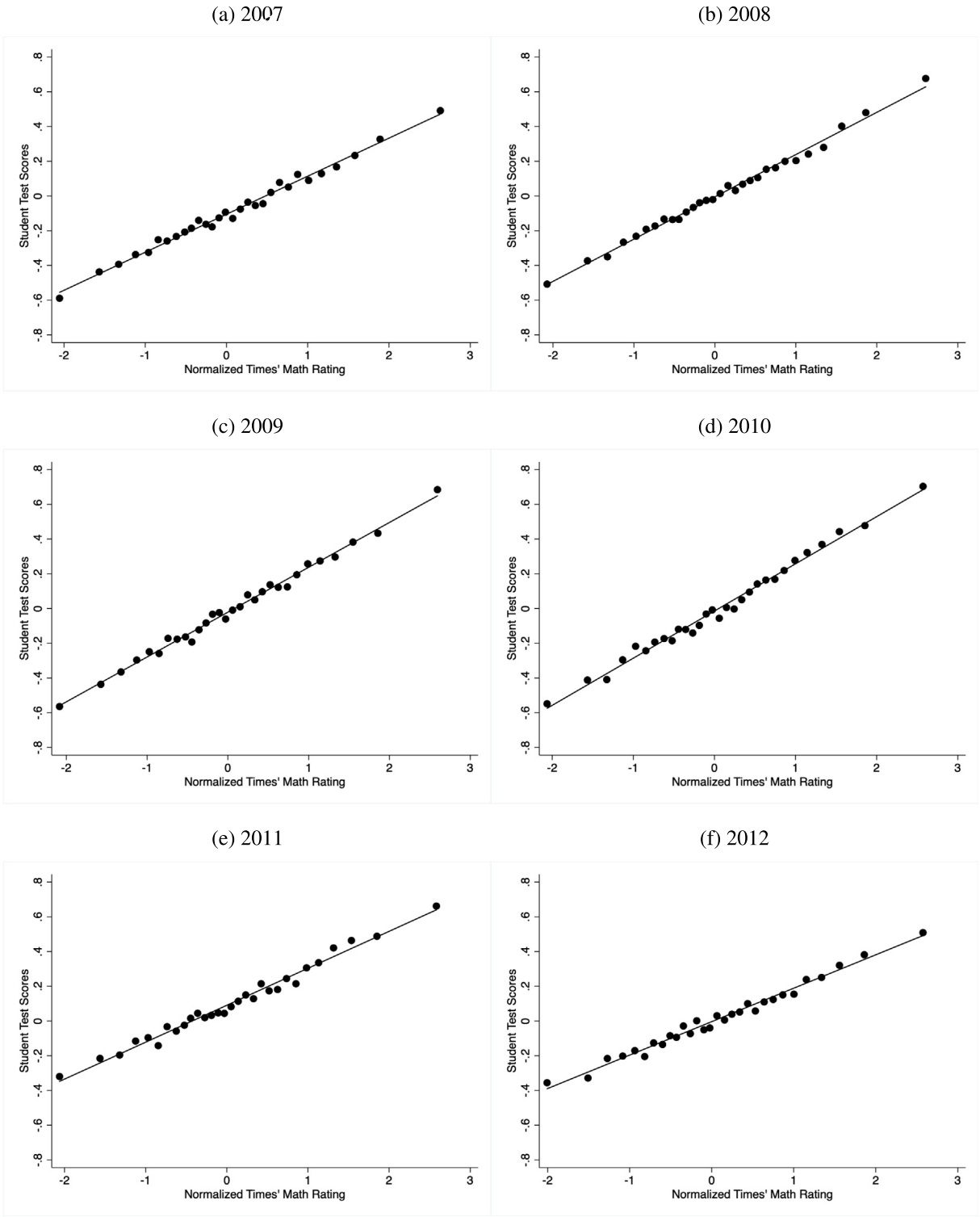

Fig. A.7. Student math test scores by teacher rating for each year. Note: Each point represents the average residualized math test scores for each bin of students in a teacher's classroom in the indicated year using the controls from Eq. (1). The solid line is the linear regression through these points. The slope of this line is estimated analogously to the points estimates shown in Fig. 5. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects. Each year contains approximately 100,000 students.

