NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

CALDER

AIR

The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality

JULIE BERRY CULLEN
CORY KOEDEL
ERIC PARSONS

WORKING PAPER 168-0717-2 • March 2019

The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality

Julie Berry Cullen
University of California, San Diego

Cory Koedel
University of Missouri
CALDER

Eric Parsons
University of Missouri

Contents

Acknowledgements

The authors gratefully acknowledge financial support from the Laura and John Arnold Foundation and the National Center for Analysis of Longitudinal Data in Education Research (CALDER) funded through grant #R305C120008 to American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education; research support from the Houston Education Research Consortium, in particular Shauna Dunn, Holly Heard, Dara Shifrer, and Ruth Turley; and research assistance from Li Tan. The authors also thank Tom Dee and seminar participants at the Center for Education Policy Analysis at Stanford University for helpful comments.

The views expressed here are those of the authors and should not be attributed to their institutions, data providers, or the funders. Any and all errors are attributable to the authors.

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality
Julie Berry Cullen, Cory Koedel & Eric Parsons
CALDER Working Paper No. 168-0717-2
March 2019

Abstract

We study how the introduction of a rigorous teacher evaluation system in a large urban school district affects the quality composition of teacher turnovers. With the implementation of the new system, we document increased turnover among the least effective teachers and decreased turnover among the most effective teachers, relative to teachers in the middle of the distribution. Our findings demonstrate that the alignment between personnel decisions and teacher effectiveness can be improved through targeted personnel policies. However, the change in the composition of exiters brought on by the policy we study is too small to meaningfully impact student achievement.

1. Introduction

Government agencies that provide services, such as education and health, operate in settings where it is difficult to observe both inputs and outputs. These are also sectors with ongoing concerns about efficiency and equity. In elementary and secondary education, efforts to improve schools have ranged from reallocation of resources via school finance reforms, to increased competition via school choice, to increased accountability via performance standards. The success of any of these depends on the quality and commitment of the workforce.

Recent research provides powerful evidence confirming that high-quality teachers are of great value to students.[1] A challenge facing school administrators in managing the teacher workforce is that effectiveness is not easy to measure and is not strongly correlated with observable characteristics. In this type of setting, improved information about quality along with pressure to use it can lead to more productive personnel policies. Given the two-sided nature of matches, there may also be equity implications because low-achieving schools struggle to attract and retain effective teachers (Bates, forthcoming; Clotfelter et al., 2006).

In this paper, we study the impact of introducing a rigorous teacher evaluation system on patterns of attrition by teacher effectiveness. The context of our study is the Houston Independent School District (HISD), the seventh largest school district in the United States. Starting in 2011, HISD phased in a new evaluation system centered on a standardized method for annually evaluating teachers, with the goal of using the comprehensive teacher performance measures to inform personnel decisions and skill development efforts. Recognizing that teacher hiring and development also play a role in overall efficacy, we focus here on how the policy impacted the level and distribution of teacher quality through more targeted retention.

Our empirical analyses rely on administrative data from the district tracking teachers for three years before the reform, two years during its phased implementation, and three years after full implementation. For the subset of teachers in tested grades and subjects, we begin by

[1] See, for example, Chetty, Friedman, and Rockoff (2014a/b), Hanushek (2011), Hanushek and Rivkin (2010), Jackson (2018), and Kraft (2019).

classifying them by quality using proxies we construct and validate for value added to student achievement. Then, using difference-in-differences style analyses, we show that the reform increased the likelihood of exit from the district of teachers in the bottom quintile of the quality distribution (by 6.2 percentage points) and decreased the likelihood of exit for teachers in the top quintile (by 4.0 percentage points), relative to teachers in the middle quintiles. The implication is that the reform improved the alignment between personnel outcomes and teacher quality.

As far as impacts on student achievement through the turnover channel, there are two issues to consider. First, overall turnover increased with the reform. Our design, which compares outcomes for more and less "treated" teachers within HISD to identify differential impacts, is not well suited to identify the causal impacts on overall turnover. Using a panel of statewide personnel data, we show that a portion of the level shift in turnover at HISD appears to be attributable to outside factors, but the majority is plausibly attributable to the policy. The disruption associated with the additional teacher turnover is likely to negatively impact student achievement (Adnot et al., 2017). Second, the impact of the reform on quality per turnover is substantively small. This is because many poorly-targeted exits remain despite the clear change in the relationship between teacher quality and exit. Though it is possible that policy efficacy may improve with time, we demonstrate that projecting forward for a decade based on the impacts we observe during the first three years of full policy implementation does not result in meaningful changes to workforce quality.

We contribute in several ways to the existing literature on policies designed to improve student achievement by strengthening the alignment between teacher effectiveness and personnel decisions. First, our findings stand in contrast to the body of simulation-based evidence suggesting such policies can meaningfully improve workforce quality through the channel of selective dismissal and retention (e.g., Hanushek, 2011; Staiger and Rockoff, 2010; Winters and Cowen, 2013). The large impacts in these studies arise in part from theoretical targeting to the tails of the distribution based directly on value-added and would apply to scenarios where teachers are evaluated exclusively on contributions to student achievement and policies are prescriptive. Our results show that turnover outcomes are likely to be more weakly tied to effectiveness in practical applications for a variety of reasons, most notably slippage between performance measures and underlying effectiveness in real-world evaluation systems and personnel policies that are more discretionary.[2]

Second, while policies implemented on the ground may not realize the full potential gains, we corroborate several recent papers showing impacts of innovations to teacher evaluation on differential retention. For example, Rockoff et al. (2012) find that providing principals in New York City with teacher value-added estimates increases the likelihood of exit for low-performing teachers, and Sartain and Steinberg (2016) find similar effects of a Chicago pilot program that evaluated teachers more rigorously via classroom observations. However, both find that the changes in the selectivity of teacher flows would not have detectable impacts on student achievement, which in the Chicago case is partly because only nontenured teachers were exited. While these two pilot interventions presumably operate primarily by influencing principal decision-making through the provision of new information, two other studies show evidence for teacher discouragement. One of these, Dee and Wyckoff (2015), shows that being subject to the threat of future dismissal following the receipt of the next-to-lowest rating under the DC IMPACT program induces voluntary exit. The other, Loeb, Miller, and Wyckoff (2015), finds that teachers denied tenure under a New York City reform that tightened standards are more likely to leave during their extended probationary periods.

Third, there has been little research of district teacher evaluation initiatives similar in scale and scope to the HISD effort, with the exceptions of the aforementioned Dee and Wyckoff (2015) study and Stecher et al. (2018). Dee and Wyckoff (2015) find promising evidence of teacher exit and effort responses at key discontinuity thresholds in the prescriptive DC IMPACT

[2] In any evaluation of program efficacy centered on student achievement, personnel decisions that incorporate information from other types of measures will appear to be mistargeted. A rationale for including multiple measures in teacher evaluations is that teacher impacts on non-test outcomes are not highly correlated with impacts on test scores (Blazar, 2018; Jackson, 2018). However, non-test-based measures used in teacher evaluations are not strongly related to teacher effectiveness as measured by non-test student outcomes (Kraft, 2019).

program, but do not attempt a holistic evaluation to assess the effect on workforce quality. The recent report from Stecher et al. (2018) most closely resembles the analysis we undertake here. These authors' study the Intensive Partnerships for Effective Teaching initiative (IPET), funded by the Bill & Melinda Gates Foundation and implemented in seven sites across the United States. Although implementation varied across sites, it is broadly similar to the HISD policy in terms of substance and scale. While exit rates of ineffective teachers increased over time as evaluation results were incorporated in tenure and contract renewal decisions, retention rates of effective teachers did not. Like ours, their null findings regarding improvements in student outcomes offer less optimism than previous research about the ability of these types of human-resource policies to meaningfully improve student achievement at scale.

The remainder of the paper proceeds as follows. In the next section we provide more background on the HISD reform and the broader policy context. Section 3 introduces the data and sample, as well as our measure of teacher quality. Sections 4 and 5 present our empirical strategy and main results, while section 6 explores heterogeneity in impacts across initially lower- and higher-achieving schools. Section 7 discusses the implications of our estimates for student achievement. Finally, section 8 offers a brief conclusion.

2. Policy background

HISD has implemented several policies designed to raise staff quality and effort since the mid-2000s. First, a merit pay program (ASPIRE) was introduced in 2006-07 to reward teachers and administrators for raising student achievement. Then, four years later, the district began phased development and implementation of the Effective Teachers Initiative (ETI). This comprehensive reform is designed to improve teacher quality through more effective recruitment at the front end, individualized professional development in the middle, and targeted retention and exit on the back end. The emphasis on tying personnel decisions more closely to quality was made explicit in differential retention goals for the least and most effective teachers, with a particular focus on improving teacher quality for high-need students. In this section, we provide an overview of the features of the two policies that are most relevant to our analysis, with much more extensive details provided in Appendix A.

The cornerstone of the ETI reform is the implementation of a rigorous teacher evaluation system intended to provide more informative reviews of teacher performance. Ratings under the old system were high and did not meaningfully differentiate teachers (e.g., see Weisberg et al., 2009). The new evaluation system was designed by the district during the 2010-11 school year with input from stakeholders and formally approved by the school board in spring 2011. The new appraisals involve three components: instructional practice, professional expectations, and student performance. Scores on the first two components are based on observations and reviews conducted inside and outside the classroom. For instructional practice, a teacher's skills are evaluated using well-defined rubrics that cover setting student expectations, lesson planning, and classroom management.[3] For the professional expectations component, teachers are evaluated relative to a set of objective measures of compliance with policies, interactions with colleagues, and participation in professional development. The student performance scores are based on estimates of a teacher's impact on student learning. Teachers are scored on a scale from 1 to 4 on each component, and these are then combined to deliver summary ratings of ineffective, needs improvement, effective, or highly effective.

The initial step in transitioning to the new system was ensuring that all teachers were formally assigned ratings in 2010-11. Prior to that year, ratings for almost one in three teachers were not recorded with the district. The 2010-11 ratings were based on at least two classroom observations that, though scored under the old system, were conducted in an environment where differentiating teachers by effectiveness was a leading district concern. This was also the first year that schools' retention rates of highly effective teachers and exit rates of ineffective teachers were measured and publicly reported.[4] In the following year, 2011-12, teachers were scored under the two new observational components: instructional practice and professional

[3] The HISD rubrics are internally developed and detailed in booklets titled "HISD Teacher Appraisal and Development: Instructional Practice and Professional Expectation Rubrics."
[4] The annual Facts and Figures brief released by HISD to the public now periodically includes these among the set of key indicators of progress, tracking trends since 2010-11.

expectations. Due to delays in approving student performance metrics for teachers in untested subjects and grades, these metrics were not formally incorporated into the ratings until 2012-13.

We focus our analysis on teachers in tested grades and subjects for whom we can measure quality in a consistent way in the pre- and post-policy periods. Constructing consistent quality measures is feasible for these teachers because the SAS Institute provided proprietary teacher-level value-added estimates to HISD as key inputs to the merit pay system. These were observable to principals prior to ETI, but the policy push to better align personnel decisions with quality was absent. The most relevant changes for our teachers with the implementation of the reform are the districtwide emphasis on tying personnel decisions more closely to quality and the addition of the improved observational components.

Given the phase-in of the policy, we treat 2010-11 and 2011-12 as years in which the policy was partially implemented, and 2012-13 onward as the period of full implementation. While the policy could have affected turnover for all teachers as early as following the 2010-11 school year, initial impacts were more likely for our teachers since information about efficacy in promoting student learning was already available at the onset of the increased emphasis on quality-driven personnel decisions.[5] In that year, though, there was no bite yet on formal ratings, as 97 percent of teachers were rated effective or better based on the old rubric.[6] The incorporation of the improved observational components in 2011-12 reduced this share noticeably to 88 percent. Then, this share fell more dramatically to 64 percent when the student performance component was added in 2012-13 under full implementation. This latter drop can be attributed to the fact that the student performance measures are relative, so that some teachers are necessarily deemed ineffective, while the other criteria are absolute and not as discriminating.

In our empirical analysis, we examine how teacher turnover evolved in ways that are

[5] Others have shown effects of policies in years prior to formal implementation in settings with structurally similar rollouts. For example, Butcher, McEwan, and Weerapana (2014) show that academic departments at Wellesley began responding to an anti-grade inflation policy during a transition year though it was made clear the policy would not be implemented until the following year.

[6] Appendix A (Table A1) shows the fraction receiving each ratings level, overall and by component.

related to quality over the partial- and full-implementation periods. Though turnover is only one channel through which the new human capital policies could affect the workforce, it is of interest because it is arguably the lever that principals can affect most, particularly in the short run. That said, any impacts on turnover reflect both demand- and supply-side responses and thus capture voluntary as well as involuntary exits.[7] Further, these responses are to the initiative as a whole, including supporting interventions bundled with the new evaluation system. For example, the district worked to streamline its recruitment procedures, such as by identifying vacancies earlier in the year, and to expand recruitment efforts, such as by increasing the number of visits to universities. The district also provided new opportunities for teacher development and established new leadership roles for effective teachers. These types of complementary changes are likely inherent to the introduction of any rigorous appraisal system.

Something more unique to the Houston context is that ETI was introduced against the backdrop of a merit pay system. Under ASPIRE, teachers in core subjects can receive bonuses for student learning gains exhibited in their classrooms and smaller bonuses for campus-wide performance.[8] Prior to ETI, nearly all teachers in tested subjects and grades received bonuses, with the average bonus on the order of $3,600 (or about 7 percent of average base salary). During the partial-implementation period, first teachers with poor attendance or low student growth were made ineligible for the campus awards in 2010-11, and then standards for both types of awards were made more stringent in 2011-12. Whereas it had been sufficient to be in the top half on at least one teacher-subject or campus measure, qualifying for an award now required being closer to the top 20 percent. In 2012-13, once ETI was fully phased in, teachers identified as ineffective or needing improvement by the appraisal system were also disqualified from receiving campus awards. Teachers experienced these changes with a one-year lag, since awards

[7] Unfortunately, we do not have access to information on reasons for exits and, even if we did, the official reasons recorded might be misleading, such as for teachers who are counseled out.

[8] The evidence on how teachers respond to ASPIRE performance incentives is mixed. Imberman and Lovenheim (2015) find that high school teachers increase effort in response to team incentives, but Brehm, Imberman, and Lovenheim (2017) do not find strategic effort responses to individual incentives among teachers in lower grades.

are announced and paid out starting in November of the following school year. Importantly, awards are paid regardless of whether the individual is still an employee.

The net effect of the post-ETI changes to ASPIRE is that the most effective teachers maintained similar levels of average awards, while average amounts fell for less effective teachers.[9] The first two post-ETI years, during the policy phase-in, can be viewed as providing insights about effects in a less discriminating merit pay regime, prior to when the changes to ASPIRE led to substantive changes in average payouts received. Beginning with full implementation of ETI in 2012-13, we capture effects that embed the increasing alignment of merit pay awards. Like with any other district policy related to teacher evaluation, such realignment would be expected in response to changes to the evaluation system, although this means that our findings may be overstated relative to implementation in an environment without complementary merit pay adjustments.

3. Data and summary statistics

We have access to detailed school, teacher, and student administrative data from HISD for school years 2007-08 through 2015-16. These data allow us to measure teacher turnover through 2014-15, leaving us with an eight-year panel that we divide into three time periods: 2007-08 through 2009-10 (three pre-policy years), 2010-11 through 2011-12 (two partial-implementation years), and 2012-13 through 2014-15 (three full-implementation years).

3.1 Measuring school disadvantage and selecting analysis schools

For our sample of schools, we begin with the 201 traditional public schools that were operational during our sample period and serve students in grades 3 to 8, which are the grade levels for which we are able to construct measures of teacher quality consistently over the course of our panel. As a summary measure of each school's context we use the achievement level, which is defined as the average of students' math and reading scores on statewide exams, standardized within grade and year, and taken over the pre-policy years. We divide schools into

[9] Appendix A details changes to the ASPIRE award standards from 2007-08 through 2014-15 (Table A3) and shows how these changes impact award payouts (Figure A4).

three groups based on pre-policy achievement levels: low (bottom quintile), middle (quintiles 2-4), and high (top quintile).

After classifying schools by achievement, we exclude a set of schools due to a concurrent intervention conducted in HISD as described by Fryer (2014). Fryer (2014) led an intervention starting in 2010-11 that introduced a bundle of best practices from effective charter schools in 15 traditional elementary and middle schools. The onset of the intervention included changes to teaching and leadership personnel. To avoid contamination, we drop the schools where Fryer intervened from our analytic sample.[10] Consistent with his description, all but one of these schools are in the bottom quintile of achievement. We assign schools to quintiles prior to dropping the Fryer schools so that our school groupings are unconditional. This allows for straightforward interpretation, with the practical consequence that our sample size of bottom-quintile schools is reduced.

Table 1 shows summary statistics for the schools included in our analysis, broken down by achievement group. The top panel shows differences in the characteristics of students served. Beyond the construct-driven differences in achievement, low-achieving schools serve a disproportionate share of black students and students with English as a second language, while high-achieving schools serve markedly fewer economically-disadvantaged students.

3.2 Measuring teacher quality and selecting analysis teachers

Critical to our analysis is the ability to measure teacher effectiveness in a comparable way over the full sample period. While teacher experience and education levels are candidate measures – and we do estimate supplementary models based on teacher experience below – the literature shows that these characteristics explain little of the variation in student learning (Aaronson, Barrow, and Sander, 2007; Hanushek and Rivkin, 2006; Harris and Sass, 2011). We also have scores from the observation-based components of the official evaluation system from 2011-12 onward, but not only are these unavailable in prior years, they are difficult to compare

[10] In Appendix E (Table E2) we show that our main findings are similar if we include these schools.

across campuses with differentially challenging environments and map more weakly to student learning (Kane et al., 2011, 2013; Steinberg and Garrett, 2016; Whitehurst, Chingos, and Lindquist, 2014).

For these reasons, we construct quality measures derived from value-added estimates available for teachers in tested grades and subjects for many years at HISD. These teacher-specific EVAAS scores are single-year measures of student test score growth produced using a propriety method developed by the SAS Institute. They are available to us back to the 2006-07 school year. Although the technical estimation details differ from standard value-added models, conceptually they are similar (Ballou, Sanders, and Wright, 2004). We restrict our analysis to teachers in grades 3-8 who are assigned math or reading EVAAS scores.

We construct measures of teacher effectiveness in math and reading by combining multiple years of teachers' scores per the following regression based on Chetty, Friedman, and Rockoff (2014a):

$$V_{ikt} = \delta_0 + \mathbf{V}_{\mathbf{ikt-}}\boldsymbol{\delta}_{\mathbf{1}} + \eta_{ikt} \tag{1}$$

In equation 1, V_{ikt} is teacher *i*'s EVAAS score in subject *k* and year *t*, $\mathbf{V}_{\mathbf{ikt-}}$ is a vector of teacher *i*'s EVAAS scores in the same subject in years prior to year *t*, and η_{ikt} is an idiosyncratic error term. The EVAAS scores are normalized by subject and year. The fitted values from the regression, $\hat{V}_{ikt} = \hat{\delta}_0 + \mathbf{V}_{\mathbf{ikt-}}\hat{\boldsymbol{\delta}}_{\mathbf{1}}$, are jackknifed quality measures where a value of one, for example, implies that the teacher is one standard deviation above average in the true distribution for teachers in the district.[11] Because not all teachers have a complete panel of prior scores to be used in the estimation of equation 1, separate regressions are estimated for all possible combinations as in Chetty, Friedman, and Rockoff (2014a). We do require, though, that the teacher have a time *t* EVAAS score to be included in the sample, which ensures the individual is

[11] The jackknifed quality measures are not renormalized to have a standard deviation of one, and in fact have a standard deviation less than one. Theoretically, a one-unit change in the jackknifed measures corresponds to a one standard deviation change in the distribution of teacher quality (see, e.g., Chetty, Friedman, and Rockoff, 2014b).

teaching the relevant subject contemporaneously.

We use scores only from years prior to t as explanatory variables in order to guard against survivor bias in our sample with measured teacher quality. An implication of including only prior scores is that first-year teachers who are newly observed in tested subjects and grades in the district are necessarily excluded from these calculations. We incorporate these teachers into our analysis by grouping them in a separate "unknown quality" category. Another issue with the backward-looking jackknife is that it relies on more observations for teachers in later years of our panel. An initial concern is that a relative reduction in noise over time could confound our analysis. Not only have we empirically confirmed that our results are robust to restricting the backward-looking windows to be comparable across years, but the implicit shrinkage in these estimates is also a theoretical argument against this concern (Jacob and Lefgren, 2008).[12] The jackknife approach also allows teacher effectiveness to drift over time, consistent with the slow-moving process documented by Chetty, Friedman, and Rockoff (2014a).

