NBER WORKING PAPER SERIES

THE COMPOSITIONAL EFFECT OF RIGOROUS TEACHER EVALUATION ON WORKFORCE QUALITY

Julie Berry Cullen
Cory Koedel
Eric Parsons

Working Paper 22805
http://www.nber.org/papers/w22805

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2016, Revised July 2017

Cullen is in the Department of Economics at the University of California, San Diego. Koedel is in the Department of Economics and Truman School of Public Affairs, and Parsons is in the Department of Economics, at the University of Missouri, Columbia. 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, the funders, or the National Bureau of Economic Research. Any and all errors are attributable to the authors.

The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality
Julie Berry Cullen, Cory Koedel, and Eric Parsons
NBER Working Paper No. 22805
November 2016, Revised July 2017
JEL No. H75,I28,J45

ABSTRACT

Improving public sector workforce quality is challenging in sectors such as education where worker productivity is difficult to assess and manager incentives are muted by political and bureaucratic constraints. In this paper, we study how providing improved information to principals about teacher effectiveness and encouraging them to use the information in personnel decisions affects the composition of teacher turnovers. Our setting is the Houston Independent School District, which recently implemented a rigorous teacher evaluation system. Prior to the new system, teacher effectiveness was negatively correlated with district exit and we show that the policy significantly strengthened this relationship, primarily by increasing the relative likelihood of exit for teachers in the bottom quintile of the quality distribution. Low-performing teachers working in low-achieving schools were especially likely to leave. However, despite the success, the implied change to the quality of the workforce overall is too small to have a detectable impact on student achievement.

Julie Berry Cullen
Department of Economics - 0508
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
and NBER
jbcullen@ucsd.edu

Eric Parsons
Dept. of Economics
University of Missouri
231 Professional Building
Columbia, Missouri 65211
parsonses@missouri.edu

Cory Koedel
Dept. of Economics
University of Missouri
118 Professional Building
Columbia, MO 65211
koedelc@missouri.edu

1. Introduction

Government agencies that provide services, such as education and health, are settings where it is difficult to observe both inputs and outputs. These are also sectors where there are ongoing concerns about efficiency and equity. In elementary and secondary education, efforts to improve the effectiveness of schools have ranged from increases in resources via school finance reforms, to increased competition via school choice, to performance standards via school accountability. 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 (Chetty, Friedman, and Rockoff, 2014a/b; Hanushek, 2011; Hanushek and Rivkin, 2010; Jackson, forthcoming). A challenge facing school administrators in managing the teacher workforce is that teacher effectiveness is not easy to measure and is not strongly correlated with observable characteristics. In this type of setting, improved information about quality can lead to more productive personnel policies. Given the two-sided nature of matches, better information may also have equity implications because low-achieving schools struggle to attract and retain good teachers (Bates, 2016; Clotfelter et al., 2006).

In this paper, we study the impact of introducing a rigorous teacher evaluation system on the level and distribution of teacher quality. The context for our study is the Houston Independent School District (HISD), which is the seventh largest school district in the United States. The new evaluation system was phased in from 2011 to 2013 and is centered on a standardized method for annually evaluating teachers. The objective is to generate comprehensive teacher performance measures and empower principals to exit ineffective and retain effective teachers at higher rates, as well as improve ongoing skill development.

Recognizing that teacher hiring and development also play a role in overall policy efficacy, we focus on how the policy impacted patterns of attrition by teacher effectiveness.

The empirical analyses rely on administrative data tracking teachers for three years before to three years after the reform. For the subset of teachers in tested grades and subjects we begin by classifying teachers by quality using proxies we construct and validate for value added to student achievement. Then, using difference-in-differences and event-study analyses, we show that the relationships between teacher quality and both school and district exit became more negative in the post-policy period. The key driver is an increase in the relative likelihood of exit from the district of teachers in the bottom quintile of the quality distribution, concentrated in low-achieving schools.