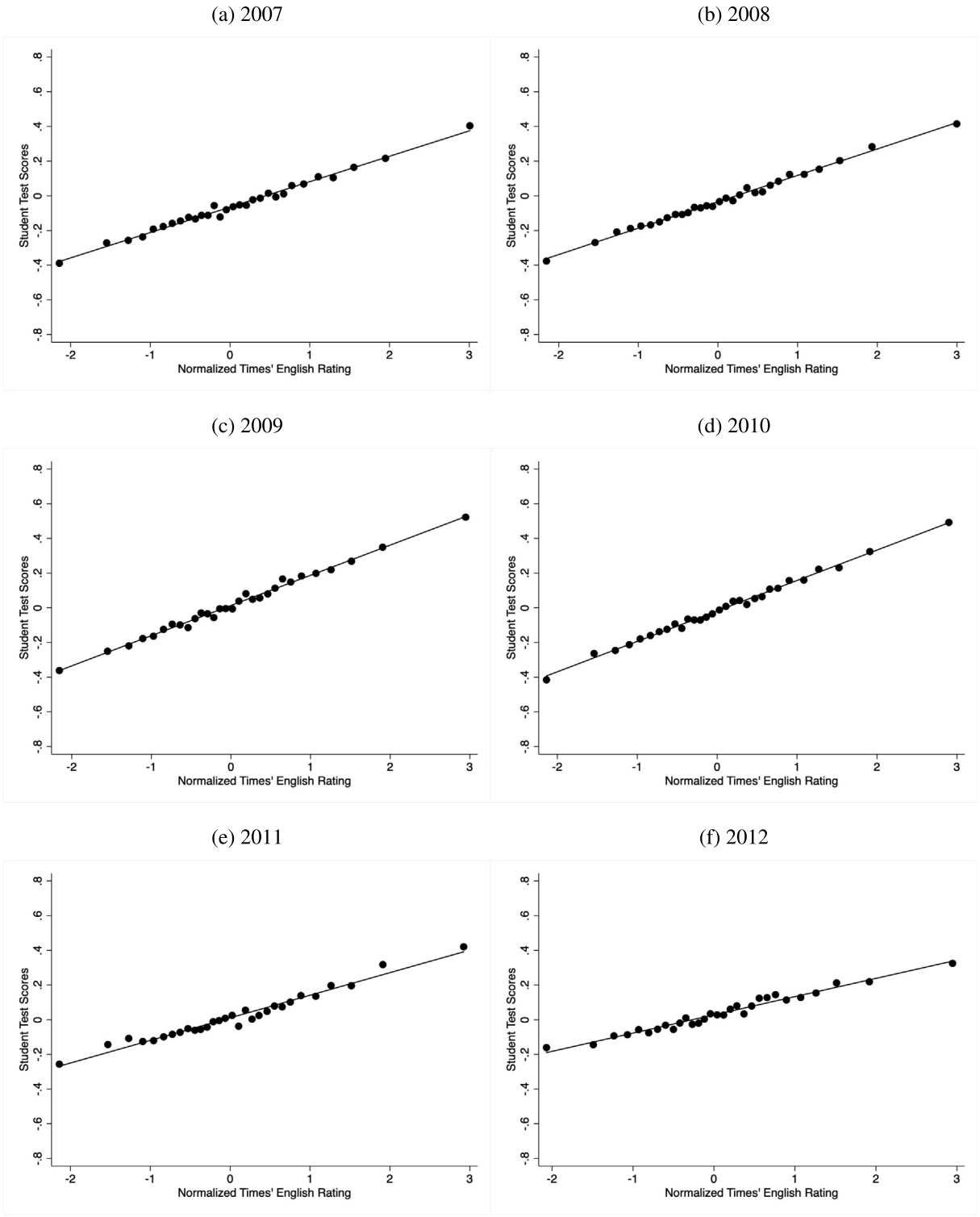

Fig. A.8. Student English test scores by teacher rating for each year. Note: Each point represents the average residualized English test scores for each bin of students in a teacher's classroom in the indicated year using the controls from Eq. (1). The solid line is the linear regression through these points. The slope of this line is estimated analogously to the points estimates shown in Fig. 5. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects. Each year contains approximately 100,000 students.