We re-norm the distribution of teacher quality each year. The re-norming is necessary because there was a state-mandated change to the testing regime during our panel (in spring 2012) that led to a downward shift in measured achievement at HISD.[13] As a consequence of the re-norming, our measure of teacher quality is comparable only in relative and not absolute terms across years. If the distribution were to shift rightward due to the policy, those classified as ineffective in later years would be more effective than their counterparts in the pre-policy period. Though HISD policies are in fact anchored around relative quality (see Appendix A), this might mute observed impacts on turnover by relative quality as compared to absolute quality. Despite this possibility, our qualitative findings regarding differential attrition for more and less effective teachers are not sensitive to adjustments in the classification of teachers that attempt to account

[12] In Appendix E (Table E3) we show our findings are qualitatively robust to estimating teacher quality using a single-lagged (i.e., common window) jackknife rather than using all available lagged scores.
[13] Backes et al. (2018) show that test changes do not substantively alter estimates of relative teacher quality.

for policy-induced shifts in the quality distribution through the turnover channel.[14]

Recent studies show that conceptually similar jackknifed measures based on value-added are forecast-unbiased estimates of teacher quality in other contexts (Bacher-Hicks, Kane, and Staiger, 2014; Chetty, Friedman, and Rockoff, 2014a). Adopting their methods, we test whether our measures have the same property by examining whether changes in teacher quality at the school-by-grade level caused by staffing changes accurately predict changes in student test scores, as would be expected if the measures are unbiased. For example, if a teacher with high measured effectiveness moves to a new school and/or different grade, test scores for students in the new school-by-grade combination should increase in the year after the change. With the caveat that our tests are less powerful than in previous studies that exploit larger datasets, our findings are consistent with the jackknifed quality measures being forecast unbiased predictors of future student achievement (see Appendix B for more details on the procedure and results). The reading-based estimates are somewhat noisier, which is consistent with previous research (Lefgren and Sims, 2012), but in both subjects the analysis indicates that our measures provide useful information about teacher effectiveness.

In the main analytic work that follows, we estimate models that pool math and reading teachers. Of all HISD teachers in our schools, 32.4 percent are charged with math or reading instruction in a tested grade.[15] Of these, roughly one in five has a current EVAAS score but no available prior scores to calculate our jackknifed quality measure. As noted above, we group these first-time teachers together into an "unknown quality" category. In order to divide teachers with jackknifed quality measures into quality bins we first assign them to one of three groups –

[14] To elaborate briefly, below we estimate models that map teacher exit rates to teacher quality measured relative to the distribution in each year. The sensitivity of our findings along this dimension can be explored by altering the distributional cut-points from year-to-year that we use to bin teachers, based on our annual estimates of how attrition patterns have changed (along with a normality assumption). For example, we can use the 19th percentile to identify "low quality" teachers in time $t+1$ rather than the 20th percentile to allow for the whole distribution to have moved. The shifts in the quality distribution implied by our annual selective attrition estimates are small enough as to make such adjustments of little consequence.

[15] The share of elementary and middle school teachers responsible for math or reading instruction in tested grades at HISD is on par with other locales, though depressed somewhat by the over-representation of bilingual and English as a second language (ESL) teachers. Among non-bilingual and non-ESL teachers in our schools, 39.0 percent are responsible for math or reading instruction in a tested grade.

math only, reading only, or both – and then use teachers' group-specific percentile rankings to assign a quality bin in each year. The percentile ranks for the latter group are based on the simple average of the jackknifed math and reading estimates. The group-based rankings facilitate comparisons across teachers whose responsibilities differ.

The middle panel of Table 1 shows how teacher quality is distributed across schools grouped by achievement level in the pre-policy period. In the first row, our measures of effectiveness are reported in standard deviations of the teacher distribution (again noting that for teachers of both subjects we use the simple average value). The table shows that teacher quality is not evenly distributed across the district. On average, less effective teachers are found at low-achieving schools. Teachers in the bottom quintile of the quality distribution are about 1.5 times more prevalent at low- relative to high-achieving schools, while teachers in the top quintile are half as prevalent. Also, teachers new to tested grades and subjects (or new to the district) are more often found in low-achieving schools. Later we show that these teachers, for whom jackknifed quality is unknown, are relatively ineffective. The table also shows that low-achieving schools employ teachers with less experience and more education. While these characteristics are not generally strongly related to outcome-based teacher quality, those with low levels of experience (i.e., 0-5 years) fall disproportionately in our bottom-quintile and unknown-quality groups.

It is important to highlight that our jackknifed quality measures are not directly available to school principals. Instead, principals have access to the inputs (i.e., the year-by-year EVAAS scores), non-value-added measures of student progress, and post-policy observational assessments, in addition to other indicators of quality we do not observe. Our measures, though, are systematically related to the summary ratings teachers receive. For example, during the full-implementation period when all components of the assessment were formally scored, our measure of quality explains 20 percent of the variation in summary ratings in our sample. The share of teachers rated ineffective or needs improvement increases steadily from 6 percent, to 22 percent, to 42 percent moving from the top, to the middle, and then to the bottom teacher quality group based on our measures (see Appendix A). Though single-year EVAAS scores are a direct input to the student performance component of the new ratings system, so predictive of final ratings through that channel, our jackknifed quality measures are more predictive of the non-test-based evaluation components.[16] Thus, in addition to being informative for how decisions under the new system are likely to affect student achievement, our measures are better aligned than single-year EVAAS scores with the other formal evaluation criteria in the system.

3.3 Measuring teacher turnover

We measure school exits and decompose school exits into exits from the district and transfers to other schools within the district. We define teacher turnover by looking forward in the data one year. A benefit of using a single-year measure instead of a multi-year measure is that we can calculate turnover for more years. The limitation is that single-year exit measures overstate exit rates because teachers – particularly young teachers – move in and out of the workforce (Grissom and Reininger, 2012). We therefore test robustness to using alternative two-year definitions for campus and district exit, where a teacher is classified as having exited only if she is also not present in year t+2.

The bottom panel of Table 1 shows single-year turnover rates for our grade 3 through 8 reading and math teachers in the pre-policy period. Pre-policy turnover is 16.7 percent at low-achieving schools, versus 14.0 and 12.7 percent at middle- and high-achieving schools.[17] The differences are driven primarily by lower rates of within-district transfer from higher-achieving schools. Unsurprisingly, the use of two-year exit rates (not shown) results in marginally lower turnover by approximately 0.4 percentage points, or 3 percent, but does not noticeably change the comparison of turnovers by school type.

4. Empirical strategy

We estimate effects on turnover using difference-in-differences (DD) models of the

[16] For example, our measures explain 10-20 percent more of the across- and within-school variance in 2012-13 teachers' instructional practice scores than single-year EVAAS scores.
[17] These pre-policy turnover rates are similar to those provided for grade 4 and 5 teachers in New York City by Ronfeldt, Loeb, and Wyckoff (2013).

following form, specified as linear probability models:

$$Y_{ist} = \delta_0 + Partial_t\delta_1 + Full_t\delta_2 + \mathbf{Q_{it}\delta_3} + (\mathbf{Q_{it}} \times Partial_t)\boldsymbol{\delta}_\mathbf{4} + (\mathbf{Q_{it}} \times Full_t)\boldsymbol{\delta}_\mathbf{5} + \mathbf{X_{st}}\boldsymbol{\beta} + \phi_s + \varepsilon_{ist} \quad (2)$$

In equation 2, Y_{ist} is a binary variable indicating whether teacher *i* at school *s* exits the school (or exits the district or transfers to another school) at the conclusion of year *t*. $Partial_t$ and $Full_t$ are indicators for the partial- and full-implementation periods, respectively. $\mathbf{Q_{it}}$ is a vector of three indicator variables for whether teacher *i* in year *t* is in the top quintile of the quality distribution, the bottom quintile, or if quality is unknown (i.e., if no prior quality measure is available for jackknifing). The omitted comparison group includes teachers in the middle three quintiles of the quality distribution.[18] Our use of the jackknifed measures to place teachers into the bins defined by $\mathbf{Q_{it}}$ guards against bias from contemporaneous circumstances that correlate with performance and exit.[19] The **X**-vector contains teacher characteristics that might have independent effects on turnover, such as race, gender, experience and education, as shown in Table 1. Finally, ϕ_s is a school fixed effect to allow for fixed school attributes that affect teacher attrition rates, and ε_{ist} is an error term. Throughout we report standard errors clustered at the school level.[20]

The objective is to identify shifts in the relationship between teacher quality and exit over time, embodied by $\boldsymbol{\delta}_\mathbf{4}$ and $\boldsymbol{\delta}_\mathbf{5}$. We also report estimates of δ_1 and δ_2 to give a sense of how the overall exit rate changes over time. These changes may be attributable to the reform, but it is difficult to rule out other time-varying factors when estimating these simple differences. Therefore, we primarily emphasize estimates of $\boldsymbol{\delta}_\mathbf{4}$ and $\boldsymbol{\delta}_\mathbf{5}$, which contain the coefficients on

[18] We show results separately for math and reading teachers in Appendix E (Tables E4 and E5) using the same quality quintile bins and also using alternative linear measures of quality (that necessarily exclude teachers of unknown quality).
[19] The implicit shrinkage in the jackknife procedure implies that our estimates should not be affected by attenuation bias from using a generated regressor (Jacob and Lefgren, 2008).
[20] Our panel includes repeat observations of teachers, so we have also estimated the models allowing for teacher clustering. We choose to report standard errors that allow for clustering at the school level both because of the conceptual appeal, since principals are the key implementers, and to be conservative, since the standard errors are slightly larger.

the interactions between the partial- and full-implementation period indicators and teacher quality. These DD parameters measure the change in the composition of exiters due to push and pull factors associated with the reform. While our model is nonstandard in the sense that there is no group of untreated individuals, the research design is functionally unaffected – that is, the design allows us to ascertain whether turnover rates diverge for low- and high-quality teachers relative to teachers in the middle.

Interpreting differential changes in turnover as causally attributable to the reform requires that differences in turnover rates across teacher quality groups would have otherwise remained stable. One way to provide evidence on whether this parallel-trends assumption holds is to explore whether these differences were stable across years prior to the reform. To this end, after presenting out baseline DD results, we show results from time-disaggregated models that include a full set of year effects in place of the indicators for the partial- and full-implementation periods. In addition to being informative about the validity of our research design by revealing whether there are pre-policy trends in differential exit rates, these models also shed light on the dynamics of any responses after policy implementation.

Even if there is no evidence of pre-policy divergence, it is still possible that turnover rates would have diverged after the reform for other reasons. A potential confounder in our context is the Great Recession and subsequent economic recovery. For example, more effective teachers might respond differently to secular changes in outside options. The existing literature does not offer much guidance for predicting the effect of the recession on turnover by quality. The only paper we are aware of that considers the role of the economy on teacher transitions by quality studies effects on selection at entry, finding that teachers who enter during a recession are more effective on average (Nagler, Piopiunik, and West, forthcoming), but it is difficult to extrapolate this finding to the exit decision. Unfortunately we do not have access to student-teacher linked data elsewhere in Texas to construct a counterfactual to HISD using outcome-based measures of teacher quality. However, we do have access to statewide personnel files that we use to construct analogous experienced-based models.[21] Although experience is a weaker measure of quality, it has the advantage of facilitating a triple-differences research design comparing HISD to other Texas districts. Moreover, in Appendix C we show that inexperience is a useful proxy for low effectiveness as teachers with five or fewer years disproportionately fall in the bottom-quintile and unknown-quality groups within HISD (see Appendix Table C1).

5. Effects of the reform on turnover

5.1 Descriptive analysis

We begin by visually documenting districtwide trends in exit and turnover rates in Figure 1. The figure shows trends for the three different mobility measures: school exit, district exit, and school transfer. The former is the sum of the two latter measures. School years in the figure, and in all figures and tables to follow, are identified by the spring year – e.g., the 2010-11 school year is labeled as 2011.

The figure shows that after modest declines through 2010, turnover by all three measures began to rise at the conclusion of the 2011 school year. Of total school exits, roughly half of the observed increase is due to an increase in district exits, and half is due to an increase in within-district school transfers. It is difficult to determine how much of the increase in overall turnover is attributable to the policy change as our data panel spans the Great Recession. Unemployment peaked in 2009 and gradually declined over the next several years. In Appendix C, using the statewide data, we show a U-shaped pattern in turnover during this period in Texas, with turnover returning to around initial levels in the later years statewide. Turnover at HISD is far above the initial level by the end of our data panel, suggesting the policy played at least some role in the post-reform increase.

Figure 2 provides similar information to Figure 1 but divides teachers into groups based on our quality measures. It is visually apparent that the school exit rate increased more quickly in the post-policy period for the least effective teachers relative to other teachers, driven primarily

[21] While these files are not linked to student achievement, they do contain information on teachers' school assignments and experience.

by district exits, with teachers of unknown quality not far behind. Although instances of school transfers are higher in the post-policy years overall, no systematic change in the relationship between teacher quality and school switching is apparent in Figure 2. This suggests that HISD has avoided a "dance of the lemons" problem, at least internal to the district, in the sense that the rise in school exits among ineffective teachers is not accompanied by a commensurate rise in these teachers shuffling between schools.[22]

5.2 Main difference-in-differences estimation results

We estimate the difference-in-differences style model described in Section 4 to assess the significance and robustness of the patterns illustrated in Figures 1 and 2. Table 2 shows results from estimating the full specification shown in equation 2 for each turnover outcome. This is our preferred model but in unreported results our estimates are very similar if we use sparser variants of the model that exclude teacher characteristics and even school fixed effects. All control variables except for the partial- and full-implementation indicators are mean-centered in the regressions, including the school indicator variables, so that the intercept and its interactions (in the first three rows) can be interpreted as exit rates at the mean values of the covariates.[23]

The general patterns from Figures 1 and 2 are reflected in the estimates in Table 2 and confirmed to be statistically significant. First, the table makes clear that turnover increased overall from the pre- to post-policy periods in terms of both district exits and school transfers, as indicated by the top three rows of the table. Turning to exits by quality, there is evidence that less effective teachers became increasingly more likely to exit HISD as the policy was phased in and implemented, as well as that more effective teachers became less likely to exit. For example, column 2 shows that a teacher in the bottom quintile of the quality distribution during the post-

[22] It would also be of interest to know whether ineffective teachers who exit HISD end up teaching in other Texas schools. Unfortunately, our unique teacher identifiers are HISD-specific and do not permit us to track teachers who move out. Using the separate statewide panel dataset of Texas teachers, we find that the post-policy period is marked by higher entry rates of former HISD teachers into other Texas districts, though it is unclear where these teachers fall in the quality distribution.

[23] The mean-centering does not affect model fit or the coefficients on the key parameters interacting time with teacher quality. It is used only to improve interpretability of the results with regard to the overall exit rate (Dalal and Zickar, 2012).

policy period was an additional 6.2 percentage points more likely to exit the district than a bottom-quintile teacher in the pre-policy period, relative to a teacher in the middle quintiles. The partial- and full-implementation periods exhibit a similar change in the relative likelihood of exit for bottom-quintile teachers compared to the pre-policy period.

Top-quintile teachers became less likely to exit, although the pattern is not as strong. This finding is directionally consistent with the general goals of ETI; moreover, the fact that the change in the exit pattern is weaker fits with the emphasis of the district on using ETI to identify and remove ineffective teachers in particular. There is little indication that ETI affected within-district school transfer rates differentially for teachers who differ by quality. Similarly, there is no systematic change in exit or mobility rates of teachers without prior teaching experience in the focal tested grades and subjects. On the whole, the findings in Table 2 point to an increased alignment between teacher quality and exit as ETI was phased in and implemented.[24]

Table 3 reports on the robustness of these findings to two adjustments to the analysis. First, in the left panel, we consider the sensitivity of our results to using a 2-year exit measure for school and district exits. That is, rather than coding exits based on looking forward just one year in the data, we look forward two years to determine whether the exiting teacher remained either out of the school or out of the district. When we make this definitional change, we are no longer able to examine outcomes for the 2015 teacher cohort. Thus, we report results from models covering the 2008-2014 cohorts using the one-year and two-year exit definitions, which are otherwise comparable to the results we show in Table 2. Although exit rates are slightly lower overall using the two-year definition, the patterns in our estimates are similar.

In the right panel of Table 3, we explore whether changes in school leadership are

[24] Though we do not emphasize the pre-policy relationships between exit and quality, Table 2 indicates that less effective teachers were less likely to transfer to a new school within the district and suggestively more likely to exit the district prior to ETI; these patterns are reversed (and suggestive only) for more effective teachers. Studies of teacher mobility in Florida (Feng and Sass, 2017; West and Chingos, 2009) and North Carolina (Goldhaber, Gross, and Player, 2010) find less effective teachers are more likely to exit the school for any reason. Within HISD we find less effective teachers are no more or less likely to stay in the same school, since the higher rate of district exit is offset by a lower rate of school transfer.

important mediators of the effects of the reform. We return to our full dataset and single-year exit measures and replicate the analysis in Table 2, but replace the school fixed effects in the model with school-by-principal fixed effects.[25] This allows us to isolate policy impacts holding the principal fixed. The results are similar to what we show in Table 2, suggesting that the reform alters teacher turnover similarly whether leadership turns over or not.

In Table 4, by way of comparison, we estimate models that match those shown in Table 2 except we use single-year EVAAS scores to assign teachers to quality bins rather than the jackknifed measures. Using single-year EVAAS scores makes the "unknown" quality category unnecessary, as all teachers in our sample have a contemporaneous measure. We view these bins to be inferior proxies for relative teacher quality, since the scores are noisy and also potentially embed unobserved shocks affecting both performance and exit. Yet, these scores are direct inputs to the ratings and are likely salient to principals making contract renewal decisions. The results reveal little changes in the targeting of turnovers in the partial-implementation period, but increased district exit for bottom-quintile teachers (of about half the magnitude as in Table 2) and reduced district exit for top-quintile teachers (of about the same magnitude as in Table 2) in the full-implementation period. Thus, we detect earlier and greater changes in the targeting of turnovers by quality according to our jackknifed teacher quality bins than bins based on current EVAAS scores.

5.3 Validity of the difference-in-differences approach

In this section, we explore the extent to which the maintained assumptions underlying our difference-in-differences approach appear to hold, as previewed in Section 4. We begin by continuing to focus on HISD in isolation and estimating event-time models that disaggregate the data into individual years. Table 5 shows these results. The partial-implementation years are shown in italics to make it easier to distinguish the three policy periods, and the intercept-by-year coefficients are suppressed (since these provide no new insights beyond what is shown in the

[25] Principal turnover in HISD is in line with patterns statewide as reported by Branch, Hanushek, and Rivkin (2012).

first three rows of Table 2). For bottom-quintile and unknown-quality teachers there is no indication of pre-policy trends for any of the types of turnover prior to policy implementation. For top-quintile teachers, while there is no pre-trend in overall school exits, this is because a dip in district exit is offset by a rise in school transfers. The detectable trend toward a reduced district exit rate leading into 2011 suggests some caution in interpreting the findings for this group of teachers, though there is a still a clear break in the full implementation period. For school transfers, transfers for top-quintile teachers were unusually low in 2008 (the base year), then rise and flatten out from 2009 onward. We interpret this pattern as showing that transfer rates are relatively flat for these teachers over the course of our panel.

The disaggregated models also show in more detail how differential exit rates by quality evolved over time after the onset of the policy. Although many of the year-to-year comparisons are only suggestive due to the reduced statistical power, two patterns stand out. First, the district exit rate for low-performing teachers jumped up in 2011, remained high through 2013, and fell thereafter. This may reflect an initial district push to exit particularly ineffective teachers, peaking in 2013 during the first year of full implementation and then softening. The other pattern apparent in the data is that the exit rate for highly effective teachers shifts noticeably between the partial- and full-implementation periods. An explanation consistent with Dee and Wyckoff (2015) is that high-performing teachers value ETI and its salience increased over time as it became clear it would not be repealed. Another potential explanation is the increasing alignment between the policy and the merit pay program, as discussed previously and documented in Appendix A.

To bolster confidence that the post-reform patterns we uncover are not attributable to confounding effects of the economic recovery, we next bring in data from the rest of the state to provide a plausible counterfactual for HISD. Recall that our best available proxy for quality in the statewide data is teacher experience and, unfortunately, experience is only systematically related to quality at the bottom end. So, having few years of experience is a reasonable proxy for being of low quality but having many years of experience does not convey much about effectiveness relative to teachers of middling experience.

Before presenting results from triple-differences models, the first two columns of Table 6 show how the policy affected patterns of school and district exits by experience within HISD. The specifications in these columns replicate our baseline models from Table 2, replacing the indicators for teacher quality groups with indicators for experience-level groups. Focusing on the results in column 2, district exit for the least experienced group is elevated in the partial- and full-implementation periods, relative to the pre-policy period and those in the middle experience group. We find opposing effects on relative turnover for highly experienced teachers in the partial (positive) and full (negative) implementation periods, underscoring that experience is likely correlated with outcomes through channels other than underlying effectiveness, including the availability of opportunities outside of teaching and salience to administrators. With this caveat to interpretation in mind, the findings for the least experienced group are consistent with the reform differentially increasing district exit for low quality teachers.

In the third and fourth columns of Table 6, we expand the sample and model to include all schools in Texas.[26] The triple-differences parameters for HISD are reported in the bottom rows of the table. They show that the pattern of exits by experience changed significantly at HISD relative to other Texas districts coinciding with the onset of ETI. Specifically, inexperienced teachers at HISD began exiting at a higher rate relative to more experienced teachers, compared to other districts. While this is a weak test in the sense that experience-based shifts in exits are not a necessary condition for the policy to have an impact, and experience-based changes were certainly not explicitly targeted by the district, Table 6 provides evidence that the human capital policies at HISD were doing something more than the average "recession effect." It is also reassuring that the DDD estimates are very similar to the DD estimates for exit-by-experience, though the estimates for total exits in HISD in the partial- and full-implementation periods are approximately halved. This latter result motivates our consideration

[26] Appendix C provides additional results for school transfers and state exits and using alternative subsets of Texas districts to construct the counterfactual for HISD.

in the discussion section of scenarios that attribute only a portion of the overall increase in turnover to the reform.