As far as impacts on student achievement through the turnover channel, there are two important issues to consider. First, overall turnover increased after the reform, though a portion of this level shift is likely attributable to the economic recovery as turnover returned to pre-recession levels statewide. Second, the reform had only moderate impacts on quality per turnover and the associated improvements in workforce quality are too small to have a detectable impact on student achievement. We demonstrate this point in illustrative models that relate observed school-by-grade teacher exits to student achievement gains and simulations that cumulate the impacts over time.

Our study contributes in a number of ways to the few existing studies of policies that are designed to improve workforce quality by providing better information on teacher effectiveness. In contrast to the rigid rules that characterize the high-profile IMPACT program in Washington DC, which is studied by Dee and Wyckoff (2015), principals in HISD have flexibility in how to

act on the performance information.[1] The presumption is that they know their own schools best and can leverage their local knowledge.[2] One way to view the HISD system is as a scaled up version of the experimental pilot interventions studied by Sartain and Steinberg (2016) and Rockoff et al. (2012), with the added feature of a district-wide emphasis on tying personnel decisions more closely to quality.[3] Beyond differences in the setting in terms of discretion and scale, we also examine the distributional impacts that can arise when an entire school system is treated.

2. Policy background

HISD has implemented several policies designed to raise staff quality and effort over the past decade. 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 effort 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.[4] A notable feature of

[1] Using a regression discontinuity design, Dee and Wyckoff (2015) find that dismissal threats associated with low ratings induce voluntary exit and raise the performance of teachers who remain.
[2] Loeb, Miller, and Wyckoff (2015) study a reform that increased principals' flexibility in decision-making in New York City and find that principals are less likely to award tenure to less effective teachers as measured by value-added and own annual assessments of teacher quality.
[3] In their New York City experiment, Rockoff et al. (2012) find that providing principals with improved information on teacher performance increases the likelihood of exit for low performing teachers. Sartain and Steinberg (2016) find similar effects of a Chicago pilot program that evaluated teachers more rigorously via classroom observations.
[4] See Appendix A for an illustration from the district's perspective of the complementary levers designed to shift the distribution of teacher effectiveness.

the new system, which is still in effect today, is that principals are the primary agents of policy implementation. Further, principals have significant latitude in exiting teachers since, unlike many districts, teachers at HISD do not have tenure and most teachers are on one-year contracts.

The cornerstone of the ETI reform is the implementation of a rigorous teacher evaluation system intended to provide more informative reviews of teacher performance. 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 principal 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. 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 averaged 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 assigned ratings in 2010-11. Prior to that year, ratings for almost one in three teachers were not recorded with the district. Further, ratings were high and did not meaningfully differentiate teachers (Weisberg et al., 2009).[5] The 2010-11 ratings were based on at least two classroom observations and, though not formally scored under the new system, these observations were

[5] Under the former appraisal system, 97% of HISD teachers were rated "proficient" or better.

conducted in an environment where differentiating teachers by quality was a leading district concern and most principals had already received training.[6] In the following year, 2011-12, teachers were scored under the two new observational components: instructional practice and professional 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.[7]

The phasing of the reform had different implications for the subset of teachers that we study, who teach in tested subjects and grades, than for other teachers. Student achievement metrics for these teachers had already been available for many years because the SAS Institute provided proprietary teacher-level value-added estimates (EVAAS® scores) to HISD as key inputs to the merit pay system. Thus, the most relevant changes are the addition of the observational components and the emphasis on differentiating teachers and tying personnel decisions more closely to quality. 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 subset because information about efficacy in promoting student learning was already available.[8]

---

[6] 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 of the report is: "In 2010-11, 373 teachers exited the organization for performance reasons. That's up from 77 in 2009."