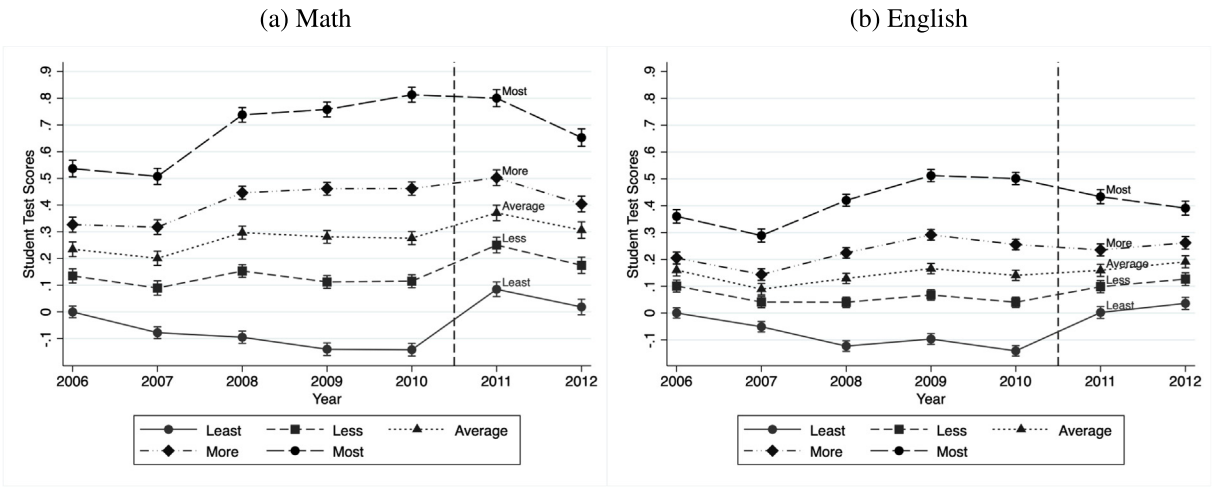

Fig. A.9. Effect of teacher ratings by quintile using leave-year-out value-added score. Note: This figure plots the estimated coefficients on the year-quintile interaction terms from Eq. (1) when V_j is replaced with a vector of binary variables for each quintile of a leave-year-out value-added score created using the school years 2005–2006 to 2011–2012. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects. Each point represents how much a teacher in the indicated quintile and year increases student test scores compared to a bottom-quintile teacher in 2006. Test scores are normalized to the 2005 test score distribution by grade. The vertical dashed line represents the release of the Times and AGT ratings. Each year contains approximately 100,000 students with 20,000 students in each quintile. All 95% confidence intervals use standard errors clustered at the teacher level.

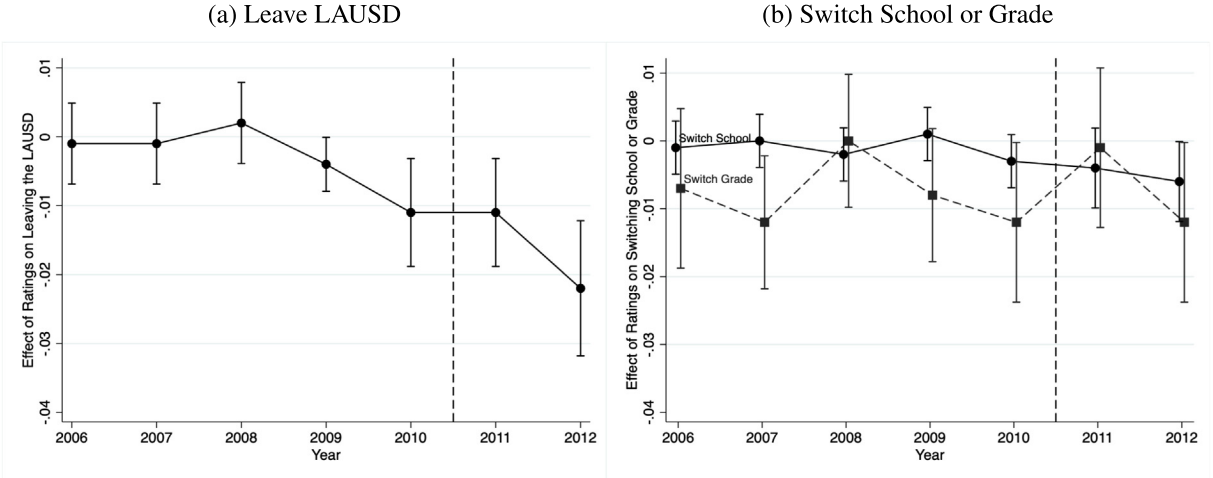

Fig. A.10. Effect of teacher ratings on teacher turnover without school fixed effects. Note: Each point represents the coefficient on the Times overall rating when the indicated binary variable is regressed on the teachers' Times ratings without a school fixed effect. For each year, the sample size of teachers is approximately 5000. The vertical dashed line represents the release of the Times and AGT ratings. All 95% confidence intervals use robust standard errors.