6. Heterogeneity by school achievement level

Next we consider the potential for effect heterogeneity across schools in order to shed light on the distributional impact of the policy. We divide schools into three groups based on their pre-policy location in the distribution of average achievement in math and reading: bottom quintile (low-achieving), middle quintiles (quintiles 2-3-4), and top quintile (high-achieving). Improving teacher quality at low-achieving schools serving high-need students is a priority under ETI, and principals at these schools might also benefit more from the information provided by the new system. However, they may also have less capacity to respond because demand for effective teachers likely increased system-wide, opening up the possibility for the best teachers to trade-up in terms of school environment and making retention tougher at the bottom.

Graphical evidence on heterogeneity in the policy impact is presented in Figures 3 and 4. Figure 3 replicates the information shown in Figure 1, but separately by school type. The figure shows that exit rates increased across all three school types in the post-policy period. A notable pattern, however, is the extent to which the intensity of policy implementation varies by school-achievement level. Specifically, there is the most action at the lowest-achieving schools, followed by middle-achieving schools, and relatively little happened at high-achieving schools. This suggests that any pressure from the reform was felt less intensely at higher-achieving schools.[27]

Figure 4 further divides teachers by quality group within the same school categories. Reading across a row in Figure 4 holds the school-achievement group fixed, and reading down a column holds the teacher quality-group fixed. We omit teachers of unknown quality for presentational convenience in the figure, focusing instead on the contrast between teachers who

[27] Appendix C shows that while the U-shaped pattern in exits across years is more pronounced for lower-achieving schools in other Texas districts as well, although the levels return only to initial levels and do not exceed these, unlike in HISD.

we can identify as being in the top, middle-three, and bottom quintiles of the quality distribution.

Although the graphs in the figure cut the data thinly, and therefore noise is an issue, they suggest some interesting patterns. For instance, the first row of graphs shows school-exit, district-exit, and school-transfer patterns at low-achieving schools, by teacher type. Low- and middle-performing teachers at these schools became much more likely to exit, and when they did exit, they mostly exited the district. In contrast, while high-performing teachers in these schools are also much more likely to exit their schools over time, they are increasingly transferring to other schools, particularly in the last year of the data panel. This suggests that among other things, ETI has increased mobility opportunities for these high-performing teachers, which is consistent with related evidence from Bates (forthcoming).[28]

We formally test the statistical significance of the patterns suggested by Figure 4 by adding interactions between the time and quality variables in equation 2 with indicators for schools that are in the bottom and top quintiles. The fully interacted model incorporates many parameters and is therefore somewhat difficult to digest and interpret, but the key takeaways discussed above are statistically distinguishable. We relegate the output from the model to Appendix E (Table E6) for brevity.

7. Discussion and interpretation

Given that we measure teacher quality in terms of effectiveness and validate the predictive power of our measures over student achievement, it is reasonable to expect gains in student learning to align with the change in the quality composition of exiters we document. However, whether or not gains are realized also depends on any impacts on the quality of teacher entrants (Rothstein, 2015).[29] Inference is further clouded by the high turnover rate among

[28] In the absence of compensating wage differentials, teachers will prefer positions that are more desirable along non-pecuniary dimensions (Greenberg and McCall, 1974). Teachers may prefer higher-achieving schools for a variety of reasons. Survey evidence suggests they prefer to work with higher-SES students, perhaps because this requires less effort, but correlated non-pecuniary benefits such as better administrative support and shorter commutes are more important in pushing teachers toward higher-achieving schools (Horng, 2009).

[29] In Appendix E (Table E1), we do not find evidence of systematic shifts in the relative quality of entrants in post-reform years, though there is no control group for this analysis and the results are imprecise. Stecher et al. (2018) also do not find any evidence of improvement despite enhanced recruitment efforts under IPET.

teachers in our sample post-ETI. Since turnover increases as much as it does, to the extent that this turnover is attributable to the policy and adversely affects achievement, it could imply net losses for students.

In order to directly gauge how the turnover aspects of the policy affect student achievement, we estimate models of changes in school-by-grade achievement of the following form:

$$\Delta\bar{A}_{sgt} = \beta_0 + \mathbf{\Delta\bar{X}_{sgt}\beta_1} + \Delta TO_{sgt}\beta_2 + [BQ_{sgt}\beta_3 + MQ_{sgt}\beta_4 + TQ_{sgt}\beta_5 + UK_{sgt}\beta_6] + \gamma_s + \tau_t + u_{sgt} \quad (3)$$

In equation 3, $\Delta\bar{A}_{sgt}$ is the difference in average test scores across student cohorts in school *s* and grade *g* from period *t-1* to *t*.[30] The focal explanatory variables are the change in the level of turnover among math and reading teachers to which the two cohorts were exposed, ΔTO_{sgt}, and the share of teachers exiting between cohorts by quality group, where teachers are weighted by the number of students taught. The quality groups are denoted by BQ_{sgt}, MQ_{sgt}, TQ_{sgt}, and UK_{sgt} for bottom quintile, middle quintiles, top quintile, and unknown teacher quality, respectively. The vector $\mathbf{\Delta\bar{X}_{sgt}}$ captures changes in student demographic characteristics across cohorts, while γ_s and τ_t are school and year fixed effects, respectively. This model is similar to the model estimated by Adnot et al. (2017) except we control not only for the levels of teacher turnover by quality across cohorts, which capture changes in the composition of teachers, but also for changes in exposure to turnover, which capture differential disruption.

The results are shown in Table 7. We estimate separate models for math and reading, along with a stacked model that pools both subjects. Across all three columns, the estimates match up well with our jackknifed quality measures, as expected. For example, the estimated coefficients on the exit shares of bottom, middle, and high quality teachers from the pooled model in Table 7 are 0.097, 0.042, and -0.146, respectively, and the analogous average

[30] Each observation is weighted by time *t* school-by-grade student enrollment, and standard errors are clustered at the school-by-grade level.

jackknifed measures for teachers in these groups are -0.085, 0.013, and 0.129 (converted to student standard deviation units). Table 7 also shows the aforementioned result that teachers new to tested grades and subjects are relatively ineffective, performing similarly to bottom-quintile teachers. And, like others, we find a disruption effect of turnover (Hanushek, Rivkin, and Schiman, 2016; Ronfeldt, Loeb, and Wyckoff, 2013). In the stacked model, we estimate the disruption effect to be 0.055 student standard deviations for a school-by-grade cell that experiences 100 percent turnover.

Next we perform back-of-the-envelope calculations to assess the impact of the reform on student achievement, drawing on the results in Table 7 in addition to results from the preceding sections. In order to carry out these calculations, we first need to repackage the estimates from Table 2 to determine the implied shares of leavers falling into each teacher quality group before and after the reform. For this, we focus on the change from the pre-policy to full-implementation periods. We first estimate the (school or district) exit rate of teachers belonging to quality group j in period m , where m indicates either the pre-policy or full-implementation period. The exit rate, ER_{jm} , is estimated from a combination of the intercept and interaction parameters shown in Table 2. For example, the district exit rate for bottom-quintile teachers in the full-implementation period is set equal to the model intercept (0.084) plus the post-policy interaction coefficient (0.108) plus the interaction coefficient that captures the differential exit rate for these teachers in the post-policy period (0.062).[31] We then convert these rates to be in terms of shares of the total teaching workforce: $ET_{jm} = (PS_j * ER_{jm})$, where PS_j is the share of teachers belonging to group j in the pre-policy period. The share of leavers falling in group j in period m is this value divided by the sum across groups (i.e., $ET_{jm}/\sum_{i=1}^{4} ET_{im}$).

Figure 5 shows the resulting estimated shares of school and district exits for each teacher quality group in the pre-policy and full-implementation periods. The directions of the shifts over

[31] We treat insignificant point estimates in Table 2 as zeros in our calculations, though results are similar if we take the insignificant coefficients at face value.

time follow from Table 2, but the figure makes clear that the implied change to the quality composition of exiters is modest. Specifically, the share of low-performing teachers among school exiters rises from just 14.1 to 17.5 percent, and among district exiters from 14.1 to 18.4 percent; while school and district exit shares for high-performing teachers decline by just 2.0 and 3.2 percentage points, respectively. Though the estimation of equation 3 reveals that teachers of unknown quality are less effective on average, their exit shares are hardly affected by the policy.

Multiplying the compositional changes from Figure 5 by the group-specific differences in teacher effectiveness from Table 7 gives an estimate of the effect of the policy on teacher quality per turnover. Focusing on district exits, this calculation indicates that the average effectiveness of leavers falls by just 0.008 student standard deviations due to the shift in the quality-composition of exits from HISD.

An important caveat to this calculation is that the district exit rate increased substantially overall from the pre- to post-policy periods. Our estimate of the effect of the policy on the level of exits is based on a single difference in our models, and thus is not well-identified, and the triple-differences estimates in Table 6 suggest it is somewhat inflated. The full rise in exits is implicitly attributed to the policy in Figure 5 since we use the pre- and post-policy intercept parameters directly from Table 2 in the calculations. This is important because a higher overall exit rate dulls the impact of differential exits by quality on the composition of exiters (i.e., for any set gaps in exit rates by teacher quality, higher overall attrition implies weaker targeting).

We consider how modifications to our calculations along this and other dimensions affect total policy impacts in Table 8, focusing on district exits. The first column in the table shows per-turnover quality effects under the alternatives indicated by the row headings. The estimate in the first row is for the baseline case reported in the text above. In row 2, we assume that the Great Recession is responsible for half of the pre-post differences we estimate in Table 2 in terms of overall exit and differential exit by quality (though there is no evidence that impacts on differential exits are overstated). Unsurprisingly, the per-turnover impact of the reform is even smaller than in the baseline case (0.006). In rows 3 and 4, we leave the differential exit rates by quality unchanged (i.e., as in the baseline condition in row 1) and change the effect of the policy on total turnover only. Row 3 lowers the parameterized effect of the policy on the level of turnover to half the size of the post-policy intercept coefficient in Table 2. In row 4 we further reduce the assumed effect of the policy on total turnover to 25 percent of the observed increase. Both of these scenarios improve the targeting of the policy, making it more impactful on a per-turnover basis, although the magnitudes are still small (0.011 in row 3 and 0.014 in row 4).

In row 5 we return to the baseline condition in row 1 but allow for a greater degree of targeting of the policy by assuming that the changes in relative exit rates we observe in the top and bottom quintiles of the quality distribution are entirely concentrated in the top and bottom *deciles*. While sample size issues make it hard to test for decile-level differences in turnover directly in our main models, it is straightforward to modify the calculations here *post hoc* in this way. If the changes in turnover we document above are concentrated deeper in the tails of the quality distribution, the results show a small uptick (to 0.010) in the per-turnover effect of the policy on teacher quality.

The second column of Table 8 presents back-of-the-envelope estimates of cumulative policy effects over 10 years on the *average quality of the workforce* through the turnover channel for these and additional alternative scenarios. Average workforce quality is more difficult to change because, in any given year, most teachers do not move. Details on the parameterizations of the iterative simulations that produce the estimates in column 2 are provided in Appendix D.

The 10-year effects shown in rows 1-5 reflect the policy parameters described above and implicitly assume replacement teachers are drawn from the same quality distribution as incumbents while the policy iterates. The remaining rows build in alternative assumptions about replacement teachers, which do not affect the estimates in column 1 (because these estimates describe exits only). Motivated by Rothstein (2015), row 6 starts with the baseline scenario (row 1) but shifts down the distribution of replacement teacher quality by 0.25 standard deviations of the incumbent distribution (which corresponds to a much smaller move in the distribution of student achievement). The result is a net negative effect of the policy cumulated over 10 years, which reflects the small effect of better-targeted outflows (as shown in row 1) coupled with lower replacement quality. Rows 7-9 incorporate negative first-year disruption effects of replacements, as estimated by the stacked model in Table 7. Their incorporation into Table 8 is somewhat awkward because the estimates end up reflecting a mix of stocks and flows. Specifically, the estimates show cumulative effects on the stock of workforce quality, conceptualized as persistent value-added, plus the temporary disruption effects of the most recent year of turnover. The results show that incorporating just a single year of disruptive turnover noticeably reduces the already small 10-year policy effects.

Overall, while the analysis in the preceding sections shows that the HISD policies have improved alignment between our validated measures of teacher quality and teacher exit, in this section we show that exits thus far have not been targeted well enough to induce meaningful achievement gains. This is true in terms of affecting the average quality of exiters, and in terms of affecting the quality of the workforce overall after 10 years. Regarding the latter calculations, our long-term estimates give effect sizes on total workforce quality measured by value-added on the order of about 0.01 standard deviations of student achievement. The range of policy parameters and conditions we consider is fairly broad. Given the consistency of the estimates, our interpretation is that under reasonable parameterizations and extrapolations, the estimates in Table 8 effectively rule out large policy impacts.

8. Conclusion

We study the effect of a new, more rigorous teacher evaluation system in the Houston Independent School District on the quality composition of teacher turnovers. The new system has increased the exit rate among low-performing teachers and reduced the exit rate among high-performing teachers, relative to teachers in the middle of the distribution. Policy activity is disproportionately concentrated at low-achieving schools within the district.

Our analysis shows that the implementation of a more rigorous evaluation system can lead to personnel decisions better aligned with teacher quality, but also highlights the challenge associated with improving workforce quality via selective attrition. In short, in the system we study there are simply too many poorly targeted exits in the post-policy period (by middle- and top-performing teachers) for the net policy effect on achievement to be meaningful. It may be that the efficacy of the program will improve as individuals within HISD gain experience with the system (Ahn and Vigdor, 2014), but based on our estimates of the effects of the policy during the first three years of full implementation, the scope for workforce improvement via selective attrition seems limited. Our results in this regard are smaller than what one might expect based on simulation studies that examine the potential for improved personnel policies to raise workforce quality (Hanushek, 2011; Staiger and Rockoff, 2010; Winters and Cowen, 2013), but in line with other recent on-the-ground evidence from Stecher et al. (2018).

Stepping back from our narrow policy context, a possible complementary intervention to help stem the tide of higher-quality teacher exits would be to offer more competitive wages that better reflect differences in teacher quality, as argued in Rothstein (2015). Increased pay for exceptional performance is a key feature of the IMPACT program in Washington DC (Dee and Wyckoff, 2015). Although HISD has attempted to better align pay with productivity, which may have contributed to the strengthening of some of our findings in the later years of our data panel (e.g., the increase in the retention rate of the most effective teachers), the merit pay program has faced challenges and taken only partial steps in that direction (Brehm, Imberman, and Lovenheim, 2017; Shifrer, Turley, and Heard, 2013). Finally, we also acknowledge there are other ways that the new system is designed to improve instruction and student outcomes about which our study is silent, notably via recruitment and greater improvement among incumbent teachers.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1), 95-135.

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2017. Teacher Turnover, Teacher Quality, and Student Achievement in DCPS. *Educational Evaluation and Policy Analysis* 39(1), 54-76.

Ahn, Thomas, and Jacob L. Vigdor. 2014. When Incentives Matter Too Much: Explaining Significant Responses to Irrelevant Information. NBER Working Paper No. 20321.

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. 2014. Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles. NBER Working Paper No. 20657.

Backes, Ben, James Cowan, Dan Goldhaber, Cory Koedel, Luke Miller, and Zeyu Xu (2018). The Common Core Conundrum: To What Extent Should We Worry that Changes to Assessments and Standards Will Affect Test-Based Measures of Teacher Performance. *Economics of Education Review* 62, 48-65.

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

Bates, Michael. Forthcoming. Public and Private Learning in the Market for Teachers: Evidence from the Adoption of Value-Added Measures. *Journal of Labor Economics*.

Blazar, David. 2018. Validating Teacher Effects on Students' Attitudes and Behaviors: Evidence from Random Assignment of Teachers to Students. *Education Finance and Policy* 13(3), 281-309.

Branch, Gregory F., Eric A. Hanushek, and Steven G. Rivkin. 2012. Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals. NBER Working Paper No. 17803.

Brehm, Margaret, Scott A. Imberman, and Michael F. Lovenhiem. 2017. Achievement Effects of Individual Performance Incentives in a Teacher Merit Pay Tournament. *Labour Economics* 44, 133-150.

Butcher, Kristin, Patrick J. McEwan, and Akila Weerapana. 2014. The Effects of an Anti-Grade-Inflation Policy at Wellesley College. *Journal of Economic Perspectives* 28(3), 189-204.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2017. Measuring the impacts of teachers: Reply. *American Economic Review* 107(6), 1685-1717.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9), 2593-2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* 104(9), 2633-79.

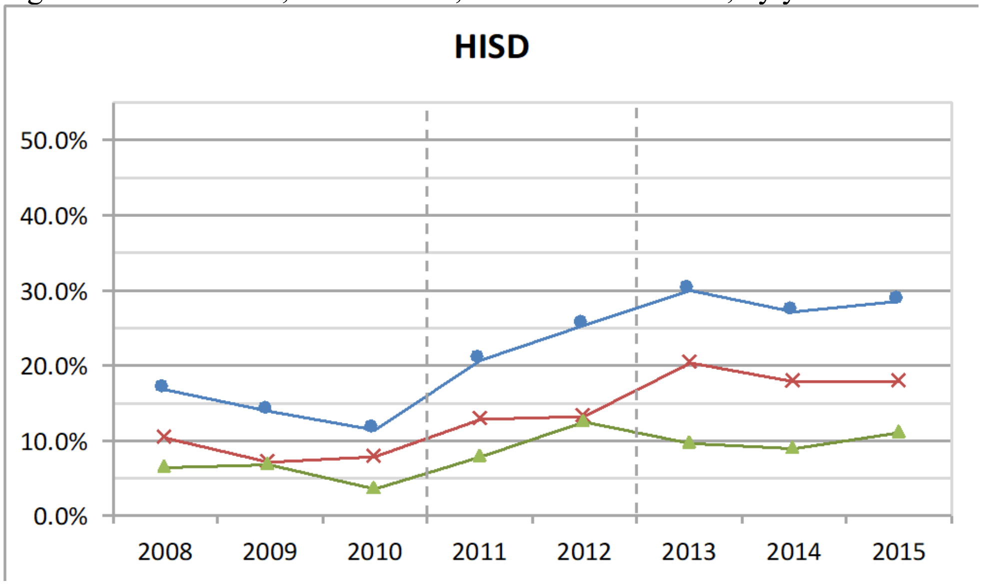
Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor, and Justin Wheeler. 2006. High-poverty Schools and the Distribution of Teachers and Principals. *North Carolina Law Review* 85: 1345-1379.

Dalal, Dev K., and Michael J. Zickar. 2012. Some Common Myths about Centering Predictor Variables in Moderated Multiple Regression and Polynomial Regression. *Organizational Research and Methods* 15(3), 339-362.

Dee, Thomas, and James Wyckoff. 2015. Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management* 34(2), 267-297.
Feng, Li, and Tim Sass. 2017. Teacher Quality and Teacher Mobility. *Education Finance and Policy* 12(3), 396-418.
Fryer, Roland G. 2014. Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *Quarterly Journal of Economics* 129(3), 1355-1407.
Goldhaber, Dan, Bethany Gross, and Daniel Player. 2010. Teacher Career Paths, Teacher Quality, and Persistence in the Classroom: Are Public Schools Keeping Their Best? *Journal of Policy Analysis and Management* 30(1), 57-87.
Greenberg, David, and John McCall. 1974. Teacher Mobility and Allocation. *Journal of Human Resources* 9(4), 480-502.
Grissom, Jason A., and Michelle Reininger. 2012. Who Comes Back? A Longitudinal Analysis of the Re-Entry Behavior of Exiting Teachers. *Education Finance and Policy* 7(4): 425-454.
Hanushek, Eric A. 2011. The Economic Value of Higher Teacher Quality. *Economics of Education Review* 30(3), 266-479.
Hanushek, Eric A., and Steven G. Rivkin. 2006. "Teacher Quality." In *Handbook of the Economics of Education* Vol. 1, ed. Eric A. Hanushek and Finis Welch, 1051-78. Amsterdam: North-Holland.
Hanushek, Eric A., and Steven G. Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review* 100(2), 267-271.
Hanushek, Eric A., Steven G. Rivkin, and Jeffrey C. Schiman. 2016. Dynamic Effects of Teacher Turnover on the Quality of Instruction. *Economics of Education Review* 55, 132-148.
Harris, Douglas N., and Tim R. Sass. 2011. Teacher Training, Teacher Quality and Student Achievement. *Journal of Public Economics* 95(7-8), 798-812.
Horng, Eileen Lai. 2009. Teacher Tradeoffs: Disentangling Teachers' Preferences for Working Conditions and Student Demographics. *American Educational Research Journal* 46(3), 690-717.
Imberman, Scott A., and Michael F. Lovenheim. 2015. Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System. *Review of Economics and Statistics* 97(2), 364-386.
Jackson, Kirabo. 2018. What do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy* 126(5), 2072-2107.
Jacob, Brian, and Lars Lefgren. 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics* 26(1), 101-136.
Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Seattle, WA: Bill and Melinda Gates Foundation.
Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources* 46(3), 587-613.
Kraft, Matthew. 2019. Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources* 54(1), 1-36.
Lefgren, Lars, and David P. Sims. 2012. Using Subject Test Scores Efficiently to Predict Teacher Value-Added. *Educational Evaluation and Policy Analysis* 34(1), 109-121.