[7] A side effect of the phased implementation is that the official ratings for teachers changed significantly from 2011-12 to 2012-13. When only the observational components were included, nearly 90% of teachers were labeled effective or highly effective, with 0.5% ineffective, 10.0% needing improvement, 57.6% effective, and 31.9% highly effective. Once student performance was explicitly factored in, these shares changed to 6.0%, 27.7%, 39.7%, and 26.6%, respectively. The downward shift is due to the fact that the student performance measures are relative, so that some teachers will necessarily be deemed ineffective, while the other criteria are absolute. Unfortunately we do not have access to personnel evaluations for earlier years to document any initial shifts.

[8] 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

We examine how the introduction of ETI has affected teacher turnover in ways that are related to quality. Though turnover is only one channel through which the new human capital policies could affect the quality of the workforce, it is arguably the lever that principals can affect most, particularly in the short run. However, it is important to recognize that any impacts on turnover reflect both demand-side and supply-side responses to the initiative as a whole, including supporting interventions bundled with the new evaluation system. Teachers are provided regular feedback on progress and opportunities for development to address their specific needs, and new leadership roles have been established for effective teachers to mentor others. Associated changes in the work environment and career opportunities could alter the relative attractiveness of teaching in the district and of teaching in more and less advantaged schools for teachers of differing effectiveness. These types of changes are likely inherent to any rigorous appraisal system.

Something that is more unique to the Houston context is that ETI was introduced against the backdrop of a merit pay system. Under the system, teachers in core subjects can receive bonuses for student learning gains exhibited in their classrooms and smaller bonuses for campus-wide performance. Prior to ETI, nearly all core teachers received bonuses, with the average bonus on the order of $3,600 (or about 7 percent of average base salary). With the onset of ETI, the standards were made more stringent. Whereas it had been sufficient to be in the top half on at least one teacher-subject or campus measure, qualifying for an award required being closer to the top 20 percent in 2011-2012. In 2012-13, once ETI was fully phased in, teachers identified as ineffective or needing improvement by the appraisal system were also disqualified from

departments at Wellesley began responding to an anti-grade inflation policy during a transition year in which the policy was discussed, though it was made clear the policy would not be implemented until the following year.

receiving campus awards. More details on changes to ASPIRE over the course of our study period are provided in Appendix A.

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 (shown in Appendix A). Teachers experienced these changes with a significant lag, since award details are not available until the spring of the academic year and the awards are announced and paid in the following year (regardless of whether the individual is still an employee). The first post-ETI year can be viewed as providing insights about effects in a less discriminating merit pay regime, while the next two years embed the increasing alignment of the merit pay program with ETI to the extent that the change was perceived by teachers.[9] For districts that have merit pay programs, such realignment would be expected in response to changes to how teachers are evaluated.

3. Data and summary statistics

We have access to detailed school, teacher, and student administrative data files for school years 2007-08 through 2013-14. These data allow us to measure teacher turnover through 2012-13 (where 2013-14 data are used to measure turnover for 2012-13 teachers), leaving us with a six-year panel centered around 2010-11, which is the first year the human capital policies began to take force.

3.1 Measuring school disadvantage and selecting analysis schools

We begin with a sample of 201 traditional public schools in HISD that were operational during our sample period and serve students in grades 3 to 8, which are the grade levels for

[9] The evidence on how ASPIRE incentives affect teacher behavior in HISD is mixed. Imberman and Lovenheim (2015) find that high school teachers increased effort in response to team incentives under ASPIRE, but Brehm, Imberman, and Lovenheim (2017) do not find any evidence of strategic effort responses to individual incentives among teachers in lower grades.

which we are able to construct measures of teacher quality consistently over the course of our data 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 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 an additional 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 the student bodies served across these schools. 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.

[10] In Appendix D, we show that our main findings are qualitatively similar if we include these schools.