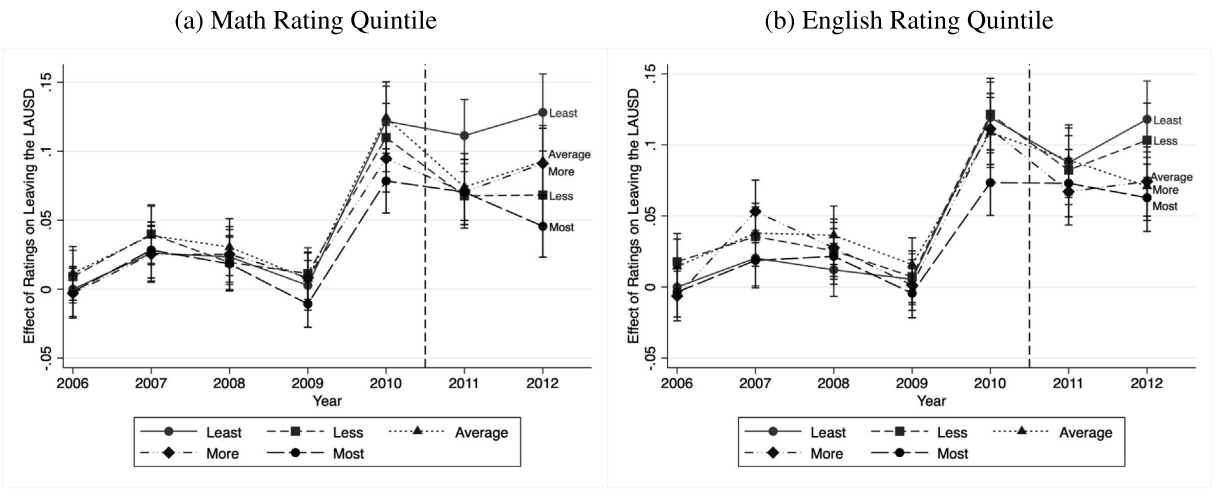

Fig. A.11. Effect of teacher rating on leaving LAUSD by quintile. Note: This figure plots the estimated coefficients on the year-quintile interaction terms when regressing a binary variable for if the teacher left the LAUSD on the year-quintile interaction terms. For each year, the sample size of teachers is approximately 5000. The vertical dashed line represents the release of the Times and AGT ratings. All 95% confidence intervals use robust standard errors.

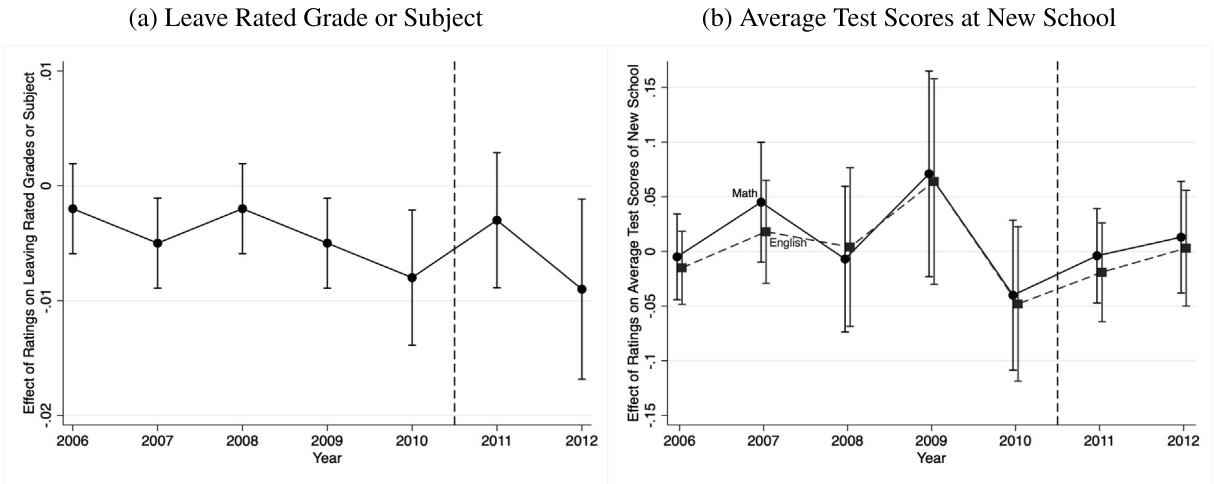

Fig. A.12. Effect of teacher ratings on specific types of teacher turnover. Note: Each point represents the coefficient on the Times overall rating when the indicated variable is regressed on the teachers' Times ratings (see Eq. (3)). For panel (a) the outcome variable is a binary variable for if the teacher moved within the same school but to a grade or subject that was not rated by the LA Times. For panel (b) the sample is restricted to teachers who switched schools within the district and the outcome variable is the average test scores of students at the school the teacher switches to. For each year, the sample size of teachers is approximately 5000 in panel (a) and approximately 200 for panel (b). The vertical dashed line represents the release of the Times and AGT ratings. All 95% confidence intervals use robust standard errors.