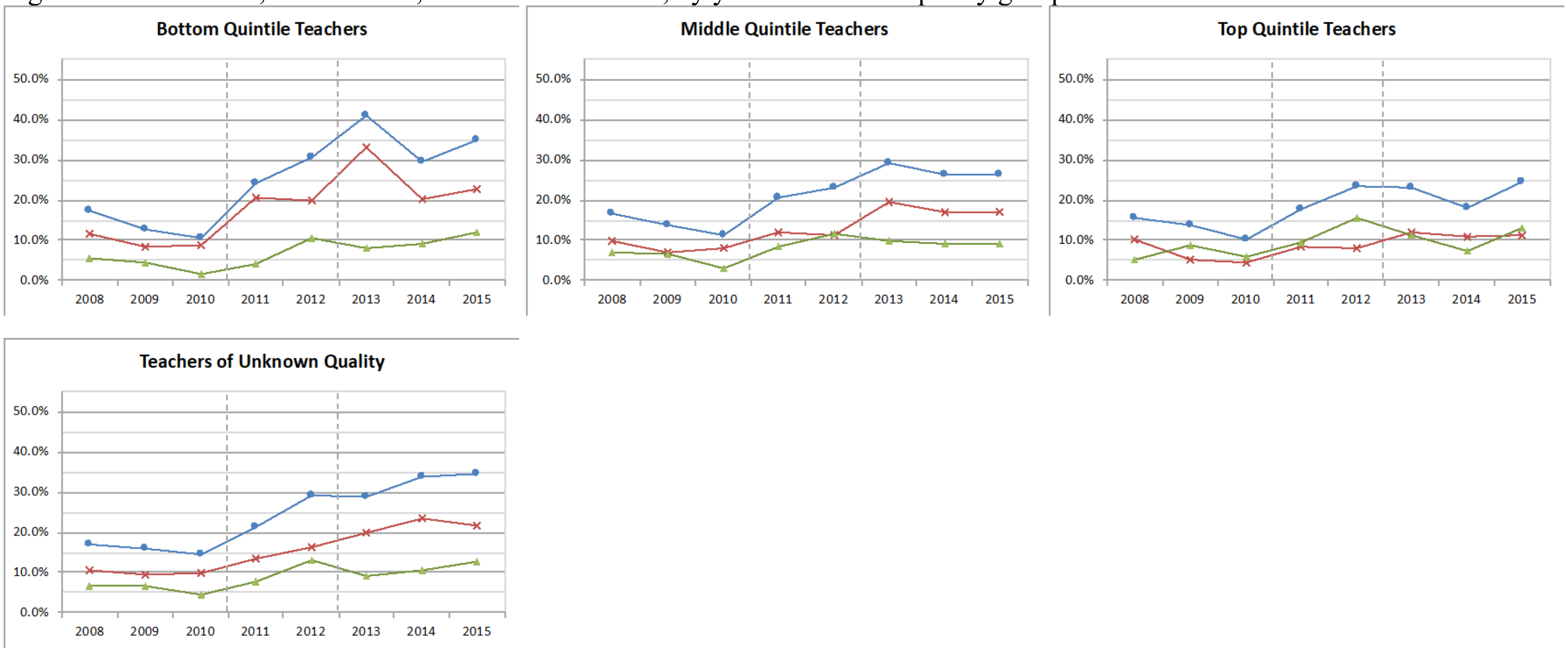
Loeb, Susanna, Luke. C. Miller, and James Wyckoff. 2015. Performance Screens for School Improvement: The Case of Teacher Tenure Reform in New York City. *Educational Researcher* 44(4), 199-212.
Nagler, Markus, Marc Piopiunik, and Martin R. West. Forthcoming. Weak Markets, Strong Teachers: Recession at Career Start and Teacher Effectiveness. *Journal of Labor Economics*.
Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2012. Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *American Economic Review* 102(7), 3184-3213.
Ronfeldt, Matthew, Susanna Loeb, and James Wyckoff. 2013. How Teacher Turnover Harms Student Achievement. *American Educational Research Journal* 50(1), 4-36.
Rothstein, Jesse. 2017. Measuring the impacts of teachers: Comment. *American Economic Review* 107(6), 1656-84.
Rothstein, Jesse. 2015. Teacher Quality Policy When Supply Matters. *American Economic Review* 105(1), 100-130.
Sartain, Lauren, and Matthew P. Steinberg. 2016. Teachers' Labor Market Responses to Performance Evaluation Reform: Experimental Evidence from Chicago Public Schools. *Journal of Human Resources* 51(3), 615-55.
Shifrer, Dara, Ruth Lopez Turley, and Holly Heard. 2013. Houston Independent School District's ASPIRE Program: Estimated Effects of Receiving Financial Awards. Policy Report. Houston Educational Research Consortium.
Staiger, Douglas O., and Jonah E. Rockoff. 2010. Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives* 24(3), 97-118.
Stecher, Brian M., Holtzman, Deborah J., Garet, Michael S., Hamilton, Laura S., Engberg, John, Steiner, Elizabeth D., Robyn, Abby, Baird, Matthew D., Gutierrez, Italo A., Peet, Evan D., Brodziak de los Reyes, Iliana, Fronberg, Kaitlin, Weinberger, Gabriel, Hunger, Gerald Paul, and Chambers, Jay. 2018. *Improving Teaching Effectiveness Final Report: The Intensive Partnerships for Effective Teaching Through 2015-16*. Santa Monica, CA: RAND Corporation.
Steinberg, Matthew P., and Rachel Garret. 2016. Classroom Composition and Measured Teacher Performance: What do Teacher Observation Scores Really Measure? *Educational Evaluation and Policy Analysis* 38(2), 293-317.
Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. New York, NY: The New Teacher Project.
West, Martin R., and Mathew W. Chingos. 2009. Teacher Effectiveness, Mobility and Attrition in Florida. In Mathew G. Spring, ed., *Performance Incentives: Their Growing Impact on American K-12 Education*, Brookings Institution Press, 251-71.
Whitehurst, Grover J., Matthew M. Chingos, and Katharine M. Lindquist. 2014. Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts. Washington, DC: Brown Center on Education Policy at Brookings.
Winters, Marcus A., and Joshua M. Cowen. 2013. Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies. *Journal of Policy Analysis and Management* 32(3), 634-654.

Figure 1. School exits, district exits, and school transfers, by year



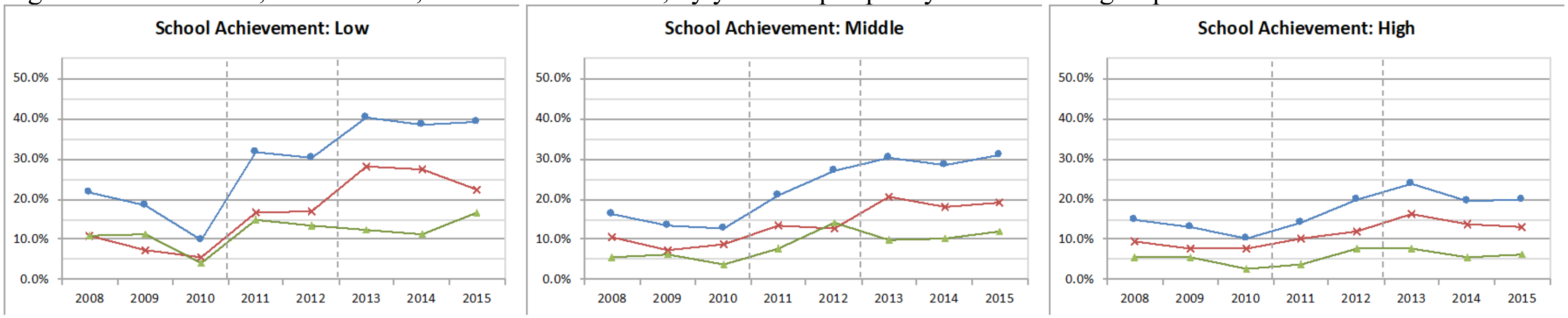
Notes: Circle markers: school exits; Cross markers: district exits; Triangle markers: school transfers. School exits are the sum of district exits and school transfers, and all three are based on transitions from year t to year t+1. The dashed vertical lines identify the partial-implementation years (2011 and 2012) and separate the pre-policy period from the full-implementation period. In this and all following figures the year refers to the fiscal year, which is the same as the spring of the school year. For example, 2008 refers to fiscal year 2008 and represents school year 2007-08.

Figure 2. School exits, district exits, and school transfers, by year and teacher quality group



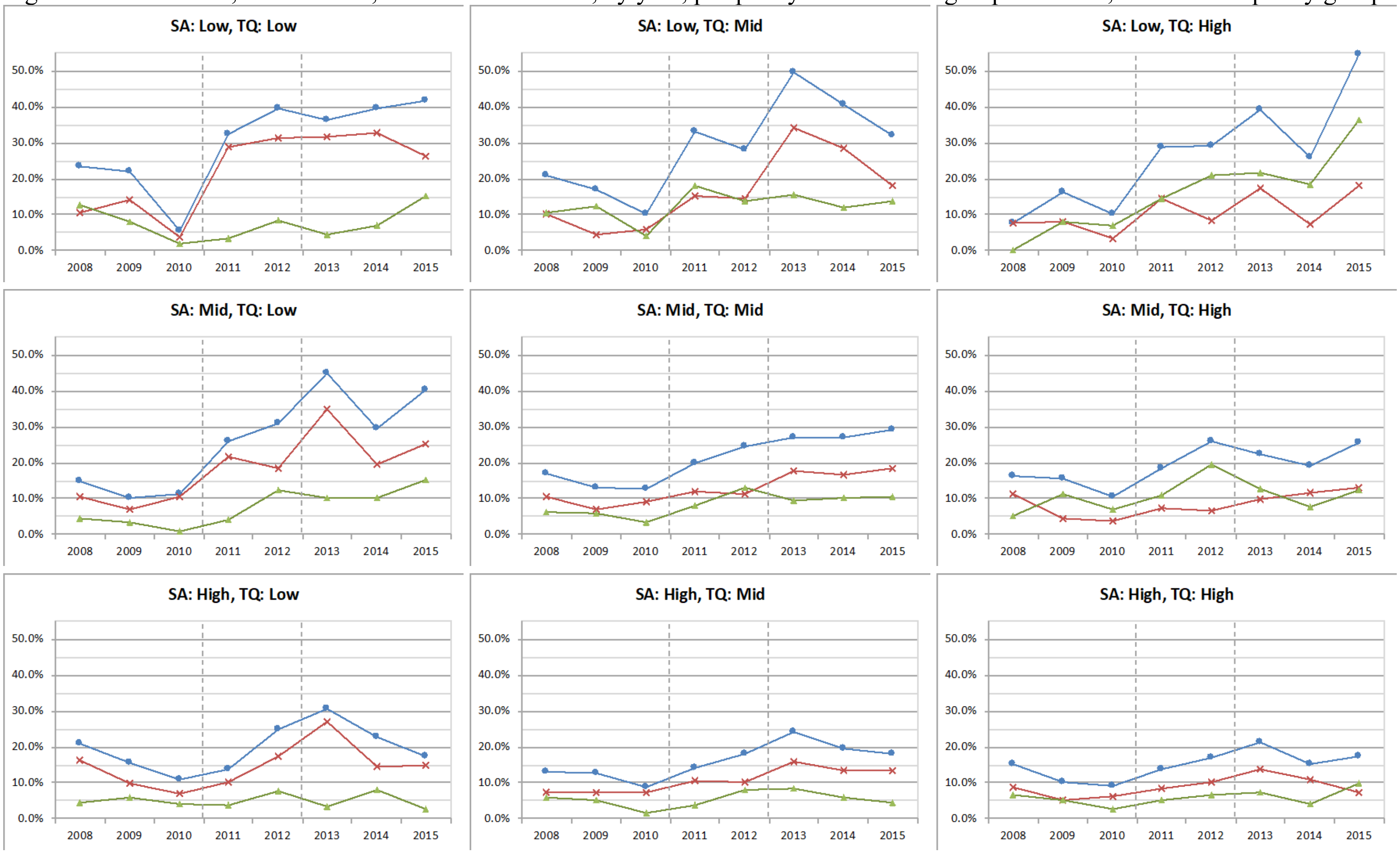
Notes: Circle markers: school exits; Cross markers: district exits; Triangle markers: school transfers. Teacher quality quintiles are determined based on the average of a teacher's jackknifed math and reading measures for teachers who have both. Teachers with a jackknifed measure in only one subject are divided into quintiles based on the distribution of all teachers with a jackknifed measure in the given subject. The bottom and top quintiles are those identified as least and most effective, respectively, while the middle-quintile teachers are those in quintiles 2-4. Teachers with no jackknifed measure but a current year EVAAS score in either subject are included in the "Unknown" category. For other details, see the notes to Figure 1.

Figure 3. School exits, district exits, and school transfers, by year and pre-policy achievement group of school



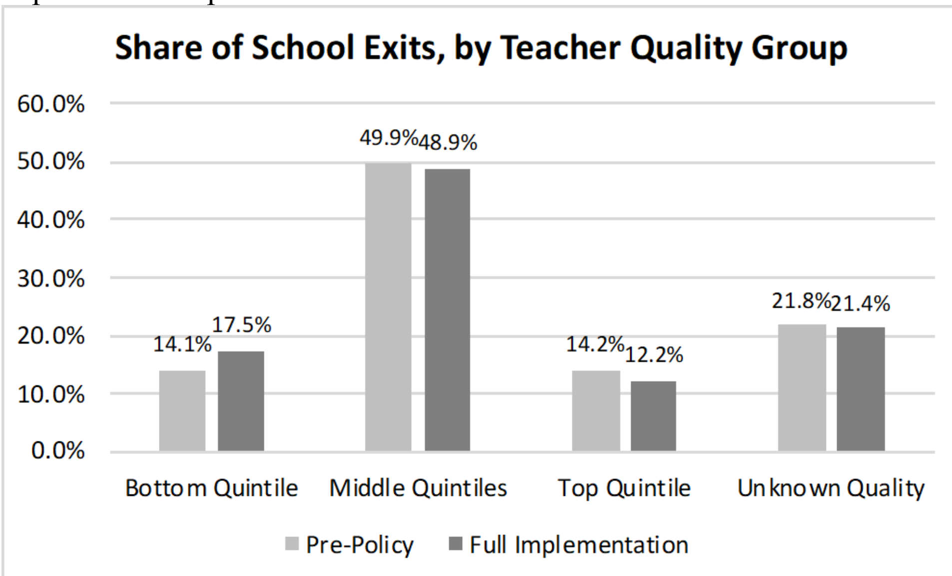
Notes: Circle markers: school exits; Cross markers: district exits; Triangle markers: school transfers. From left to right, the charts show turnover rates at bottom-quintile, middle-quintiles (2-4), and top-quintile schools in the pre-policy (average of reading and math) achievement distribution. For other details, see the notes to Figure 1.

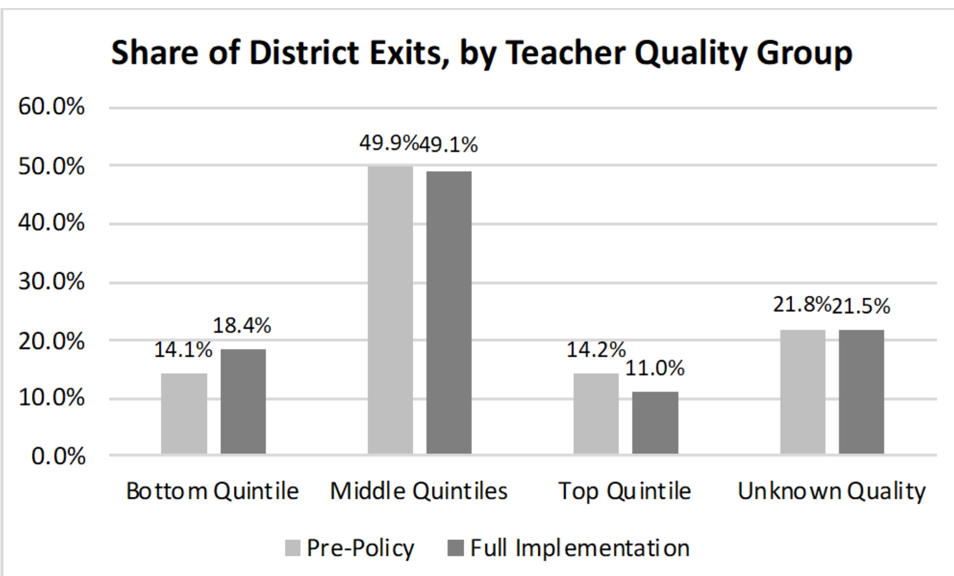
Figure 4. School exits, district exits, and school transfers, by year, pre-policy achievement group of school, and teacher quality group



Notes: Circle markers: school exits; Cross markers: district exits; Triangle markers: school transfers. SA abbreviates "school achievement" and TQ abbreviates "teacher quality" in the chart titles. Low (quintile 1), mid (quintiles 2-4), and high (quintile 5) indicate the distributional placement of teachers and schools. Each chart shows exit patterns for teachers in a teacher quality group crossed with a school achievement group. For other details, see the notes to Figure 1.

Figure 5. Share of school and district exits by teacher quality group, pre-policy and full-implementation periods





Notes: The figures are derived from model-based estimates of the shares of exiting teachers falling in each quality-quintile group in the pre-policy and full-implementation periods, from columns 1 and 2 of Table 2. The top panel shows the shares for school exits, and the bottom panel shows the shares for the subset of school exits that are also district exits. In each case, only the point estimates that are statistically significant (at least at the 10 percent level) are taken into account. Since there are no significant differential exit rates across groups during the pre-policy period, the pre-policy shares of exits reflect each group's representation in the full sample of teachers and do not differ across school and district exits. In the full-implementation period, all groups' exit rates increase, with a differentially greater (smaller) increase for teachers in the bottom (top) quintile. For example, the school exit rate for bottom-quintile teachers in the full-implementation period is calculated as 0.134 + 0.163 + .080, as per the estimates in column 1 of Table 2.

Table 1. Summary statistics for pre-policy years, 2008-2010

	Schools by achievement level		
	Low	Middle	High
Student characteristics			
Average achievement z-scores	-0.329	-0.082	0.436
Percent free lunch	55.8%	57.4%	34.3%
Percent reduced-price lunch	8.8%	11.0%	11.1%
Percent Black	39.6%	21.4%	20.5%
Percent Hispanic	58.2%	74.4%	47.8%
Percent ESL	16.5%	8.1%	4.9%
Total number of grade 3-8 students tested	32,968	109,311	55,239
Teacher quality			
Jackknifed quality measure	-0.012	0.030	0.089
Percent bottom-quintile quality	14.3%	15.6%	10.4%
Percent quintiles 2-4 quality	52.0%	48.9%	51.2%
Percent top-quintile quality	8.6%	14.2%	17.4%
Percent with unknown quality	25.0%	21.4%	21.1%
Percent new to focal grades/subjects	13.7%	11.1%	10.3%
Percent new to HISD	11.3%	10.3%	10.8%
Teacher characteristics			
Percent with 0-2 years of experience	24.6%	22.1%	19.6%
Percent with 3-5 years of experience	18.7%	17.5%	18.1%
Percent with 6-12 years of experience	28.6%	28.9%	28.8%
Percent with 13-20 years of experience	13.7%	16.0%	15.3%
Percent with 21-27 years of experience	7.6%	7.7%	8.7%
Percent with > 27 years of experience	6.8%	7.7%	9.6%
Percent with master's degree	34.1%	30.3%	31.2%
Percent with doctoral degree	1.9%	1.4%	1.3%
Percent Hispanic	21.7%	34.9%	22.6%
Percent American Indian	3.0%	2.8%	1.6%
Percent Asian/Pacific Islander	6.1%	5.3%	4.4%
Percent Black	55.5%	38.2%	24.0%
Percent White	37.8%	55.1%	71.5%
Percent Female	74.6%	78.0%	83.6%
Teacher turnover			
Exited the school in t+1	16.7%	14.0%	12.7%
Exited the district in t+1	8.0%	8.8%	8.1%
Transferred to another school in t+1	8.7%	5.2%	4.5%
N (Teacher-year)	1,067	4,505	1,849

Notes: The columns present summary statistics for 2008-2010 analysis teachers. Teachers are divided into three groups based on pre-policy achievement levels of their schools, averaged across reading and math: bottom quintile, quintiles 2-4, and top quintile. The bottom-quintile sample is smaller because treated schools in Fryer's 2014 study are omitted and low-achieving schools have lower enrollment on average. Years are identified by the fiscal year in this and all following tables, which is the same as the spring of the school year. For example, 2008 refers to school year 2007-08. Student characteristics are derived from student-level data and are averages for grade 3-8 students taught by analysis teachers in the pre-period. Teacher characteristics are for all teachers with current EVAAS scores in reading and/or math, other than the jackknifed quality measure which is only available for a subset of these. Teacher quality quintiles are determined based on the average of a teacher's jackknifed math and reading measures for teachers who have both. Teachers with a jackknifed measure in only one subject are divided into quintiles based on the distribution of all teachers with a jackknifed measure in the given subject. Teachers with no jackknifed measure but a current year EVAAS score in either subject are included in the "Unknown" category. School exits are the sum of district exits and school transfers.