3.2 Measuring teacher quality and selecting analysis teachers

Critical to our analysis is being able to measure teacher effectiveness in a comparable way over the full sample period. While teacher experience and education levels might be candidates, the literature has consistently shown that these observable characteristics explain little of the variation in student learning and are not consistently linked to teacher quality (Aaronson, Barrow, and Sander, 2007; Hanushek and Rivkin, 2006; Harris and Sass, 2011). We also have scores from principal appraisals for the components that were part of the official evaluation system in 2011-12 and 2012-13, but not only are these unavailable in prior years, the observational components are difficult to compare 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 the value-added estimates that have been provided to principals 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. Although the technical estimation details differ from standard value-added models, conceptually they are similar (Ballou, Sanders, and Wright, 2004). Teachers' EVAAS® scores are estimated from regressions of student achievement on a set of indicators for each teacher the student had in the current and prior two years, as well as indicators for subject, grade, and year. These scores are available to us back to the 2006-07 school year, and we restrict our analysis to teachers in grades 3-8 who have been assigned math or reading EVAAS® scores.

Because the single-year estimates are quite noisy, we construct a more informative measure of teacher effectiveness by combining multiple years of teachers' scores per the

following regression based on Chetty, Friedman, and Rockoff (2014a):

$$V_{ikt} = \delta_0 + \mathbf{V_{ikt\text{-}}}\boldsymbol{\delta_1} + \eta_{ikt} \quad (1)$$

In equation (1), V_{ikt} is teacher *i*'s EVAAS® score in subject *k* and year *t*, $\mathbf{V_{ikt\text{-}}}$ 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, $\hat{V}_{ikt} = \hat{\delta}_0 + \mathbf{V_{ikt\text{-}}}\hat{\boldsymbol{\delta}}_{\mathbf{1}}$, 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 regression models 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 teaching the relevant subject contemporaneously. An implication of including only scores from years prior to *t* as explanatory variables in equation (1) is that first-year teachers are necessarily excluded from the sample. However, our reliance only on prior-year performance guards against introducing survivor bias to our turnover analysis, since otherwise teachers who persist would be more likely to have quality measures available and thus be overrepresented in our sample.[12] Another implication of our strategy is that teacher effectiveness is allowed to drift

[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).

[12] Our jackknifed measures rely on more observations for teachers in later years of our panel, so it may seem that a relative reduction in noise could confound our estimated relationships over time. 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 is also a theoretical argument against this concern (Jacob and Lefgren, 2008).

over time, consistent with the slow-moving process documented by Chetty, Friedman, and Rockoff (2014a).

It is important to demonstrate that our measures meaningfully capture teacher effectiveness in raising student achievement. 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 a different grade, test scores for students in the new school-by-grade combination should increase in the year after the change. Moreover, if the quality metric is properly scaled, the magnitude of the change in teacher quality should predict the magnitude of the change in student achievement.

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, as shown in Appendix B. The reading-based teacher quality estimates appear to be less informative (i.e., noisier), however, which is consistent with findings in previous research (Backes et al., 2016; Lefgren and Sims, 2012). Thus, we choose to present results restricted to teachers for whom we can observe effectiveness in teaching math.[13] Across our sample years, one-fifth of the teachers in our schools are teaching math in a tested grade, and one-fifth of these have no available prior math scores to calculate our jackknifed

[13] That said, Appendix D shows that results for reading teachers are qualitatively similar.

measure. Thus, in the end, our analysis sample represents 16.4 percent of teachers in our schools.[14]

The middle panel of Table 1 shows how math teacher quality is distributed across schools grouped by achievement level in the pre-policy period. In addition to our measure of effectiveness, where the numbers reported are in standard deviations of the teacher distribution, we also include other observable teacher characteristics. Based on our measure, Table 1 shows that teacher quality is not evenly distributed across the district. More low-quality teachers and fewer high-quality teachers are found at low-achieving schools. Low-achieving schools also employ teachers with less experience and more education, but we place little emphasis on these differences in observed qualifications because a simple regression of our jackknifed quality measure in math on teacher experience and indicators for education levels yields an R-squared of just 0.01.