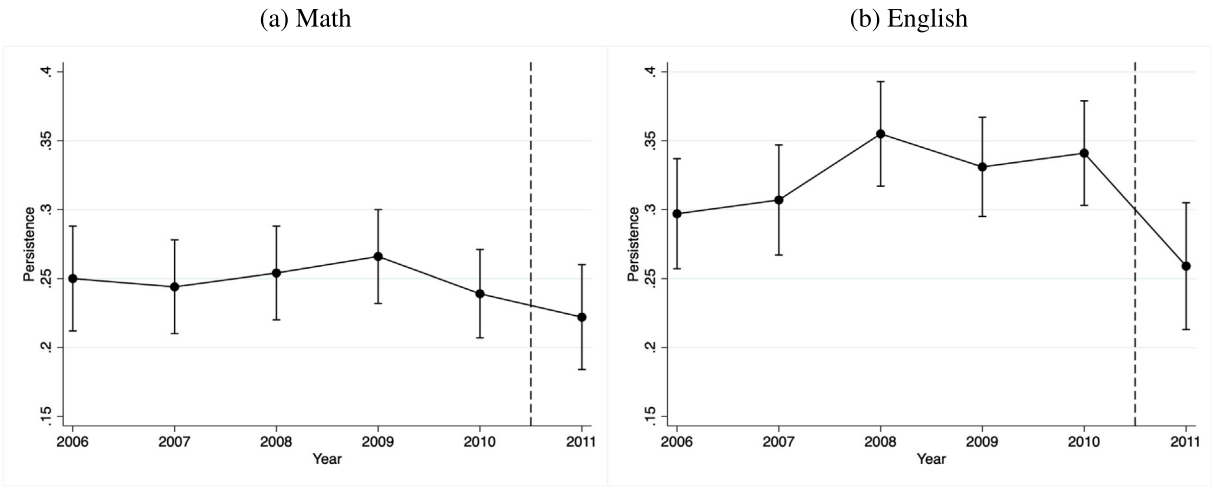

Fig. A.13. Persistence of teacher effects over time using the times ratings. Note: This figure reports the persistence of effects on students' math and English tests in the following year from having a standard-deviation-higher Times rated teacher. For example, the first estimate for math shows that 25% of the positive effect of having a standard-deviation-higher-rated teacher in 2006 persisted one year later in 2007. Each year contains approximately 90,000 students. Standard errors are clustered at the teacher level.

Table A.1
Robustness checks.

	Math					English				
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Post*rating	−0.024***	−0.041***	−0.041***	−0.040***	−0.061***	−0.019***	−0.041***	−0.041***	−0.040***	−0.045***
	[0.007]	[0.005]	[0.004]	[0.004]	[0.006]	[0.007]	[0.004]	[0.004]	[0.004]	[0.005]
Post	0.293***	0.077***	0.084***	0.083***	0.106***	0.278***	0.046***	0.040***	0.040***	0.032***
	[0.007]	[0.004]	[0.004]	[0.004]	[0.006]	[0.007]	[0.004]	[0.003]	[0.003]	[0.004]
Rating	0.316***	0.271***	0.275***	–	0.308***	0.217***	0.175***	0.177***	–	0.194***
	[0.006]	[0.002]	[0.002]	–	[0.004]	[0.007]	[0.003]	[0.001]	–	[0.004]
Prior math CST		0.580***	0.555***	0.557***	0.558***		0.282***	0.156***	0.155***	0.153***
		[0.002]	[0.002]	[0.002]	[0.003]		[0.002]	[0.001]	[0.001]	[0.002]
Prior ELA CST		0.230***	0.236***	0.234***	0.232***		0.531***	0.603***	0.604***	0.601***
		[0.002]	[0.002]	[0.002]	[0.003]		[0.002]	[0.002]	[0.002]	[0.003]
ELL			−0.047***	−0.045***	−0.055***			−0.161***	−0.160***	−0.191***
			[0.003]	[0.003]	[0.005]			[0.002]	[0.002]	[0.004]
Parents educ FE			X	X	X			X	X	X
Grade FE			X	X	X			X	X	X
Peer effect controls			X	X	X			X	X	X
School FE			X		X			X		X
Teacher FE				X					X	
2010/2011 sample					X					X
Observations	662,538	662,538	662,538	662,538	201,460	662,538	662,538	662,538	662,538	201,460
R-squared	0.101	0.652	0.663	0.676	0.647	0.058	0.675	0.724	0.731	0.714

Note: This table shows the robustness of the results for different specifications. All estimates come from the following equation: $Y_{i,t} = \alpha + \beta Post_{i,t} * V_j + \theta Post_{i,t} + \pi V_j + \delta P_{i,t-1} + \gamma X_{i,t} + C_{i,t} + \theta_{s,t} + \varepsilon_{i,t}$ where the binary variable $Post_{i,t}$ is equal to one if the year is after the Times ratings were released. Each row indicates the independent variables included in the specification. Data from the school years 2005–2006 to 2011–2012 are used, except for columns (5) and (10) which restrict the sample to the school years 2009–2010 and 2010–2011. Standard errors are clustered at the teacher level.
*** $p < 0.01$.

Table A.2
Stability of the CST.

	2006	2007	2008	2009	2010	2011	2012
Panel A: Math							
Grade 3	0.714	0.706	0.746	0.769	0.810	0.762	0.741
	[0.006]	[0.006]	[0.006]	[0.006]	[0.006]	[0.006]	[0.006]
Grade 4	0.701	0.626	0.656	0.648	0.641	0.625	0.628
	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]
Grade 5	0.901	0.861	0.879	0.863	0.858	0.830	0.816
	[0.007]	[0.006]	[0.007]	[0.006]	[0.006]	[0.006]	[0.006]
Panel B: English							
Grade 3	0.707	0.651	0.659	0.730	0.744	0.696	0.733
	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]
Grade 4	0.762	0.708	0.720	0.797	0.768	0.725	0.713
	[0.005]	[0.005]	[0.005]	[0.006]	[0.006]	[0.005]	[0.006]
Grade 5	0.823	0.724	0.736	0.815	0.718	0.727	0.726
	[0.005]	[0.004]	[0.004]	[0.005]	[0.004]	[0.004]	[0.005]

Note: Each cell reports the estimated coefficient on prior test scores from Eq. (2) for a given test subject, grade, and year subgroup. The model only controls for classroom fixed effects. All math and English test scores are normalized to the 2005 test score distribution by grade. Each test subject, grade, and year subgroup contains between 29,000 and 39,000 observations. Standard errors are clustered at the classroom level.