Table 2. Impacts on turnover, by teacher quality group and evaluation period

	School exit	District exit	School transfer
	(1)	(2)	(3)
Intercept	0.134	0.084	0.050
	(0.006)**	(0.004)**	(0.004)**
Intercept*PARTIAL	0.101	0.049	0.053
	(0.011)**	(0.007)**	(0.007)**
Intercept*FULL	0.163	0.108	0.055
	(0.012)**	(0.009)**	(0.007)**
Bottom quintile	-0.006	0.013	-0.019
	(0.011)	(0.010)	(0.007)**
Bottom quintile*PARTIAL	0.068	0.076	-0.008
	(0.020)**	(0.020)**	(0.013)
Bottom quintile*FULL	0.080	0.062	0.019
	(0.022)**	(0.021)**	(0.014)
Top quintile	0.003	-0.010	0.014
	(0.013)	(0.009)	(0.009)
Top quintile*PARTIAL	-0.001	-0.011	0.010
	(0.023)	(0.015)	(0.016)
Top quintile*FULL	-0.035	-0.040	0.005
	(0.020)*	(0.014)**	(0.014)
Unknown	-0.018	-0.013	-0.004
	(0.011)	(0.010)	(0.007)
Unknown*PARTIAL	0.007	0.006	0.001
	(0.020)	(0.016)	(0.013)
Unknown*FULL	0.015	0.014	0.001
	(0.018)	(0.015)	(0.012)
R-squared	0.087	0.067	0.058
N (Teacher-year)	18,681	18,681	18,681

Notes: Each column reports coefficient estimates, with standard errors clustered at the school level in parentheses, from a regression of the dependent variable indicated in the column heading. Whether an exit or transfer has occurred is based on teacher transitions observed between years t and t+1. In addition to the variables shown, the models also include the teacher characteristics shown in Table 1 and school fixed effects. All control variables other than the period indicators are mean-centered so the intercept can be interpreted as the exit rate at the mean values of all other covariates. The omitted teacher quality group includes teachers in quality quintiles 2, 3, and 4. Observations for teachers from 2011 and 2012 are coded as "PARTIAL", while observations from 2013, 2014, and 2015 are coded as "FULL". The parameters for the variables interacted with PARTIAL and FULL are estimated relative to the pre-policy years (2008-2010). ** Denotes statistical significance at the 5 percent level; * Denotes statistical significance at the 10 percent level.

Table 3. Tests for robustness of impacts on turnover, by teacher quality group and evaluation period

	1-year exit definition 2008-2014 cohorts		2-year exit definition 2008-2014 cohorts		Principal-by-school fixed effects 1-year exit definition 2008-2015 cohorts		
	School exit	District exit	School exit	District exit	School exit	District exit	School transfer
	(1a)	(2a)	(1b)	(2b)	(3)	(4)	(5)
Intercept	0.136	0.085	0.130	0.076	0.140	0.089	0.051
	(0.006)**	(0.004)**	(0.006)**	(0.004)**	(0.007)**	(0.005)**	(0.004)**
Intercept*PARTIAL	0.101	0.049	0.101	0.046	0.098	0.041	0.057
	(0.011)**	(0.007)**	(0.011)**	(0.006)**	(0.012)**	(0.007)**	(0.009)**
Intercept*FULL	0.165	0.114	0.162	0.108	0.150	0.101	0.049
	(0.013)**	(0.010)**	(0.013)**	(0.010)**	(0.014)**	(0.011)**	(0.009)**
Bottom quintile	-0.007	0.013	-0.003	0.017	-0.003	0.013	-0.016
	(0.011)	(0.010)	(0.011)	(0.009)*	(0.011)	(0.010)	(0.007)**
Bottom quintile*PARTIAL	0.068	0.075	0.064	0.066	0.057	0.069	-0.012
	(0.020)**	(0.020)**	(0.020)**	(0.019)**	(0.019)**	(0.019)**	(0.014)
Bottom quintile*FULL	0.078	0.071	0.076	0.063	0.070	0.058	0.012
	(0.025)**	(0.025)**	(0.025)**	(0.024)**	(0.022)**	(0.021)**	(0.013)
Top quintile	0.003	-0.011	0.004	-0.009	0.003	-0.011	0.014
	(0.013)	(0.009)	(0.013)	(0.008)	(0.013)	(0.009)	(0.009)
Top quintile*PARTIAL	-0.001	-0.011	-0.003	-0.012	0.004	-0.008	0.012
	(0.023)	(0.015)	(0.023)	(0.015)	(0.023)	(0.015)	(0.016)
Top quintile*FULL	-0.055	-0.044	-0.056	-0.044	-0.037	-0.043	0.006
	(0.022)**	(0.017)**	(0.021)**	(0.017)**	(0.020)*	(0.015)**	(0.015)
Unknown	-0.018	-0.015	-0.016	-0.013	-0.020	-0.015	-0.005
	(0.011)	(0.010)	(0.011)	(0.009)	(0.011)*	(0.010)	(0.007)
Unknown*PARTIAL	0.007	0.006	0.011	0.016	0.004	0.004	0.001
	(0.020)	(0.016)	(0.020)	(0.015)	(0.021)	(0.016)	(0.013)
Unknown*FULL	-0.003	0.006	0.000	0.009	0.007	0.010	-0.003
	(0.022)	(0.019)	(0.022)	(0.018)	(0.019)	(0.016)	(0.012)
R-squared	0.085	0.068	0.085	0.066	0.116	0.086	0.090
N (Teacher-year)	16,670	16,670	16,670	16,670	18,681	18,681	18,681

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. In the left panel, we examine teacher cohorts in 2008-14 only, for whom we can define exits looking forward both 1 and 2 years, to examine sensitivity to the exit definition holding all else equal. In the right panel, we return to using our primary single-year exit definition and 2008-15 cohorts but expand the model to include principal-by-school fixed effects.

Table 4. Impacts on turnover, by teacher quality group and evaluation period, using current-year EVAAS scores in place of jackknifed measures

	School exit (1)	District exit (2)	School transfer (3)
Intercept	0.130	0.081	0.049
	(0.005)**	(0.004)**	(0.003)**
Intercept*PARTIAL	0.103	0.051	0.053
	(0.010)**	(0.007)**	(0.007)**
Intercept*FULL	0.166	0.111	0.055
	(0.011)**	(0.008)**	(0.006)**
Bottom quintile	0.040	0.030	0.010
	(0.012)**	(0.009)**	(0.008)
Bottom quintile*PARTIAL	0.006	0.007	-0.002
	(0.022)	(0.016)	(0.014)
Bottom quintile*FULL	0.029	0.035	-0.006
	(0.021)	(0.017)**	(0.013)
Top quintile	-0.006	-0.003	-0.004
	(0.010)	(0.008)	(0.007)
Top quintile*PARTIAL	-0.007	-0.021	0.014
	(0.019)	(0.014)	(0.014)
Top quintile*FULL	-0.039	-0.049	0.010
	(0.018)**	(0.014)**	(0.012)
R-squared	0.088	0.066	0.057
N (Teacher-year)	18,681	18,681	18,681

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. The specifications are the same other than that the teacher quality quintiles are based on current-year EVAAS scores. There are no teachers in the analysis sample with unknown quality according to this measure, which is why those additional variables are not included.

Table 5. Impacts on turnover, by teacher quality group and year

	School exit (1)	District exit (2)	School transfer (3)
Bottom quintile	0.001 (0.022)	0.022 (0.018)	-0.021 (0.014)
Bottom quintile*2009	-0.018 (0.031)	-0.012 (0.026)	-0.006 (0.019)
Bottom quintile*2010	-0.005 (0.029)	-0.016 (0.024)	0.011 (0.017)
*Bottom quintile*2011*	*0.044 (0.034)*	*0.066 (0.030)***	*-0.022 (0.019)*
*Bottom quintile*2012*	*0.078 (0.033)***	*0.068 (0.029)***	*0.010 (0.024)*
Bottom quintile*2013	0.108 (0.041)**	0.110 (0.039)**	-0.003 (0.025)
Bottom quintile*2014	0.030 (0.035)	0.013 (0.032)	0.017 (0.024)
Bottom quintile*2015	0.082 (0.042)*	0.033 (0.033)	0.049 (0.026)*
Top quintile	-0.003 (0.024)	0.014 (0.019)	-0.017 (0.015)
Top quintile*2009	0.018 (0.032)	-0.026 (0.024)	0.044 (0.024)*
Top quintile*2010	0.002 (0.029)	-0.043 (0.022)*	0.045 (0.022)**
*Top quintile*2011*	*-0.013 (0.034)*	*-0.039 (0.025)*	*0.026 (0.023)*
*Top quintile*2012*	*0.025 (0.034)*	*-0.031 (0.025)*	*0.056 (0.027)***
Top quintile*2013	-0.039 (0.040)	-0.078 (0.032)**	0.039 (0.023)*
Top quintile*2014	-0.056 (0.34)*	-0.060 (0.027)**	0.004 (0.024)
Top quintile*2015	0.012 (0.037)	-0.053 (0.029)*	0.066 (0.025)**
Unknown	-0.031 (0.018)*	-0.013 (0.016)	-0.018 (0.013)
Unknown*2009	0.008 (0.025)	0.000 (0.021)	0.009 (0.017)
Unknown*2010	0.022 (0.027)	-0.007 (0.021)	0.029 (0.017)
*Unknown*2011*	*-0.004 (0.028)*	*-0.009 (0.023)*	*0.005 (0.019)*
*Unknown*2012*	*0.035 (0.029)*	*0.017 (0.024)*	*0.019 (0.020)*
Unknown*2013	-0.037 (0.031)	-0.033 (0.027)	-0.004 (0.020)
Unknown*2014	0.063 (0.034)*	0.050 (0.030)*	0.012 (0.022)
Unknown*2015	0.081 (0.034)**	0.036 (0.028)	0.044 (0.026)*
R-squared	0.090	0.070	0.062
N (Teacher-year)	18,681	18,681	18,681

Notes: The specifications are the same as in Table 2, other than that indicators for each year replace the evaluation period indicators and the estimates of turnover by year for the baseline middle-quintile categories are not shown for brevity. In this table, the year-specific parameters are estimated relative to 2008, which is the first year of the panel. The italicized rows are for the partial-implementation period, and thus separate years from the pre-policy and full-implementation periods. For more details, see the notes to Table 2.

Table 6. Impacts on turnover, by teacher experience level, HISD compared to the rest of Texas

	HISD only		All TX districts	
	School exit	District exit	School exit	District exit
	(1)	(2)	(3)	(4)
Intercept	0.131	0.082	0.152	0.112
	(0.006)**	(0.004)**	(0.001)**	(0.001)**
Intercept*PARTIAL	0.103	0.049	0.004	0.000
	(0.010)**	(0.007)**	(0.001)**	(0.001)
Intercept*FULL	0.164	0.109	0.043	0.035
	(0.011)**	(0.008)**	(0.001)**	(0.001)**
Exp≤5	0.041	0.036	0.045	0.037
	(0.011)**	(0.008)**	(0.001)**	(0.001)**
(Exp≤5)*PARTIAL	0.032	0.034	0.009	0.006
	(0.017)*	(0.014)**	(0.002)**	(0.002)**
(Exp≤5)*FULL	0.007	0.026	0.000	0.005
	(0.017)	(0.014)*	(0.002)	(0.002)**
Exp>20	0.001	0.024	-0.005	0.009
	(0.012)	(0.009)**	(0.002)**	(0.002)**
(Exp>20)*PARTIAL	0.019	0.031	0.040	0.041
	(0.020)	(0.018)*	(0.003)**	(0.002)**
(Exp>20)*FULL	-0.065	-0.031	0.012	0.016
	(0.018)**	(0.016)*	(0.002)**	(0.002)**
HISD*PARTIAL			0.060	0.030
			(0.008)**	(0.005)**
HISD*FULL			0.076	0.048
			(0.008)**	(0.005)**
HISD*(Exp≤5)			0.004	0.012
			(0.006)	(0.005)**
HISD*(Exp≤5)*PARTIAL			0.027	0.027
			(0.010)**	(0.008)**
HISD*(Exp≤5)*FULL			0.018	0.016
			(0.009)*	(0.008)**
HISD*(Exp>20)			0.008	0.012
			(0.009)	(0.008)
HISD*(Exp>20)*PARTIAL			0.015	0.022
			(0.013)	(0.012)*
HISD*(Exp>20)*FULL			-0.048	-0.033
			(0.012)**	(0.011)**
R-squared	0.082	0.057	0.002	0.009
N (Teacher-year)	18,681	18,681	1,716,949	1,716,949

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. In the first two columns, the sample and specifications are the same other than that the indicators for teacher quality quintiles are replaced by indicators for levels of experience. The omitted experience group includes teachers with 6 to 20 years of experience. The estimates in the last two columns use statewide data and the primary analytic variables in the model (teacher experience and policy implementation period) are additionally interacted with an indicator variable for HISD. Note that there is no main effect for HISD since this is subsumed by the school fixed effects.

Table 7. Estimated effects of turnover on student achievement

	Dependent variable: Difference in average test scores across consecutive cohorts within a school-grade		
	Stacked math and reading	Math only	Reading only
	(1)	(2)	(3)
Difference in teacher turnover	-0.055	-0.072	-0.040
	(0.013)**	(0.019)**	(0.015)**
Teacher share, bottom-quintile exits	0.097	0.133	0.072
	(0.027)**	(0.040)**	(0.034)**
Teacher share, middle-quintiles exits	0.042	0.038	0.056
	(0.019)**	(0.031)	(0.023)**
Teacher share, top-quintile exits	-0.146	-0.196	-0.068
	(0.035)**	(0.052)**	(0.036)*
Teacher share, unknown-quality exits	0.104	0.116	0.099
	(0.024)**	(0.039)**	(0.027)**
R-squared	0.097	0.069	0.180
N (School-grade-year)	8,095	4,038	4,057

Notes: Each column shows estimates from a separate regression with standard errors clustered at the school-by-grade level in parentheses. The dependent variable is the one-year difference in the school-by-grade average math and reading z-scores in columns 2 and 3, respectively, with results for these two subjects stacked shown in column 1. Teacher turnover for a school-by-grade cohort is defined as the share of teachers exiting the grade at the end of the prior year, and the difference across cohorts is intended to capture differential exposure to disruption. The shares of teachers by school-grade exiting between cohorts and falling in each of the teacher quality groups is intended to capture changes in teacher quality. Teachers are scaled by instructional percentages in the given subject prior to exit. The quality groupings, inclusive of the unknown quality group (i.e., teachers without jackknifed quality measures), are exhaustive. The control set also includes the differences in the school-grade averages across cohorts for the student characteristics shown in Table 1 and school fixed effects.

Table 8. Policy effects on per-turnover quality and district workforce quality after 10 years, in student test score standard deviations

	Scenarios	Effect on teacher quality per turnover (1)	Effect on average workforce quality after 10 years (2)
1	Baseline	0.008	0.009
2	Great Recession assumed to be responsible for half of the estimated policy impacts	0.006	0.006
3	Effect on level of turnover assumed to be 50% of baseline estimate	0.011	0.011
4	Effect on level of turnover assumed to be 25% of baseline estimate	0.014	0.013
5	Policy effects concentrated in the top and bottom deciles of the teacher quality distribution	0.010	0.012
6	Replacement teacher quality distribution shifted so that average replacement teacher is 0.25 standard deviations below average, baseline policy impacts (as in row 1)	0.008†	-0.007
7	Replacement teachers have negative first-year disruption effect, assuming policy impact on level of turnover at baseline (as in row 1)	0.008†	0.005
8	Replacement teachers have negative first-year disruption effect, assuming policy impact on level of turnover at 50% of baseline (as in row 3)	0.011†	0.007
9	Replacement teachers have negative first-year disruption effect, assuming policy impact on level of turnover at 25% of baseline (as in row 4)	0.014†	0.008

Notes: These calculations are based on the regression estimates in Table 2 for district exit as described in the text. Column 1 shows how the average quality of district exiters changes concurrently, while column 2 simulates the associated change in average workforce quality after 10 years. Appendix D provides details on the simulations used to populate column 2. †Indicates the value shown is a repeat value from a previous row of the table. Replacement teachers do not factor into the calculations in column 1 because column 1 is only concerned with the effect on the quality of exiters.

Appendices

Appendix A. Additional background on HISD policies

The two major policies affecting teachers in HISD during our sample period, which spans schools years 2007-08 through 2014-15, are the effective teacher initiative (ETI) and the merit pay system (ASPIRE). Figure A1 presents a timeline of the most relevant ETI and ASPIRE events. The new teacher evaluation system was phased in from school years 2010-11 to 2012-13. Though the merit pay system was in effect throughout, there were periodic adjustments to the eligibility criteria and award amounts. For the teachers we study, who are teaching core subjects in grades 3-8, the two systems are tied primarily by reliance to differing degrees on teachers' EVAAS scores.

The teacher evaluation reform: ETI

One of the priorities of the ETI reform was to encourage retention of more effective educators and exit of less effective educators. Combined with attempting to hire better teachers and offering individualized training to existing teachers, the hope was to shift the distribution of teacher effectiveness in the district (as illustrated in Figure A2). Though high rates of turnover had always been a concern, the focus on selective retention/exit was new and has continued to be an important focus in HISD. Prior to 2010-11, the district did not produce indicators for differential turnover by teacher effectiveness, in large part because teacher ratings were not universally recorded with the district. Since then, the retention rate of highly effective teachers and exit rate of ineffective teachers have been among the set of key indicators of progress reported in the annual Facts and Figures brief released by HISD to the public.[32]

The new teacher evaluation system, known as the Teacher Appraisal and Development System (TADS), is the cornerstone of the ETI. Under this system, teachers are evaluated on three components: instructional practice (IP), professional expectations (PE), and student performance (SP). Teachers are rated on a scale of 1 to 4 on each of these components. The ratings are then combined to provide a summary rating of ineffective (I), needs improvement (NI), effective (E), or highly effective (HE).

While the IP and PE ratings are based on scoring observed teacher practices against well-

[32] The superintendent's message in the district's 2011 Annual Report conveys this priority: "In 2011, we took bold steps to transform the way teachers are recruited, trained, evaluated, and retained. […] HISD is committed to recognizing and rewarding top teachers. And teachers whose students consistently demonstrate weak academic growth are asked to exit the organization." Further, highlighted in a box on the first page is: "In 2010-11, 373 teachers exited the organization for performance reasons. That's up from 77 in 2009."

defined rubrics, the SP rating is based on composite measures of student performance. For teachers in subjects and grades where EVAAS scores are available, part is based on these scores (VA SP) and part is based on non-value added measures (non-VA SP). The intermediate VA SP is assigned a performance level of 1 to 5 based on a teacher's cumulative gain index (CGI) which combines EVAAS scores across all grades and subjects taught by a teacher in a given year.[33] The CGI can be interpreted as standard deviations relative to the average HISD teacher. The lowest performance level is assigned to those far below average (CGI below -2), the highest level to those far above average (CGI above 2), while the middle level corresponds to those who are not detectably different from average (CGI between -1 and 1). The intermediate non-VA SP is assigned a performance level of 1 to 4 based on student progress on districtwide assessments relative to other students with the same prior year score. Students are assigned percentile ranks based on relative progress, and teachers' are then rated based on the median percentile rank across students taught.

Prior to 2014-15, the process by which the ratings across the three components were combined into a summative rating ended up giving nearly 50 percent weight to SP. First, the level ratings ranging from 1-4 were calculated for IP and PE. These were then combined into an IP x PE rating, according to the "Step 1" matrix shown below. In 2011-12, this was also a teacher's summative rating. When SP was incorporated starting in 2012-13, the performance levels based on value-added measures and non-value added measures were each calculated and then combined into an overall SP rating according to the "Step 2" matrix. The IP x PE rating and SP rating were then combined using the "Step 3" matrix.

Step 1. IP x PE rating

		IP			
		1	2	3	4
PE	1	1	2	2	3
	2	1	2	3	3
	3	1	2	3	4
	4	2	2	3	4

Step 2. SP rating

		VA SP				
		1	2	3	4	5
Non-VA SP	1	1	2	2	2	3
	2	1	2	3	3	3
	3	1	2	3	4	4
	4	2	3	3	4	4

Step 3. Summative rating

		SP			
		1	2	3	4
IP x PE	1	I	NI	NI	NI
	2	I	NI	E	E
	3	NI	NI	E	HE
	4	NI	E	E	HE

In response to teacher concerns about transparency and year-to-year volatility due to the high

[33] The district voted not to renew the contract with the SAS Institute in June 2016, beyond the range of our study period, as a reaction to teacher concerns about opaqueness of the proprietary EVAAS measures. After a gap year where teachers were again only rated on the observational components, the district developed its own value-added metrics for this component.

reliance on test scores, the calculation was simplified to set the weights at 50% for IP, 20% for PE, and 30% for SP (with VA SP weighted 20% and non-VA SP weighted 10%) in 2014-15.

Table A1 shows the fractions of teachers in our analysis sample receiving each rating level for the summative rating and the three main components, as well as the intermediate SP measures. The least discriminating component is PE; no teacher receives the lowest rating and only a few percent receive the second lowest. IP is somewhat more discriminating, though more than 4 out of 5 teachers receive one of the top two ratings. For SP, this is true for only about 3 out of 5 teachers. Not surprisingly, the inclusion of SP starting in 2012-13 shrinks the share rated effective and expands the share rated needs improvement, which then reverses to some extent with the reweighting in 2015. Table A2 documents correspondences in the ratings teachers receive across several of the components and reveals that in practice the SP rating is almost perfectly predicted by the VA SP rating, which is based on current year EVAAS scores. Also, it is only those who receive the lowest VA SP rating that are at risk of being classified as ineffective.

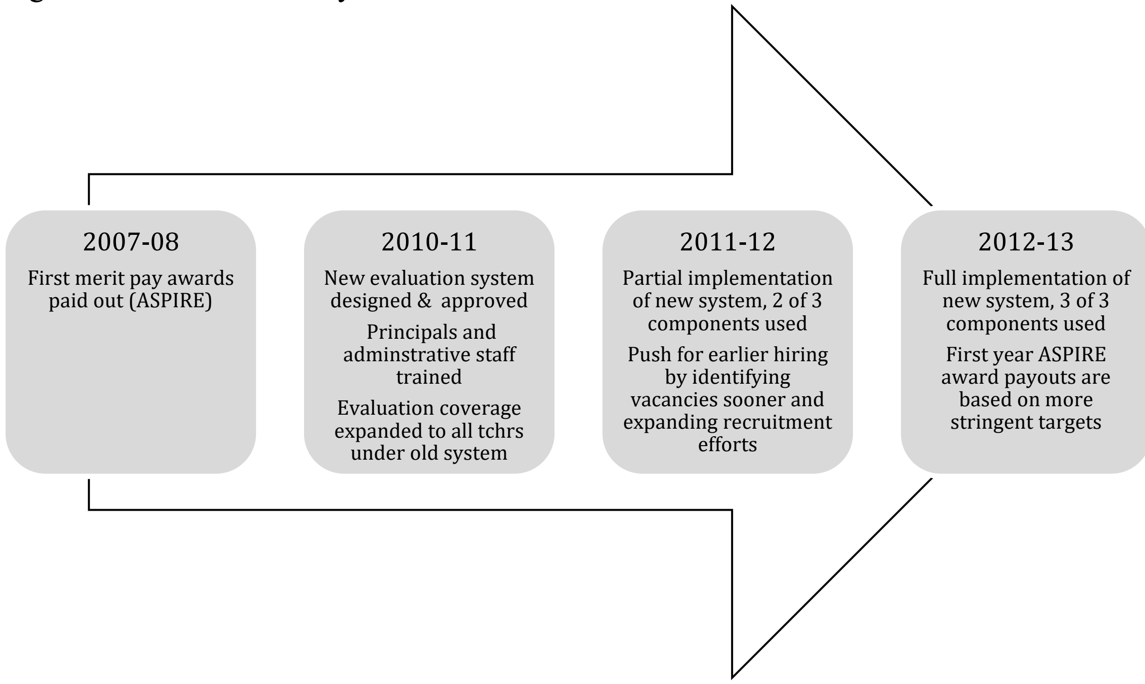
Figure A3 shows the how ratings are related to the teacher quality groupings based on our jackknife method, which are not as sensitive to year-to-year fluctuations in test scores. The shares classified in more effective categories increase monotonically moving from lower to higher quality groups. However, more than half of teachers in the bottom quintile of the quality distribution are classified as effective or highly effective, and there is some slippage at the top end as well. The teachers of unknown quality, who are either new to the district or new to teaching reading or math in a tested grade, look very similar to bottom-quintile teachers in terms of the ratings they receive.

The merit pay system: ASPIRE

The merit pay system was first introduced in 2006-07, one year before the start of our data panel, with the associated payouts first received in 2007-08. Table A3 shows the evolution of the provisions that are relevant to teachers of core subjects in grades 3-8 over our study period. As far as generosity of the awards for recipients, amounts were increased in 2008-09 and then reduced in 2011-12, before another increase in 2012-13 and then a decline for the last year of our data panel in 2014-15. Eligibility was tightened somewhat in 2010-11 due to the introduction of a teacher attendance requirement and minimum threshold for student growth. There were then more dramatic changes in eligibility in 2011-12 when performance targets jumped, and in 2012-13 when teachers rated needs improvement or ineffective were precluded from receiving awards. The latter requirement coincided with full implementation of the new evaluation system and directly tied ASPIRE award eligibility to system ratings. Teachers experienced all of these changes with a one year lag, since awards are announced and paid the following year (see the last column of the table). Importantly, awards are paid regardless of whether the teacher is still an employee.

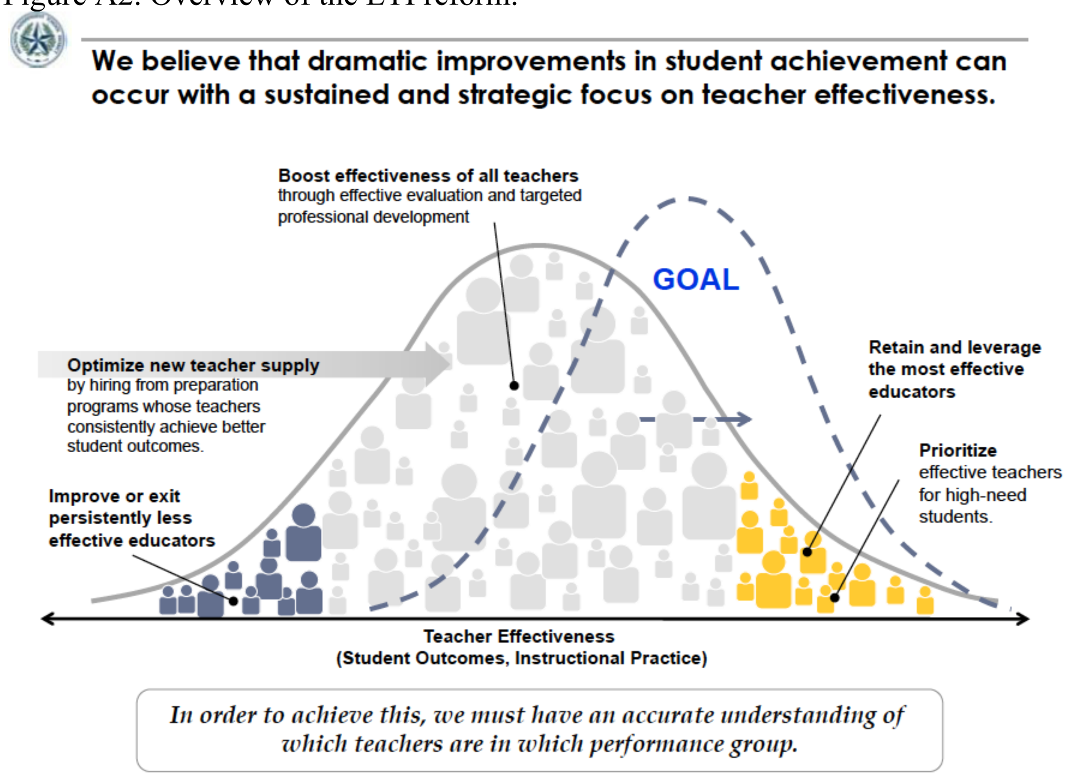
Figure A4 shows the implications for the teachers who are the focus of our analysis. In the figure, teachers are assigned to quality quintiles using the jackknife method described in the main text. The first graph shows that award receipt was nearly universal in the pre-policy period. The share receiving awards fell steadily across years after the reform, with greater drops for the lower-quality groups. The fluctuations in average award amounts across years in the middle graph primarily reflect the statutory changes to the generosity of the maximums shown in Table A3, with the cut in 2011-12 and the increase in 2012-13, realized in payouts in 2012-13 and 2013-14, respectively. Finally, the last graph shows average award amounts unconditional on receipt. Here it is clear that the changes to provisions in 2010-11 (realized in 2011-12 payouts) had minimal effects, while subsequent changes ultimately lowered average awards for bottom-quintile teachers by more than two-thirds and for middle-quintile teachers by almost one-third. The top quintile was more or less held harmless.

Figure A1. Timeline of key ETI and ASPIRE events at HISD



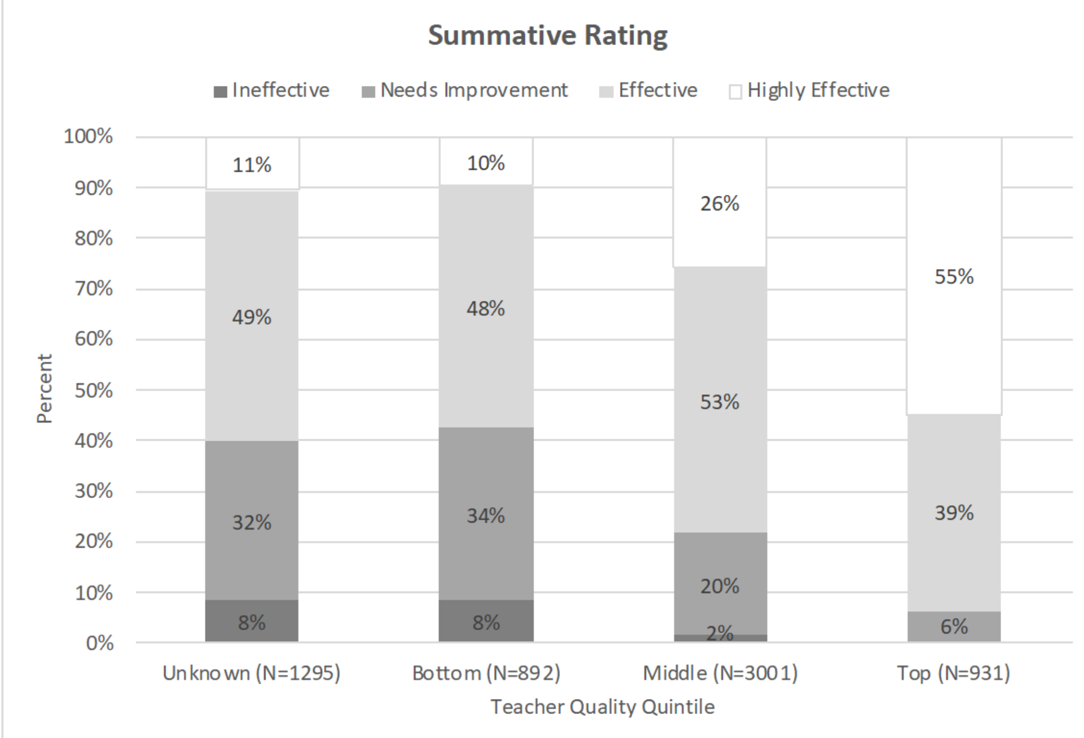
Notes: ASPIRE award timing is based on when award payouts are made, which is the year following the evaluation period.

Figure A2. Overview of the ETI reform.



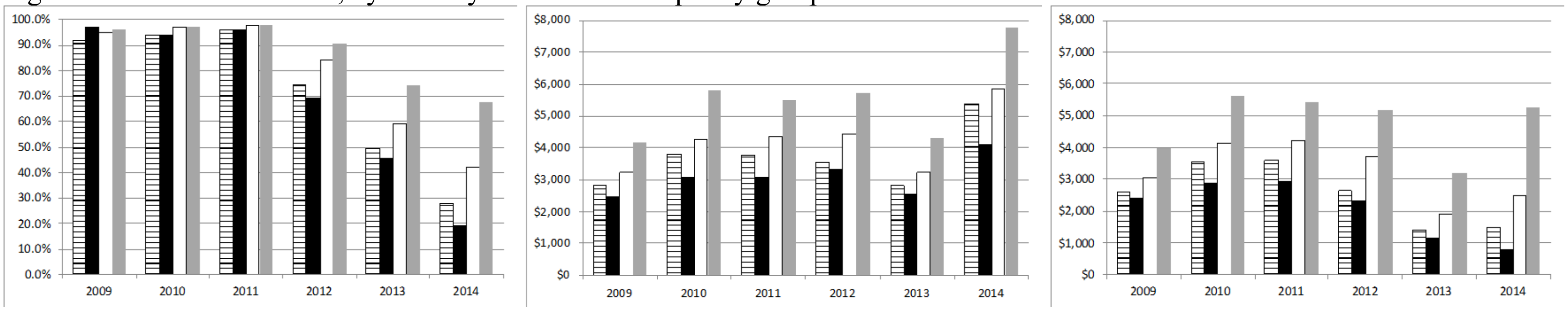
Notes: The source for this figure is “Teacher appraisal systems: how one urban school district is linking effective teaching to student achievement,” presented by Superintendent Grier at the American Association of School Administrators meeting on February 17, 2012.

Figure A3. Summative ratings by teacher quality group



Notes: Each column corresponds to the teacher quality group indicated in the heading. Shown are the shares of analysis teachers in the quality group receiving each summative rating across school years 2011-12 to 2014-15.

Figure A4. ASPIRE Awards, by award year and teacher quality groups



Notes: Striped bars: teachers of unknown quality; Black bars: bottom-quintile teachers; Clear bars: middle-quintile teachers; Gray bars: top-quintile teachers in the quality distribution. From left to right, the graphs show the share of teachers receiving awards, the average award per recipient, and the average award per teacher unconditional on receipt. See notes to Figure 2 for details on how teachers are assigned to quality quintiles. The years reflect the fiscal year of award receipt. For example, 2009 refers to awards paid out in the middle of school year 2008-2009 based on performance in school year 2007-2008. We do not have access to the ASPIRE award microdata for 2008 and 2015, so the time period is restricted to payout years 2009 through 2014.

Table A1. Appraisal components ratings distributions for analysis teachers

	2011 (1)	2012 (2)	2013 (3)	2014 (4)	2015 (5)
Summative rating					
Ineffective (I)	0.00	0.01	0.07	0.04	0.00
Needs improvement (NI)	0.03	0.11	0.29	0.22	0.16
Effective (E)	0.67	0.59	0.42	0.43	0.63
Highly effective (HE)	0.30	0.29	0.22	0.30	0.21
Instructional practice (IP)					
Level 1	n/a	0.01	0.02	0.01	0.01
Level 2	n/a	0.11	0.18	0.13	0.11
Level 3	n/a	0.69	0.62	0.63	0.61
Level 4	n/a	0.20	0.18	0.23	0.27
Professional expectations (PE)					
Level 1	n/a	0.00	0.00	0.00	0.00
Level 2	n/a	0.03	0.04	0.02	0.02
Level 3	n/a	0.76	0.75	0.74	0.66
Level 4	n/a	0.21	0.21	0.25	0.31
Student performance (SP)					
Level 1	n/a	n/a	0.21	0.15	0.15
Level 2	n/a	n/a	0.16	0.12	0.31
Level 3	n/a	n/a	0.40	0.41	0.31
Level 4	n/a	n/a	0.23	0.32	0.23
Value-added student performance (VA SP)					
Level 1	n/a	n/a	0.21	0.16	-
Level 2	n/a	n/a	0.16	0.12	-
Level 3	n/a	n/a	0.37	0.38	-
Level 4	n/a	n/a	0.11	0.13	-
Level 5	n/a	n/a	0.15	0.21	-
Non-value added student performance (non-VA SP)					
Level 1	n/a	n/a	0.05	0.05	-
Level 2	n/a	n/a	0.33	0.27	-
Level 3	n/a	n/a	0.45	0.46	-
Level 4	n/a	n/a	0.17	0.22	-

Notes: The cells show the fraction of teachers receiving each rating level by fiscal year. The fiscal year is the same as the spring of the school year. For example, 2011 refers to fiscal year 2011 and represents school year 2010-11. The first column is drawn from HISD reports and is based on all core subject teachers, while the remaining columns are based on the math and reading teacher subset that is included in our analysis. Note that the 2012 summative ratings distribution for all core teachers is very similar to that for our analysis teachers: the fractions rated ineffective, needs improvement, effective, and highly effective are 0.01, 0.12, 0.61, and 0.26, respectively. In 2011, teachers were not formally scored under the new rubric for IP and PE but most principals had been trained. The SP metrics were not included in the system until 2013, so are not available before that year. While the intermediate VA SP and non-VA SP ratings were produced in 2015, HISD did not make these available to us in that year.

Table A2. Ratings distributions for analysis teachers, pooling all available years

		Professional expectations (PE) rating			
		1	2	3	4
Instructional practice (IP) rating	1	0.00	0.01	0.00	0.00
	2	0.00	0.02	0.11	0.00
	3	0.00	0.00	0.53	0.10
	4	0.00	0.00	0.08	0.14
		Non-VA student performance (non-VA SP) rating			
		1	2	3	4
VA student performance (VA SP) rating	1	0.04	0.10	0.05	0.01
	2	0.01	0.06	0.06	0.01
	3	0.01	0.11	0.20	0.05
	4	0.00	0.02	0.07	0.04
	5	0.00	0.02	0.07	0.09
		Overall student performance (SP) rating			
		1	2	3	4
VA student performance (VA SP) rating	1	0.18	0.01	0.00	0.00
	2	0.00	0.13	0.01	0.00
	3	0.00	0.01	0.36	0.00
	4	0.00	0.00	0.02	0.11
	5	0.00	0.00	0.02	0.16
		Summative rating			
		I	NI	E	HE
VA student performance (VA SP) rating	1	0.06	0.13	0.00	0.00
	2	0.00	0.11	0.02	0.00
	3	0.00	0.01	0.36	0.00
	4	0.00	0.00	0.02	0.10
	5	0.00	0.00	0.02	0.16

Notes: The cells show the fraction of analysis teachers receiving each cross-combination of ratings. For example, the top cell shows the fraction receiving an IP rating of 1 and a PE rating of 1. The cells are based on all available years for each component or subcomponent.

Table A3. ASPIRE program details

School year	Campus Performance Awards: Distributional Targets	Campus Performance Awards: Award Max	Individual Performance Awards: Distributional Targets	Individual Performance Awards: Award Max	Ineligible Teachers	Award Max	Awards Announced/ Paid
2007-08	Top 50% in campus growth within HISD, plus bonuses for top 50% in growth within comparable Texas schools and for attaining a state accountability rating above Acceptable	$2,600	Top 50% in student growth, pro-rated per subject if teaches multiple tested subjects	$5,000	None	$7,600	Nov-08/ Jan-09
2008-09	Same as prior year	$3,100	Same as prior year	$7,000	None	$10,100	Nov-09/ Jan-10
2009-10	Same as prior year	$3,100	Same as prior year	$7,000	None	$10,100	Nov-10/ Jan-11
2010-11	Same as prior year	$3,100	Same as prior year	$7,000	Those missing >10 days or with low student growth	$10,100	Nov-11/ Jan-12
2011-12	Top 20% in campus growth within HISD, plus bonuses for high growth or achievement in shares scoring above national medians in reading and/or math	$2,000	Top 15% in student growth	$7,000	Same as prior year	$9,000	Nov-12/ Jan-13
2012-13	Same as prior year	$3,000	Same as prior year	$10,000	In addition, those rated below effective	$13,000	Nov-13/ Jan-14
2013-14	Same as prior year	$3,000	Same as prior year	$10,000	Same as prior year	$13,000	Nov-14/ Jan-15
2014-15	Same as prior year	$2,250	Same as prior year	$7,500	Same as prior year	$9,750	Nov-15/ Jan-16

Notes: This table is constructed by the authors using various sources of information published by HISD.

Appendix B. Validating our teacher quality measures

In order to validate our teacher quality measures, we test whether changes in teacher quality at the school-by-grade level caused by staffing changes accurately predict changes in student test scores, as would be expected if our quality measures are unbiased (Chetty, Friedman and Rockoff, 2014a).[34] We implement the forecasting test by estimating the following regression model, weighted by time *t* school-by-grade enrollment:

$$\Delta\bar{A}_{sgkt} = \gamma_0 + \Delta\bar{\hat{V}}'_{sgkt}\gamma_1 + \mathbf{\Delta\bar{X}_{sgt}\gamma_2} + \phi_t + \varepsilon_{sgkt} \tag{B1}$$

The dependent variable, $\Delta\bar{A}_{sgkt}$, is the change in the average test score on the statewide exam (standardized by grade and year) between years *t* and *t-1* for school *s* and grade *g* in subject *k*. Only students taught by a teacher with an available effectiveness measure at time *t* are included in the regression and used to calculate $\Delta\bar{A}_{sgkt}$. In addition to year effects, the control set includes $\Delta\bar{\hat{V}}'_{sgkt}$, which is the change in average measured teacher quality, and $\mathbf{\Delta\bar{X}_{sgt}}$, which captures the change in student demographics between years *t* and *t-1*.

For the purposes of the validation exercise, we make adjustments to the way teacher quality is measured, which is why the variable is denoted with a prime in equation (B1). First, we rescale teachers' EVAAS scores to student exam score units. This permits one-to-one forecasting between the teacher quality metrics and the dependent variable. Then, we calculate leave-two-year-out jackknife estimates, where neither the time *t* nor the *t-1* teacher scores are included in $\mathbf{V_{ikt-}}$ (from equation 1 in the main text). This is important to remove the influence of the mechanical correlation between the change in average student test scores between those two periods and the estimation error in the annual teacher scores. We conduct the test for both purely backward-looking quality measures and, to increase precision, for measures that also allow post-period performance data to inform the current-year quality measure (as in Chetty, Friedman, and Rockoff, 2014a). Jackknifing based on pre- and post-period data is not a problem for this

[34] There is an ongoing debate between Chetty, Friedman, and Rockoff (2017) and Rothstein (2017) about the informational value of this test. Rothstein (2017) implements various parametric solutions to potential problems and concludes that the necessary assumptions do not hold. Chetty, Friedman, and Rockoff (2017) argue that his parametric approaches likely generate biases themselves, and that non-parametric tests do not indicate any problems with the methodology. They further note that even Rothstein's estimates of forecast bias range from just 5-15 percent across specifications, which still implies that these are meaningful measures of effectiveness.

exercise because internally valid estimates of forecasting bias can still be obtained even if survivors are oversampled.

We test the null hypothesis $\gamma_1 = 1$ separately for math and reading and report the results in Table B1. We cannot reject that the coefficient on the change in teacher quality is unity at the 5 percent level in any of the models and reject at the 10 percent level in only one case. Equally importantly, even in the case where unbiasedness is rejected, there is still substantial information contained in the value-added measures, as indicated by the large coefficient.

Table B1. Test for bias in jackknifed teacher quality measures

	$\hat{\gamma}_1$ (1)	P-value (*H₀: γ₁=1*) (2)	Number of school-grade-year cells (3)
Backward looking			
Grades 4-8, math	0.820 (0.092)	0.051	3,624
Grades 4-8, reading	0.814 (0.152)	0.221	3,627
Backward and forward looking			
Grades 4-8, math	0.904 (0.067)	0.152	3,973
Grades 4-8, reading	0.891 (0.116)	0.347	3,948

Notes: Coefficients and standard errors as estimated by equation B1 are reported in column 1. Column 2 reports p-values from tests of the null hypothesis of forecast-unbiasedness, and column 3 reports the number of school-by-grade-by-year cells used in the regressions. The backward-looking measures include only teacher scores from year *t*-2 and earlier. The backward- and forward-looking measures also include scores from year t+1 and later.