Table A.3
Estimates for students of low- and high-educated parents.

Parents education level	Math	English
Low education	−0.038***	−0.040***
	[0.005]	[0.004]
High education	−0.049***	−0.045***
	[0.006]	[0.005]
Difference	0.010	0.005
p-Value	0.173	0.421

Note: Both the low education and high education rows use the same equation used in columns 3 and 8 of Table A.1. This model includes controls for lagged student test scores, parents' education level, ELL status, and the classroom average of these variables for other students in the classroom. The model also includes grade, year, and school fixed effects.The coefficient on the interaction between the binary variable for being after the release of the Times ratings and the Times rating are reported. Low-educated parents have a high school degree or less. High-educated parents have some college or more. Standard errors are clustered at the teacher level.
*** $p < 0.01$.

Table A.4
Effect of teacher ratings on performance by year.

2006	2007	2008	2009	2010	2011	2012
Panel A: Math						
0.248	0.252	0.276	0.291	0.306	0.245	0.224
[0.004]	[0.005]	[0.005]	[0.005]	[0.006]	[0.006]	[0.007]
Panel B: English						
0.165	0.161	0.170	0.192	0.193	0.148	0.122
[0.004]	[0.005]	[0.005]	[0.005]	[0.005]	[0.006]	[0.006]

Note: This table reports the point estimates and standard errors from Fig. 5.

Table A.5
Effect of teacher ratings by quintile for math.

Math						
2006	2007	2008	2009	2010	2011	2012
Panel A: Most effective quintile						
0.729	0.683	0.829	0.834	0.887	0.889	0.755
[0.014]	[0.014]	[0.014]	[0.014]	[0.015]	[0.016]	[0.016]
Panel B: More effective quintile						
0.465	0.440	0.546	0.539	0.548	0.604	0.507
[0.012]	[0.012]	[0.011]	[0.012]	[0.012]	[0.014]	[0.014]
Panel C: Average effectiveness quintile						
0.308	0.282	0.390	0.390	0.369	0.465	0.388
[0.012]	[0.011]	[0.012]	[0.012]	[0.012]	[0.014]	[0.015]
Panel D: Less effective quintile						
0.200	0.166	0.246	0.218	0.211	0.356	0.277
[0.011]	[0.011]	[0.012]	[0.012]	[0.012]	[0.014]	[0.014]
Panel E: Least Effective Quintile						
0.003	−0.048	0.022	−0.021	−0.019	0.159	0.097
[0.010]	[0.010]	[0.011]	[0.011]	[0.012]	[0.014]	[0.015]

Note: This table reports the point estimates and standard errors from panel (a) of Fig. 6.

Table A.6
Effect of teacher ratings by quintile for English.

English						
2006	2007	2008	2009	2010	2011	2012
Panel A: Most effective quintile						
0.497	0.429	0.485	0.563	0.541	0.511	0.464
[0.012]	[0.011]	[0.011]	[0.011]	[0.012]	[0.013]	[0.013]
Panel B: More effective quintile						
0.307	0.245	0.290	0.359	0.319	0.305	0.337
[0.009]	[0.009]	[0.009]	[0.010]	[0.010]	[0.010]	[0.011]
Panel C: Average effectiveness quintile						
0.221	0.158	0.195	0.255	0.216	0.233	0.255
[0.009]	[0.009]	[0.009]	[0.009]	[0.009]	[0.010]	[0.010]
Panel D: Less effective quintile						
0.156	0.088	0.115	0.154	0.117	0.162	0.195
[0.009]	[0.009]	[0.009]	[0.009]	[0.009]	[0.010]	[0.011]
Panel E: Least effective quintile						
0.017	−0.042	−0.017	0.002	−0.034	0.071	0.106
[0.009]	[0.009]	[0.009]	[0.010]	[0.010]	[0.011]	[0.011]

Note: This table reports the point estimates and standard errors from panel (b) of Fig. 6.

Table A.7
Effect of school ratings on performance by year.

2006	2007	2008	2009	2010	2011	2012
Panel A: Math						
0.036	0.017	0.028	0.033	0.027	0.004	0.016
[0.004]	[0.005]	[0.005]	[0.005]	[0.005]	[0.006]	[0.007]
Panel B: English						
0.033	0.011	0.013	0.042	0.022	0.021	0.054
[0.003]	[0.004]	[0.004]	[0.004]	[0.004]	[0.005]	[0.005]

Note: This table reports the point estimates and standard errors from Fig. 7.

Table A.8
Change in classroom composition.

2006	2007	2008	2009	2010	2011	2012
Panel A: Prior math CST scores						
0.042	0.054	0.061	0.078	0.090	0.100	0.080
[0.009]	[0.009]	[0.009]	[0.010]	[0.010]	[0.010]	[0.011]
Panel B: Prior English CST scores						
0.036	0.041	0.051	0.063	0.084	0.096	0.091
[0.011]	[0.010]	[0.010]	[0.011]	[0.011]	[0.011]	[0.013]
Panel C: Parents' education						
−0.016	−0.011	0.026	0.030	0.038	0.052	0.025
[0.020]	[0.016]	[0.017]	[0.017]	[0.017]	[0.017]	[0.019]
Panel D: Classroom size						
0.101	0.182	0.151	0.294	0.230	0.294	0.321
[0.075]	[0.066]	[0.069]	[0.077]	[0.077]	[0.082]	[0.087]

Note: This table reports the point estimates and standard errors from Fig. 8.

Table A.9
Effect of teacher ratings on teacher turnover.

2006	2007	2008	2009	2010	2011	2012
Panel A: Leave LAUSD						
−0.003	0.000	0.002	−0.003	−0.012	−0.011	−0.021
[0.003]	[0.003]	[0.003]	[0.003]	[0.005]	[0.005]	[0.005]
Panel B: Switch school						
−0.001	−0.001	−0.002	0.000	0.000	−0.003	−0.006
[0.003]	[0.002]	[0.002]	[0.002]	[0.002]	[0.003]	[0.003]
Panel C: Switch grade						
−0.009	−0.018	0.001	−0.006	−0.014	−0.003	−0.015
[0.006]	[0.006]	[0.005]	[0.005]	[0.006]	[0.006]	[0.007]

Note: This table reports the point estimates and standard errors from Fig. 9.

Table A.10
Persistence of teacher effects over time.

2006	2007	2008	2009	2010	2011
Panel A: Math					
0.193	0.198	0.207	0.232	0.189	0.200
[0.021]	[0.019]	[0.019]	[0.019]	[0.018]	[0.017]
Panel B: English					
0.250	0.256	0.319	0.297	0.292	0.244
[0.024]	[0.025]	[0.023]	[0.022]	[0.022]	[0.021]

Note: This table reports the point estimates and standard errors from Fig. 10.

References

Anderson, S., Rodin, J., 1989. Is bad news always bad?: cue and feedback effects on intrinsic motivation. J. Appl. Soc. Psychol. 19 (6), 449–467.

Banchero, S., 2012. Teacher ratings aired in New York. Wall Street J. February 25, 2012.

Barankay, I., 2014. Rank incentives: evidence from a randomized workplace experiment. Working Paper.

Bergman, P., Hill, M.J., 2018. The effects of making performance information public: regression discontinuity evidence from Los Angeles teachers. Econ. Educ. Rev. 66, 104–113. forthcoming.

Black, S.E., Lynch, L.M., 2001. How to compete: the impact of workplace practices and information technology on productivity. Rev. Econ. Stat. 83 (3), 434–445.

Buddin, R., 2010. How effective are Los Angeles elementary teachers and schools? MPRA Working Paper, pp. 27366.

Buddin, R., 2011. Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools. MPRA Working Paper, pp. 31963.

Chetty, R., Friedman, J.N., Rockoff, J.E., 2014a. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. Am. Econ. Rev. 104 (9), 2593–2632.

Chetty, R., Friedman, J.N., Rockoff, J.E., 2014b. Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. Am. Econ. Rev. 104 (9), 2633–2679.

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L., 2006. Teacher-student matching and the assessment of teacher effectiveness. J. Hum. Resour. 41 (4), 778–820.

Deci, E.L., Koestner, R., Ryan, R.M., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol. Bull. 125 (6), 627.

Dee, T.S., Wyckoff, J., 2015. Incentives, selection, and teacher performance: evidence from IMPACT. J. Policy Anal. Manage. 34 (2), 267–297.

DeNisi, A., Kluger, A.N., 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol. Bull. 119, 254–284.

Dixit, A., 2002. Incentives and organizations in the public sector: an interpretative review. J. Hum. Resour. 37 (4), 696–727.

Eisenberger, R., Fasolo, P., Davis-LaMastro, V., 1990. Perceived organizational support and employee diligence, commitment, and innovation. J. Appl. Psychol. 75 (1), 51–59.

Engellandt, A., Riphahn, R.T., 2011. Evidence on incentive effects of subjective performance evaluations. Ind. Labor Relat. Rev. 64 (2), 241–257.

Fleisher, L., 2012. Teacher rankings are slated for release. Wall Street J. February 24, 2012.