Appendix C. Statewide trends in teacher turnover

Our empirical strategy estimates overall and differential changes in turnover by teacher quality within HISD associated with the partial- and full-implementation periods, relative to the pre-period. There are challenges to interpreting both the overall and differential changes as causally attributable to the reform. First, the overall increase we observe in teacher turnover in HISD may be driven by secular changes over time. The period we study is one during which economic conditions were not stable. With unemployment rates peaking in September 2009, and then steadily declining after, our post-policy period overlaps with an economic recovery. Second, a maintained assumption for the estimates of impacts on differential turnover is that any differences across teacher quality groups would otherwise have been stable. There is the possibility, though, that sensitivity to economic conditions varies for more- and less-effective teachers.

In this appendix, we construct counterfactuals for HISD from a statewide panel of personnel and campus data we compiled. The personnel files are from the Texas Public Education Information Management System (PEIMS) and include all full-FTE teachers in traditional public schools that cover any grade from 3 to 8. In order to classify schools by achievement levels comparably to our analysis of HISD, we combine TAKS pass rate data for three springs (2008, 2009, and 2010). The pass rate is the average across math and reading for all grades. Consistent with our analysis of HISD data, we divide schools into three groups based on their placement in their respective districts' pass rate distributions: bottom quintile, middle quintiles, and top quintile. While it is straightforward to measure teacher turnover in other districts, we unfortunately do not have access to the type of restricted-use data that would allow us to measure teacher quality. Thus, we use teacher experience as our best available proxy.

We begin by exploring patterns for levels of teacher turnover, regardless of quality, over our sample period. In Figure C1, we show trends for the three measures we focus on in our analysis of HISD – school exits, district exits, and school transfers (within district) – across three samples of Texas school districts: (a) all districts other than HISD, (b) the five largest districts excluding HISD, and (c) districts adjacent to HISD. The state data also allow us to track exits from the Texas public schools, which we additionally include in the charts. Across all three samples, there is evidence of a U-shaped pattern in turnovers from before to after the recession, with continued gradual increases in the more recent years. Notably, though, turnover rates tend to return to levels in the initial year by 2013, before then continuing to rise somewhat beyond. This differs from the case for HISD shown in Figure 1, where rates jump up in 2011 and are well above the pre-policy baseline by 2013.

Figures C2a, C2b and C2c are comparable to Figure 3. Each figure shows turnover rates broken down by school achievement group (within district) for one of the three district samples. As in HISD, the U-shape is more marked for lower-achieving schools. However, once again, turnover rates are only slightly above initial levels by the end of the period whereas in HISD they far exceed them, with the most striking increase at low-achieving HISD schools.

In order to more formally compare overall and differential turnover patterns in HISD to the rest of the state, we estimate difference-in-difference-in-differences (DDD) style regression models evaluating changes over time in turnover by teacher experience levels in HISD relative to other districts in the state. Before discussing the DDD results, we first establish how experience relates to quality in HISD. Table C1 shows the shares of teachers with differing levels of experience that fall in each of our teacher quality bins. The middle (6 to 20 years of experience) and high (more than 20 years) groups exhibit similar patterns, but the low (5 or less years) group is substantially more likely to fall in the bottom-quintile and unknown-quality groups. The unknown-quality group is mechanically tied to experience, since this is the group for which current but not prior EVAAS scores are available. Table 7 in the main text shows this group appears to be comparable to bottom-quintile teachers in terms of effectiveness in raising student achievement. Thus, while having high levels of experience is not systematically related to our estimates of quality, inexperience can be viewed as a reasonable proxy for being likely to be of low quality. We still break results down by low- and high- relative to middle-experience levels for completeness.

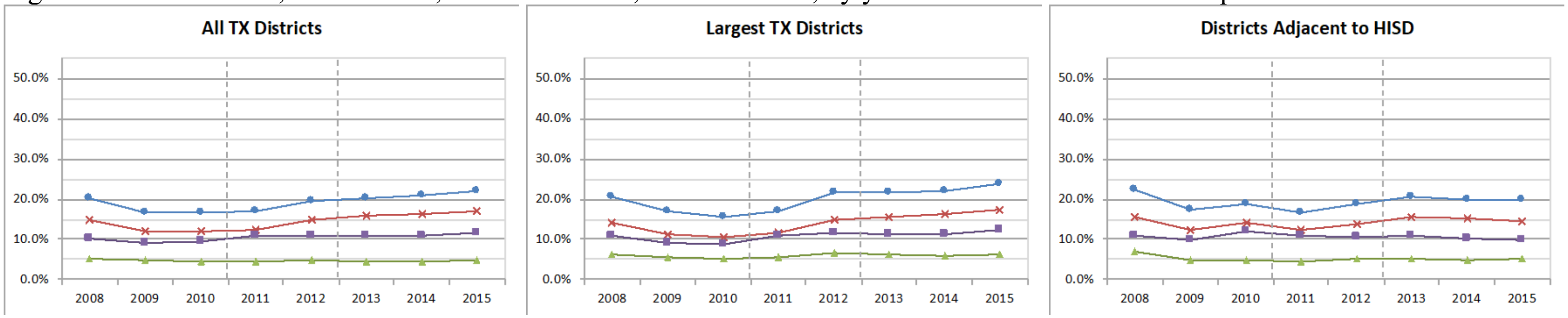
Figure C3 shows patterns of exit across years for teachers by experience level in HISD. The visually apparent relative increase for the least-experienced group is confirmed in the difference-in-differences style estimates shown in Table C2 (which repeats the first two columns of Table 6 in the main text with an added column to show school transfer outcomes). As noted in the text, the findings for the least-experienced group are consistent with the reform differentially increasing district exit for low-quality teachers.

The DDD results comparing HISD to the rest of the state are presented in Table C3 (which overlaps with and extends the results shown in the last two columns of Table 6).

Focusing on estimates for district exit in column 2, we find that turnover for those with middle experience was unchanged during the partial-implementation period and rose by 3.5 percentage points during the full-implementation period statewide. Relative increases in district exit rates for the least-experienced teachers were only slightly higher, by 0.6 and 0.5 percentage points in the partial- and full-implementation periods, respectively, while rates were higher by a few percentage points in both periods for the most-experienced. As noted in the text, for HISD relative to the rest of the state, the estimates are qualitatively similar to those shown in Table C2. Relative to the rest of the state, district exit for teachers with middle experience increased in HISD (by 4.8 percentage points in the full-implementation period), with effects magnified (by an additional 1.6 percentage points) for the least-experienced teachers.

Table C4 shows the sensitivity of the DDD results for district exit comparing HISD to different subsamples of Texas districts. We prefer the estimates based on all districts, since we know of at least one large district (Dallas) that had its own personnel reforms over the study period, and the labor markets of neighboring districts might be indirectly affected by happenings in HISD.

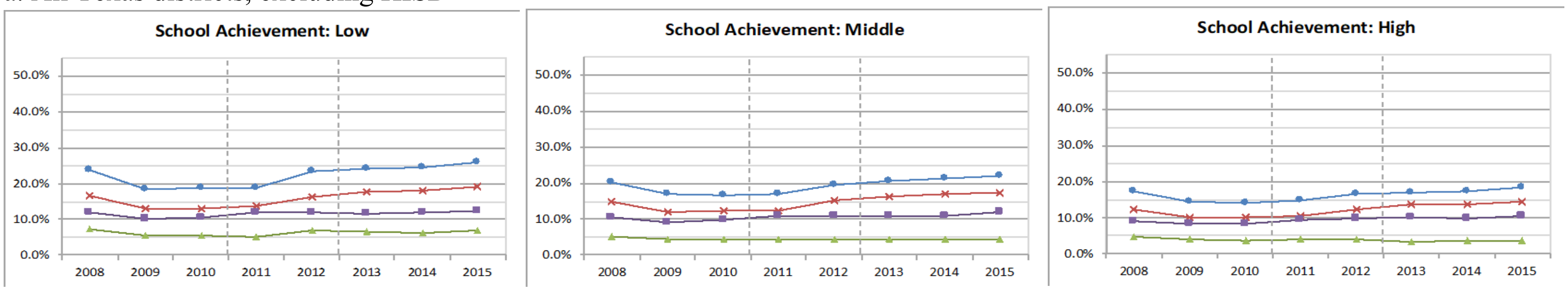
Figure C1. School exits, district exits, school transfers, and state exits, by year and Texas district subsample



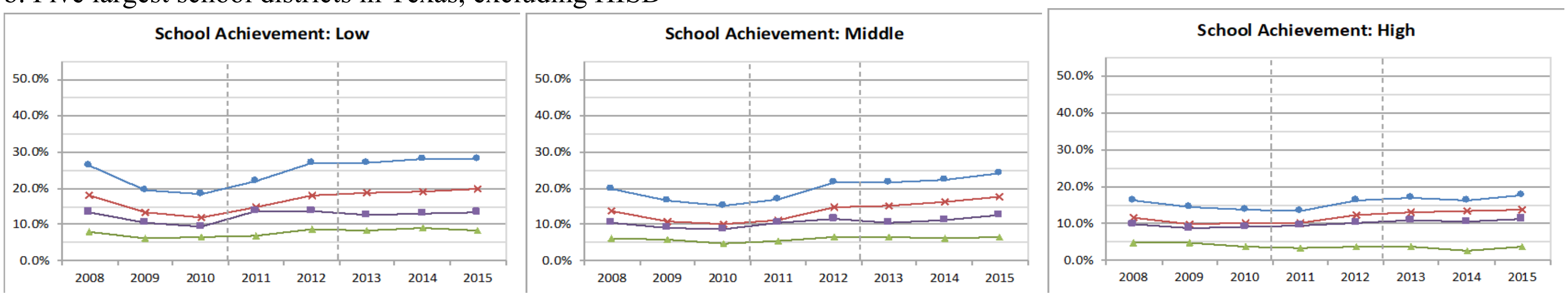
Notes: Circle markers: school exits; Cross markers: district exits; Triangle markers: school transfers (within district); Square markers: state exits. School exits are the sum of district exits and school transfers, and state exits are a subset of district exits. From left to right, the charts show turnover rates at all Texas school districts other than HISD, at the five largest school districts excluding HISD (Dallas, Cypress-Fairbanks, Northside, Austin, and Fort Worth), and at districts adjacent to HISD (Aldine, Alief, Alvin, Cypress-Fairbanks, Fort Bend, Galena Park, Humble, Katy, Pasadena, Pearland, Sheldon, Spring Branch, and Stafford). The turnover rates are teacher-weighted so they can be interpreted as the likelihood of exit for the typical teacher in each sample of districts.

Figure C2. School exits, district exits, school transfers, and state exits, by year and pre-policy achievement group of school

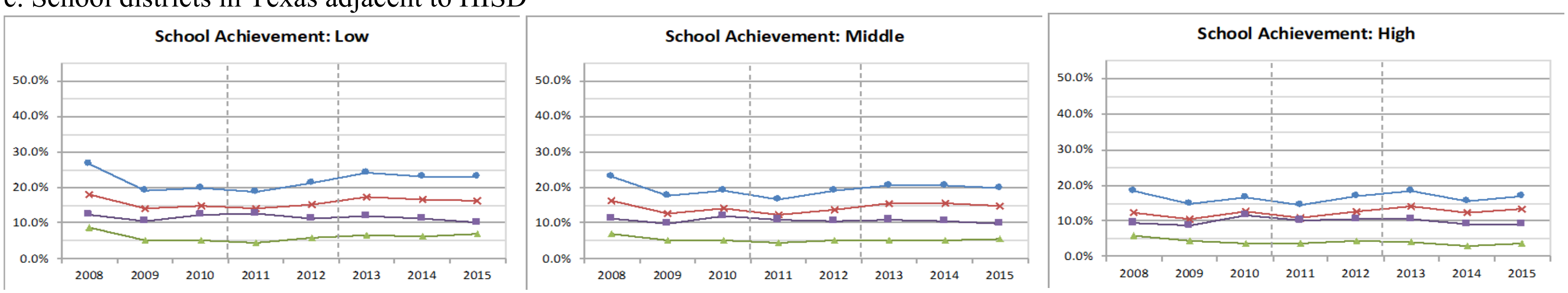
a. All Texas districts, excluding HISD



b. Five largest school districts in Texas, excluding HISD

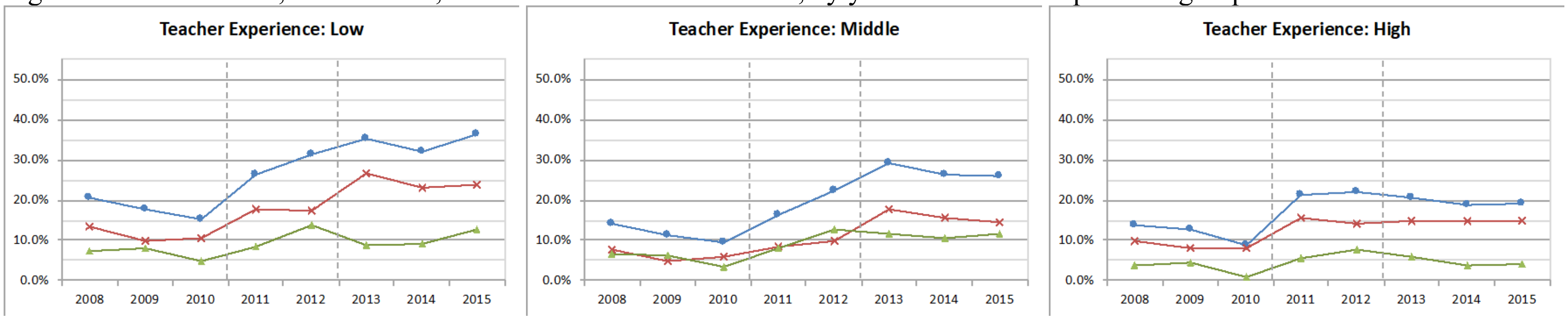


c. School districts in Texas adjacent to HISD



Notes: Circle markers: school exits; Cross markers: district exits; Triangle markers: school transfers (within district); Square markers: state exits. From left to right, the graphs show turnovers at bottom-quintile, middle-quintiles (2-3-4), and top-quintile schools according to pre-policy achievement levels. Panel a includes schools in all Texas districts other than HISD; Panel b includes the five largest school districts excluding HISD; Panel c includes districts adjacent to HISD. For other details, see the notes to Figure C1.

Figure C3. School exits, district exits, and school transfers for HISD, by year and teacher experience group



Notes: Circle markers: school exits; Cross markers: district exits; Triangle markers: school transfers. From left to right, the charts show turnover rates for HISD teachers with low (0-5), middle (6-20), and high (more than 20) years of experience. For other details, see the notes to Figure 1.

Table C1. Distribution of teacher quality by experience levels

		Quintile Bin			
		Unknown	Bottom	Middle	Top
Experience Level	Exp≤5	0.33	0.12	0.43	0.11
	5<Exp≤20	0.14	0.16	0.54	0.17
	Exp>20	0.11	0.17	0.56	0.17

Notes: The cells show the shares of our analysis teachers within each experience level that fall in each of the teacher quality bins. The shares sum to one across rows.

Table C2. Impacts on turnover, by teacher experience group and evaluation period, at HISD

	School exit	District exit	School transfer
	(1)	(2)	(3)
Intercept	0.131	0.082	0.049
	(0.006)**	(0.004)**	(0.003)**
Intercept*PARTIAL	0.103	0.049	0.053
	(0.010)**	(0.007)**	(0.007)**
Intercept*FULL	0.164	0.109	0.055
	(0.011)**	(0.008)**	(0.006)**
Exp≤5	0.041	0.036	0.004
	(0.011)**	(0.008)**	(0.007)
Exp≤5*PARTIAL	0.032	0.034	-0.002
	(0.017)*	(0.014)**	(0.012)
Exp≤5*FULL	0.007	0.026	-0.019
	(0.017)	(0.014)*	(0.012)
Exp>20	0.001	0.024	-0.023
	(0.012)	(0.009)**	(0.007)**
Exp>20*PARTIAL	0.019	0.031	-0.013
	(0.020)	(0.018)*	(0.012)
Exp>20*FULL	-0.065	-0.031	-0.034
	(0.018)**	(0.016)*	(0.012)**
R-squared	0.082	0.057	0.056
N (Teacher-year)	18,681	18,681	18,681

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. The specifications are the same other than that the indicators for teacher quality quintiles are replaced by indicators for levels of experience. The omitted experience group includes teachers with 6 to 20 years of experience.

Table C3. Impacts on turnover, by teacher experience level, HISD compared to the rest of Texas

	School exit (1)	District exit (2)	School transfer (3)	State exit (4)
Intercept	0.152	0.112	0.040	0.087
	(0.001)**	(0.001)**	(0.001)**	(0.006)**
Intercept*PARTIAL	0.004	0.000	0.004	0.003
	(0.001)**	(0.001)	(0.001)**	(0.001)**
Intercept*FULL	0.043	0.035	0.008	0.011
	(0.001)**	(0.001)**	(0.001)**	(0.001)**
Exp≤5	0.045	0.037	0.008	0.015
	(0.001)**	(0.001)**	(0.001)**	(0.001)**
(Exp≤5)*PARTIAL	0.009	0.006	0.003	0.009
	(0.002)**	(0.002)**	(0.001)**	(0.001)**
(Exp≤5)*FULL	0.000	0.005	-0.005	-0.004
	(0.002)	(0.002)**	(0.001)**	(0.001)**
Exp>20	-0.005	0.009	-0.014	0.028
	(0.002)**	(0.002)**	(0.001)**	(0.002)**
(Exp>20)*PARTIAL	0.040	0.041	0.000	0.037
	(0.003)**	(0.002)**	(0.001)	(0.002)**
(Exp>20)*FULL	0.012	0.016	-0.003	0.029
	(0.002)**	(0.002)**	(0.001)**	(0.002)**
HISD*PARTIAL	0.060	0.030	0.030	0.020
	(0.008)**	(0.005)**	(0.006)**	(0.004)**
HISD*FULL	0.076	0.048	0.027	0.032
	(0.008)**	(0.005)**	(0.005)**	(0.004)**
HISD*(Exp≤5)	0.004	0.012	-0.008	0.021
	(0.006)	(0.005)**	(0.004)**	(0.004)**
HISD*(Exp≤5)*PARTIAL	0.027	0.027	0.000	0.016
	(0.010)**	(0.008)**	(0.006)	(0.007)**
HISD*(Exp≤5)*FULL	0.018	0.016	0.002	0.000
	(0.009)*	(0.008)**	(0.005)	(0.007)
HISD*(Exp>20)	0.008	0.012	-0.004	0.003
	(0.009)	(0.008)	(0.005)	(0.008)
HISD*(Exp>20)*PARTIAL	0.015	0.022	-0.007	0.034
	(0.013)	(0.012)*	(0.007)	(0.012)**
HISD*(Exp>20)*FULL	-0.048	-0.033	-0.015	-0.018
	(0.012)**	(0.011)**	(0.006)**	(0.011)*
R-squared	0.010	0.011	0.002	0.009
N (Teacher-year)	1,716,949	1,716,949	1,716,949	1,716,949

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. The differences are that statewide data are used in the regression, teacher quality is proxied by teacher experience level, and the primary analytic variables in the model (teacher quality and policy implementation period) are additionally interacted with an indicator variable for HISD. The omitted experience group includes teachers with 6 to 20 years of experience. Note that there is no main effect for HISD since this is subsumed by the school fixed effects.

Table C4. Impacts on district exit, by teacher experience level, HISD compared to other districts

	Sample of non-HISD TX school districts		
	All	Five largest	Adjacent to HISD
	(1)	(2)	(3)
Intercept	0.112	0.115	0.130
	(0.001)**	(0.002)**	(0.002)**
Intercept*PARTIAL	0.000	0.012	-0.015
	(0.001)	(0.003)**	(0.003)**
Intercept*FULL	0.035	0.040	0.014
	(0.001)**	(0.003)**	(0.003)**
Exp≤5	0.037	0.031	0.020
	(0.001)**	(0.003)**	(0.003)**
(Exp≤5)*PARTIAL	0.006	-0.005	0.011
	(0.002)**	(0.004)	(0.005)**
(Exp≤5)*FULL	0.005	0.016	0.000
	(0.002)**	(0.004)**	(0.004)
Exp>20	0.009	0.013	0.015
	(0.002)**	(0.005)**	(0.006)**
(Exp>20)*PARTIAL	0.041	0.043	0.019
	(0.002)**	(0.008)**	(0.009)**
(Exp>20)*FULL	0.016	0.010	-0.006
	(0.002)**	(0.007)	(0.008)
HISD*PARTIAL	0.030	0.019	0.046
	(0.005)**	(0.006)**	(0.006)**
HISD*FULL	0.048	0.044	0.070
	(0.005)**	(0.006)**	(0.006)**
HISD*(Exp≤5)	0.012	0.017	0.029
	(0.005)**	(0.006)**	(0.006)**
HISD*(Exp≤5)*PARTIAL	0.027	0.038	0.022
	(0.008)**	(0.009)**	(0.009)**
HISD*(Exp≤5)*FULL	0.016	0.003	0.020
	(0.008)**	(0.009)	(0.009)**
HISD*(Exp>20)	0.012	0.009	0.005
	(0.008)	(0.009)	(0.010)
HISD*(Exp>20)*PARTIAL	0.022	0.019	0.043
	(0.012)*	(0.014)	(0.015)**
HISD*(Exp>20)*FULL	-0.033	-0.029	-0.012
	(0.011)**	(0.013)**	(0.013)
R-squared	0.011	0.014	0.010
N (Teacher-year)	1,716,949	262,972	257,509

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. The estimates in the first column are from Table C3 column 2, while columns 2 and 3 restrict the sample of comparison districts as indicated in the column heading.