Fryer, R.G., 2013. Teacher incentives and student achievement: evidence from New York City public schools. J. Labor Econ. 31 (2), 373–407.

Goldhaber, D., Hansen, M., 2010. Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. Am. Econ. Rev. Pap. Proc. 100 (2), 250–255.

Goodman, S., Turner, L., 2013. The design of teacher incentive pay and educational outcomes: evidence from the New York City bonus program. J. Labor Econ. 31 (2), 409–420.

Gordon, R., Kane, T.J., Staiger, D.O., 2006. Identifying Effective Teachers Using Performance on the Job. The Brookings Institution, Washington, DC.

Hanushek, E.A., 1971. Teacher characteristics and gains in student achievement: estimation using micro data. Am. Econ. Rev. Pap. Proc. 61 (2), 280–288.

Hanushek, E.A., 2011. Valuing teachers: how much is a good teacher worth? Educ. Next 11 (3), 41–45.

Hanushek, E.A., Rivkin, S.G., 2010. Generalizations about using value-added measures of teacher quality. Am. Econ. Rev. Pap. Proc. 100 (2), 267–271.

Ichniowski, C., Shaw, K., Prennushi, G., 1997. The effects of human resource management practices on productivity: a study of steel finishing lines. Am. Econ. Rev. 87 (3), 291–313.

Imberman, S.A., Lovenheim, M.F., 2015. Incentive strength and teacher productivity: evidence from a group-based teacher incentive pay system. Rev. Econ. Stat. 97 (2), 364–386.

Imberman, S.A., Lovenheim, M.F., 2016. Does the market value value-added? Evidence from housing prices after a public release of school and teacher value-added. J. Urban Econ. 91, 104–121.

Jackson, C.K., 2013. Match quality, worker productivity, and worker mobility: direct evidence from teachers. Rev. Econ. Stat. 95 (4), 1096–1116.

Jackson, C.K., Bruegmann, E., 2009. Teaching students and teaching each other: the importance of peer learning for teachers. Am. Econ. Rev. Pap. Proc. 1 (4), 85–108.

Jacob, B.A., 2013. The effect of employment protection on worker effort: evidence from public schooling. J. Labor Econ. 31 (4), 727–761.

Jacob, B.A., Lefgren, L., 2007. What do parents value in education? An empirical investigation of parents' revealed preferences for teachers. Q. J. Econ. 122 (4), 1603–1637.

Jacob, B.A., Lefgren, L., Sims, D.P., 2010. The persistence of teacher-induced learning. J. Hum. Resour. 45 (4), 915–943.

Kuhnen, C.M., Tymula, A., 2012. Feedback, Self-Esteem, and Performance in Organizations. Manag. Sci. 58 (1), 94–113.

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B., Le, V.-N., Martinez, J.F., 2007. The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. J. Educ. Meas. 44 (1), 47–67.

Neal, D., 2011. The design of performance pay in education. In: Hanushek, E.A., Machin, S., Woessmann, L. (Eds.), Handbook of the Economics of Education Volume 4. Elsevier, Amsterdam: North-Holland, pp. 495–550.

Ost, B., 2014. How do teachers improve? The relative importance of specific and general human capital. Am. Econ. J. Appl. Econ. 6 (2), 127–151.

Papay, J.P., 2011. Different tests, different answers the stability of teacher value-added estimates across outcome measures. Am. Educ. Res. J. 48 (1), 163–193.

Pearce, J.L., Porter, L.W., 1986. Employee responses to formal performance appraisal feedback. J. Appl. Psychol. 71, 211–218.

Pop-Eleches, C., Urquiola, M., 2013. Going to a better school: effects and behavioral responses. Am. Econ. Rev. 103 (4), 1289–1324.

Rockoff, J.E., 2004. The impact of individual teachers on student achievement: evidence from panel data. Am. Econ. Rev. 94, 247–252.

Rockoff, J.E., Staiger, D.O., Kane, T.J., Taylor, E.S., 2012. Information and employee evaluation: evidence from a randomized intervention in public schools. Am. Econ. Rev. 102 (7), 3184–3213.

Rothstein, J., 2010. Teacher quality in educational production: tracking, decay, and student achievement. Q. J. Econ. 125 (1), 175–214.

Sartain, L., Steinberg, M.P., 2016. Teachers' labor market responses to performance evaluation reform: experimental evidence from Chicago public schools.. J. Hum. Res. 51 (3), 615–655.

Song, J., 2010. Teachers blast L.A. Times for releasing effectiveness rankings. Los Angeles Times August 30, 2010.

Springer, M.G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J.R., McCaffrey, D.F., Pepper, M., Stecher, B.M., 2010. Teacher Pay for Performance: Experimental Evidence From the Project on Incentives in Teaching. Nashville: National Center on Performance Incentives at Vanderbilt University.

Taylor, E.S., Tyler, J.H., 2012. The effect of evaluation on teacher performance. Am. Econ. Rev. 102 (7), 3628–3651.