Appendix D. Technical details of the simulation

This appendix describes the technical details of the simulation framework used to generate estimates of the 10-year change in overall workforce quality presented in column 2 of Table 8.

We start by defining a vector of pre-policy teacher quality sets, $T_0 = (T_{0,top}, T_{0,mid}, T_{0,bot}, T_{0,all})$. The first three elements in this vector represent the sets of teachers within the top quintile, three middle quintiles, and bottom quintile, respectively. The final element is the set of all teachers, i.e. the union of the first three sets.[35] The distribution of $T_{0,all}$ is specified as normal with the mean and standard deviation taken from the pre-policy empirical teacher quality distribution in HISD, measured in student exam score standard deviations. We then define a probability vector, $p = (p_{top}, p_{mid}, p_{bot}, p_{all})$, the elements of which are post-policy full-implementation district exit probabilities of teachers in the corresponding set, taken as estimated in Table 2. Note that $p_{all} = 0.2 \times p_{top} + 0.6 \times p_{mid} + 0.2 \times p_{bottom}$.

Next, we define T_i to be the i_{th} iteration of the teacher quality sets during the simulation, where $i = 1$ to 10. T_i is constructed from T_{i-1} by drawing $(1 - p_{top})$, $(1 - p_{mid})$, and $(1 - p_{bot})$ shares of teachers from $T_{i-1,top}$, $T_{i-1,mid}$, and $T_{i-1,bot}$, respectively. Then, the same number of replacement teachers are drawn for exiters from the pre-policy teacher quality distribution by drawing the share p_{all} of teachers from $T_{0,all}$. The effect on average workforce quality after 10 years is calculated as $mean(T_{10,all}) - mean(T_{0,all})$.

In rows 1 to 4 of Table 8, the basic simulation framework remains the same, but we alter the exit probability vectors to coincide with the given scenario. The probability vectors used for each row are presented in Table D1. In the baseline scenario, the probabilities are taken directly from column 2 of Table 2. For example, p_{mid} is calculated by adding the coefficient on *Intercept*FULL* (0.108) to the intercept value (0.084), and p_{bot} further adds the coefficient of *Bottom quintile*FULL* (0.062) to the p_{mid} value. For row 2, all of the full-implementation effects are halved (*Intercept*FULL*, *Bottom quintile*FULL*, *Top quintile*FULL*) when calculating the exit probabilities for each group, while in rows 3 and 4 only the level effect (*Intercept*FULL*) is modified.

[35] We ignore teachers of unknown quality in the simulation framework as noted in the main text because there is little evidence they are affected by the policy and it is not straightforward to build them into the simulations.

In row 5 of Table 8, we explore the impact of concentrating the policy effects further in the tails of the teacher quality distribution by redefining the T vector such that T_{top} is the top decile of the teacher quality distribution, T_{mid} represents the eight middle deciles, and T_{bot} is the bottom decile. To account for the full effect on differential attrition with this concentration, we also double the post-period policy effects in the tails (*Bottom quintile*FULL*, *Top quintile*FULL*) when calculating the exit probabilities for each group. The revised probability vector for this scenario is presented in row 5 of Table D1.

In row 6 of Table 8, the simulation is carried out using the baseline exit probability vector (row 1 of Table D1). However, motivated by Rothstein (2015), replacement teachers are drawn from $T'_{0,all}$, where the distribution is shifted downward so $mean(T'_{0,all}) = mean(T_{0,all}) - 0.25 * sd(T_{0,all})$.

Rows 7-9 apply the teacher exit probability vectors from rows 1, 3, and 4 of Table D1, respectively, and additionally incorporate a one-year disruption effect (Table 7) for each replacement teacher. Specifically, a teacher's quality in the first year is reduced by 0.055 but then returns to its full value from the second year onward. Exit probabilities for all years, even in the first year, are based on true teacher quality (i.e., teachers' disruption effects do not influence their retention likelihoods).[36]

Table D1. Exit probability vectors for the different simulation scenarios

	p_{top}	p_{mid}	p_{bot}	p_{all}
Row 1	0.152	0.192	0.254	0.196
Row 2	0.118	0.138	0.169	0.140
Row 3	0.098	0.138	0.200	0.142
Row 4	0.071	0.111	0.173	0.115
Row 5	0.112	0.192	0.316	0.196

Notes: For row 5, the "top" and "bottom" bins are deciles, not quintiles, as described in the appendix text.

[36] This is equivalent to applying the relevant baseline simulation for years 1 to 9 and then drawing the year 10 replacement teachers from $T''_{0,all}$ where $mean(T''_{0,all}) = mean(T_{0,all}) - 0.055$. The simulation is operationalized using this equivalent method.

Appendix E. Supplementary tables

Table E1 shows how the EVAAS scores of teachers who are new to teaching math and reading in a tested grade in HISD compare to others teaching those subjects in the same years. Since EVAAS scores are normed on an annual basis, any patterns should be interpreted as reflecting relative and not absolute performance on this metric. Column 1 shows results for the full sample of teachers with first-time EVAAS scores, while column 2 restricts the sample to those new to the district and not just new to the subject and grade. Notably, entrants in the base year of 2008 were negatively selected, with EVAAS scores -0.15 and -0.33 standard deviations below the average teacher in HISD. Since one standard deviation in the teacher distribution maps to approximately one-tenth of a standard deviation in the student test score distribution, these are equivalent to about -0.02 and -0.03 student standard deviations. In column 1, there is no evidence that the performance of teachers new to tested grades and subjects is systematically better or worse in the post-reform years. However, in column 2 those new to the district tend to perform better than in 2008 in the full-implementation period, and there was also a rebound in 2009 and 2010 that predates the reform. Thus, there is no clear evidence that the reform has attracted relatively more- or less-effective teachers than the existing stock, although it is important to keep in mind that the existing stock may be improving.

Table E2 replicates Table 2 but adds the schools where Fryer (2014) intervened back into the sample. The results are similar to what we report in Table 2, which is expected because the Fryer schools make up a small fraction of district schools.

Table E3 replicates Table 2 but assigns the teacher quality quintiles using jackknifed measures based only on a single-lagged EVAAS score rather than all available lagged EVAAS scores. Note that teachers who move in and out of tested grades and subjects are more likely to show up in the "unknown quality" bin in these models.

Tables E4 and E5 replicate our main findings from Table 2, which pools math and reading teachers, but present estimates separately by subject. Columns 1a, 2a, and 3a match the results in Table 2 structurally; columns 1b, 2b, and 3b show results from a model where mathematics (or reading) teacher quality is entered into the regression as a continuous, linear variable. Teachers in the "Unknown" bin are necessarily omitted from the linear models, which is reflected in the sample sizes reported at the bottom of the table.

Table E6 adds a full set of interactions with the achievement quintile of the school (bottom and top, with the middle 2-4 quintiles as the omitted benchmark category), which matches the structure of Figure 4 in the main text. The estimates are noisy but confirm the key themes illustrated in the figure. Most notably, column 3 shows that top-quintile teachers at bottom-quintile schools were significantly more likely to transfer schools within HISD in the full implementation period, and less likely to exit the district. Also, the bottom rows of the first two columns show that top quintile schools did not experience the increased targeting of district exits toward bottom quintile teachers as was the case in other schools.

Table E1. EVAAS scores of new teachers by year, controlling for teacher characteristics and school fixed effects

	First year with EVAAS score (1)	First year in HISD (2)
Intercept	-0.152	-0.325
	(0.043)**	(0.067)**
Intercept*2009	0.034	0.202
	(0.064)	(0.099)**
Intercept*2010	0.030	0.229
	(0.064)	(0.091)**
Intercept*2011	-0.129	-0.039
	(0.074)*	(0.107)
Intercept*2012	0.000	0.043
	(0.069)	(0.111)
Intercept*2013	-0.028	0.433
	(0.067)	(0.219)**
Intercept*2014	-0.111	0.101
	(0.080)	(0.124)
Intercept*2015	0.016	0.204
	(0.071)	(0.109)*
R-squared	0.122	0.229
N (Teacher-year)	3,848	1,528

Notes: The estimates in this table are taken from regressions of single-year EVAAS scores on school and year fixed effects and the same vector of teacher characteristics used in the teacher exit regressions and reported in Table 1. Standard errors are clustered by school. The single-year EVAAS scores used in the regressions are the average of a teacher's math and reading scores when both subjects are available in the given year and are subject-specific values otherwise. The sample used to estimate the results shown in column 1 consists of all teachers who are new to a tested grade and subject in the given year (i.e., those for whom it is their first year with an EVAAS score). The estimation sample for column 2 consists only of those teachers who are new to HISD in the given year.

Table E2. Impacts on turnover, by teacher quality group, including Fryer schools

	School exit	District exit	School transfer
	(1)	(2)	(3)
Intercept	0.137	0.084	0.052
	(0.006)**	(0.004)**	(0.004)**
Intercept*PARTIAL	0.109	0.055	0.054
	(0.011)**	(0.007)**	(0.007)**
Intercept*FULL	0.171	0.110	0.061
	(0.012)**	(0.009)**	(0.007)**
Bottom quintile	-0.004	0.014	-0.019
	(0.011)	(0.010)	(0.007)**
Bottom quintile*PARTIAL	0.074	0.084	-0.010
	(0.019)**	(0.019)**	(0.013)
Bottom quintile*FULL	0.073	0.057	0.017
	(0.021)**	(0.020)**	(0.014)
Top quintile	0.000	-0.009	0.009
	(0.013)	(0.009)	(0.009)
Top quintile*PARTIAL	0.003	-0.014	0.018
	(0.023)	(0.014)	(0.016)
Top quintile*FULL	-0.030	-0.045	0.016
	(0.020)	(0.014)**	(0.014)
Unknown	-0.014	-0.012	-0.002
	(0.011)	(0.010)	(0.007)
Unknown*PARTIAL	0.013	0.006	0.007
	(0.019)	(0.015)	(0.013)
Unknown*FULL	0.001	0.008	-0.007
	(0.018)	(0.015)	(0.012)
R-squared	0.092	0.067	0.062
N (Teacher-year)	20,128	20,128	20,128

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. The only difference is that the sample here includes the schools affected by the Fryer (2014) intervention.

Table E3. Impacts on turnover, by teacher quality group and evaluation period, single-lag jackknife

	School exit	District exit	School transfer
	(1)	(2)	(3)
Intercept	0.135	0.084	0.051
	(0.006)**	(0.005)**	(0.004)**
Intercept*PARTIAL	0.101	0.051	0.050
	(0.011)**	(0.007)**	(0.008)**
Intercept*FULL	0.158	0.102	0.055
	(0.013)**	(0.009)**	(0.007)**
Bottom quintile	-0.008	0.013	-0.021
	(0.011)	(0.010)	(0.007)**
Bottom quintile*PARTIAL	0.033	0.035	-0.002
	(0.020)	(0.020)*	(0.014)
Bottom quintile*FULL	0.072	0.059	0.013
	(0.022)**	(0.021)**	(0.014)
Top quintile	-0.001	-0.006	0.005
	(0.013)	(0.009)	(0.009)
Top quintile*PARTIAL	-0.022	-0.029	0.007
	(0.023)	(0.016)*	(0.016)
Top quintile*FULL	-0.030	-0.030	-0.001
	(0.019)	(0.016)*	(0.013)
Unknown	-0.022	-0.013	-0.009
	(0.010)**	(0.009)	(0.007)
Unknown*PARTIAL	0.006	-0.004	0.010
	(0.018)	(0.013)	(0.012)
Unknown*FULL	0.035	0.033	0.002
	(0.017)**	(0.014)**	(0.011)
R-squared	0.086	0.065	0.058
N (Teacher-year)	18,681	18,681	18,681

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. The only difference is that the jackknifed measures used to construct the teacher quality quintiles are estimated based only on a single-lagged EVAAS score, rather than all available lagged EVAAS scores. Teachers who move in and out of tested grades and subjects are more likely to show up in the “unknown quality” bin in these models.

Table E4. Impacts on turnover, by mathematics teacher quality group and linear mathematics teacher quality

	School exit		District exit		School transfer	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Intercept	0.127	0.118	0.078	0.072	0.049	0.045
	(0.007)**	(0.007)**	(0.004)**	(0.005)**	(0.004)**	(0.004)**
Intercept*PARTIAL	0.105	0.112	0.043	0.053	0.061	0.059
	(0.013)**	(0.013)**	(0.008)**	(0.008)**	(0.009)**	(0.009)**
Intercept*FULL	0.159	0.173	0.104	0.121	0.055	0.052
	(0.013)**	(0.014)**	(0.010)**	(0.010)**	(0.008)**	(0.008)**
Linear quality		-0.008		-0.031		0.024
		(0.015)		(0.011)**		(0.009)**
Linear quality*PARTIAL		-0.035		-0.047		0.012
		(0.024)		(0.019)**		(0.019)
Linear quality*FULL		-0.068		-0.081		0.013
		(0.024)**		(0.020)**		(0.015)
Bottom quintile	0.007		0.020		-0.013	
	(0.014)		(0.012)*		(0.009)	
Bottom quintile*PARTIAL	0.056		0.078		-0.022	
	(0.024)**		(0.022)**		(0.017)	
Bottom quintile*FULL	0.073		0.063		0.009	
	(0.026)**		(0.024)**		(0.014)	
Top quintile	-0.002		-0.012		0.010	
	(0.015)		(0.010)		(0.010)	
Top quintile*PARTIAL	0.006		0.005		0.001	
	(0.024)		(0.017)		(0.018)	
Top quintile*FULL	-0.004		-0.032		0.028	
	(0.027)		(0.019)*		(0.018)	
Unknown	-0.006		-0.006		0.001	
	(0.013)		(0.011)		(0.008)	
Unknown*PARTIAL	-0.011		0.011		-0.021	
	(0.021)		(0.018)		(0.015)	
Unknown*FULL	0.002		0.003		-0.001	
	(0.022)		(0.019)		(0.015)	
R-squared	0.091	0.102	0.074	0.085	0.067	0.074
N (Teacher-year)	12,418	9,787	12,418	9,787	12,418	9,787

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. For the models where teacher quality enters linearly, teachers of unknown quality are necessarily omitted.

Table E5. Impacts on turnover, by reading teacher quality group and linear reading teacher quality

	School exit		District exit		School transfer	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Intercept	0.124	0.113	0.079	0.070	0.088	0.043
	(0.007)**	(0.007)**	(0.005)**	(0.005)**	(0.009)**	(0.004)**
Intercept*PARTIAL	0.100	0.104	0.048	0.053	0.052	0.052
	(0.012)**	(0.012)**	(0.008)**	(0.008)**	(0.009)**	(0.009)**
Intercept*FULL	0.172	0.181	0.114	0.122	0.058	0.058
	(0.014)**	(0.014)**	(0.010)**	(0.011)**	(0.008)**	(0.008)**
Linear quality		0.000		-0.012		0.012
		(0.017)		(0.013)		(0.012)
Linear quality*PARTIAL		-0.055		-0.060		0.006
		(0.031)*		(0.026)**		(0.019)
Linear quality*FULL		-0.101		-0.094		-0.007
		(0.031)**		(0.026)**		(0.022)
Bottom quintile	-0.016		-0.007		-0.009	
	(0.015)		(0.012)		(0.009)	
Bottom quintile*PARTIAL	0.036		0.053		-0.017	
	(0.026)		(0.023)**		(0.016)	
Bottom quintile*FULL	0.055		0.048		0.007	
	(0.027)**		(0.022)**		(0.020)	
Top quintile	0.001		-0.014		0.015	
	(0.015)		(0.010)		(0.010)	
Top quintile*PARTIAL	-0.022		-0.018		-0.004	
	(0.026)		(0.018)		(0.017)	
Top quintile*FULL	-0.055		-0.051		-0.004	
	(0.024)**		(0.018)**		(0.016)	
Unknown	-0.016		-0.019		0.002	
	(0.013)		(0.011)		(0.009)	
Unknown*PARTIAL	0.027		0.024		0.003	
	(0.025)		(0.020)		(0.016)	
Unknown*FULL	0.020		0.018		0.002	
	(0.022)		(0.019)		(0.015)	
R-squared	0.093	0.102	0.073	0.084	0.066	0.075
N (Teacher-year)	12,739	9,879	12,739	9,879	12,739	9,879

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. For the models where teacher quality enters linearly, teachers of unknown quality are necessarily omitted.

Table E6. Impacts on turnover across school quality quintiles, by teacher quality quintile and evaluation period

	School exit	District exit	School transfer
	(1)	(2)	(3)
Intercept	0.119 (0.009)**	0.102 (0.006)**	0.017 (0.005)**
Bottom-quintile teacher	-0.022 (0.014)	0.001 (0.012)	-0.023 (0.008)**
Top-quintile teacher	-0.001 (0.018)	-0.024 (0.012)**	0.023 (0.012)*
Unknown teacher quality	-0.037 (0.013)**	-0.025 (0.011)**	-0.012 (0.008)
Bottom-quintile school	-0.033 (0.023)	-0.014 (0.016)	-0.020 (0.015)
Bottom-quintile school*Bottom-quintile teacher	0.029 (0.035)	0.021 (0.029)	0.008 (0.022)
Bottom-quintile school*Top-quintile teacher	-0.033 (0.039)	0.026 (0.033)	-0.060 (0.027)**
Bottom-quintile school*Unknown teacher	0.057 (0.039)	0.029 (0.028)	0.028 (0.022)
Top-quintile school	0.063 (0.014)**	-0.077 (0.011)**	0.140 (0.009)**
Top-quintile school*Bottom-quintile teacher	0.072 (0.026)**	0.046 (0.024)*	0.026 (0.017)
Top-quintile school*Top-quintile teacher	0.011 (0.028)	0.029 (0.018)*	-0.018 (0.018)
Top-quintile school*Unknown teacher	0.047 (0.022)**	0.032 (0.021)	0.015 (0.015)
Intercept*PARTIAL	0.102 (0.014)**	0.041 (0.009)**	0.060 (0.010)**
Intercept*FULL	0.166 (0.016)**	0.103 (0.011)**	0.063 (0.009)**
Bottom-quintile teacher *PARTIAL	0.087 (0.026)**	0.085 (0.024)**	0.002 (0.017)
Bottom-quintile teacher *FULL	0.128 (0.027)**	0.090 (0.027)**	0.038 (0.017)**
Top-quintile teacher*PARTIAL	0.006 (0.029)	-0.011 (0.018)	0.018 (0.021)
Top-quintile teacher*FULL	-0.033 (0.027)	-0.023 (0.019)	-0.010 (0.017)
Unknown teacher*PARTIAL	0.036 (0.026)	0.020 (0.020)	0.016 (0.017)
Unknown teacher*FULL	0.046 (0.023)**	0.035 (0.019)*	0.011 (0.015)
Bottom-quintile school*PARTIAL	0.090 (0.043)**	0.051 (0.022)**	0.039 (0.032)
Bottom-quintile school*FULL	0.134 (0.046)**	0.121 (0.030)**	0.013 (0.032)
Bottom-quintile school*Bottom-quintile teacher*PARTIAL	-0.032 (0.063)	0.064 (0.072)	-0.096 (0.037)**
Bottom-quintile school*Bottom-quintile teacher*FULL	-0.147 (0.071)**	-0.083 (0.053)	-0.064 (0.049)
Bottom-quintile school*Top-quintile teacher*PARTIAL	0.015 (0.085)	-0.007 (0.056)	0.021 (0.066)
Bottom-quintile school*Top-quintile Teacher*FULL	0.035 (0.096)	-0.114 (0.058)*	0.149 (0.067)**
Bottom-quintile school*Unknown teacher*PARTIAL	-0.118 (0.059)**	-0.054 (0.043)	-0.063 (0.041)
Bottom-quintile school*Unknown teacher*FULL	-0.143 (0.061)**	-0.084 (0.046)*	-0.059 (0.037)
Top-quintile school*PARTIAL	-0.035 (0.022)	0.005 (0.016)	-0.041 (0.015)**
Top-quintile school*FULL	-0.047 (0.026)*	-0.012 (0.022)	-0.034 (0.015)**
Top-quintile school*Bottom-quintile teacher*PARTIAL	-0.086 (0.043)**	-0.086 (0.042)**	0.000 (0.0310
Top-quintile school*Bottom-quintile teacher*FULL	-0.159 (0.050)**	-0.097 (0.050)*	-0.062 (0.031)**
Top-quintile school*Top-quintile teacher*PARTIAL	-0.006 (0.050)	0.013 (0.032)	-0.018 (0.031)
Top-quintile school*Top-quintile Teacher*FULL	0.011 (0.041)	-0.007 (0.030)	0.019 (0.027)
Top-quintile school*Unknown teacher*PARTIAL	-0.053 (0.043)	-0.020 (0.039)	-0.033 (0.026)
Top-quintile school*Unknown teacher*FULL	-0.059 (0.039)	-0.051 (0.037)	-0.008 (0.025)
R-squared	0.091	0.070	0.061
N (Teacher year)	18,681	18,681	18,681

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table2 apply. The specifications are the same other than that interactions with top- and bottom-quintile schools have been added as shown.