Finally, we note that our quality measure is not directly available to school principals. Instead, principals have access to year-by-year EVAAS® scores and post-policy observational assessments, in addition to other indicators of quality that we do not observe. In 2012-13, the first year that all components of the assessment were formally scored, our measure of quality explains 24 percent of the variation in overall appraisal ratings among our sample of teachers. It explains 13, 4, and 23 percent of the instructional practice, professionalism, and student performance components, respectively. One reason that the correlation with the classroom-observation component is not higher may be that scores on these types of best-practices metrics have been shown to be sensitive to the composition of students in the classroom (Steinberg and

[14] Teachers who are not responsible for math instruction in a tested grade primarily teach students below grade 3 or students in other subjects (particularly in middle schools), or are non-classroom teachers who focus on special populations such as special needs students and English language learners (ELLs). The share of ELL teachers is especially large at HISD.

Garrett, 2016; Whitehurst, Chingos, and Lindquist, 2014). That said, both across and within schools, our jack-knifed quality measures are more predictive than single-year EVAAS® scores of the non-test-based evaluation components.[15] Thus, in addition to being an informative measure of how decisions under the new system are likely to affect student achievement, our measures are also better aligned than single-year EVAAS® scores with the other formal evaluation criteria in the system.

3.3 Measuring teacher turnover

In addition to measuring school exits, we decompose school exits into exits from the district and transfers to other schools within the district. A complication we face is that the staffing data provided to us in 2013-14 include only teachers. In all previous years the staffing data include all positions. For consistency, throughout our analysis we identify a teacher as having exited the school at the conclusion of year t if the teacher is not observed teaching in the school in year t+1. Thus, we code switches to non-teaching positions (e.g., school leadership) as exits. We have confirmed that our results are substantively the same if we exclude the last year of data and code position changes as non-exits.[16]

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. That said, the limitation of the single-year exit measures is that they are noisy and overstate exit rates. It has been well documented that 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

[15] 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.
[16] While the overall annual exit rates decline slightly (1-2 percentage points) if we recode position changes as non-exits, the across-year differences are hardly affected. See Appendix D.

having exited if she is also not present in year *t*+2.

The bottom panel of Table 1 shows single-year turnover rates for math teachers in grades 3-8 in the pre-policy period. Pre-policy turnover is 13.7 percent at low- and middle-achieving schools, versus 12.1 percent at high-achieving schools. The difference is driven primarily by a lower rate of within-district transfer from high-achieving schools. Unsurprisingly, two-year exit rates (not shown) indicate marginally lower turnover by approximately 0.4 percentage points, or 3 percent.[17]

4. Empirical strategy

To estimate effects on turnover, we begin with difference-in-differences models of the following form, specified as linear probability models:

$$Y_{ist} = \delta_0 + \delta_1 Post_t + \delta_2 Q_{it} + \delta_3 Post_t \times Q_{it} + \mathbf{X_{st}}\boldsymbol{\beta} + \phi_s + \varepsilon_{ist} \quad (2)$$

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*, $Post_t$ is an indicator set to one for 2010-11 and later years, and Q_{it} is our measure of teacher quality.[18] In some variants of the model, we replace the continuous measure of teacher quality with a vector of indicator variables for the bottom, middle-three, and top quality quintiles. While these variants have the advantage of allowing for nonlinear effects, we lead with the more parametric model since it has the advantage of parsimony and nicely summarizes whether the link between quality and turnover strengthened on average following the reform. The **X**-vector contains teacher

[17] The pre-policy turnover statistics in Table 1 are similar to turnover statistics provided for grade 4 to 5 teachers in New York City by Ronfeldt, Loeb, and Wyckoff (2013).

[18] Shrinkage is implicit in the jackknifing procedure and thus our estimates will not be affected by attenuation bias from using a generated regressor as would be the case with a standard linear predictor (Jacob and Lefgren, 2008).

characteristics that might have independent effects on turnover, such as race, gender, experience, and education. Our findings are not sensitive to which subset of these characteristics we include in the regressions, nor are they sensitive to omitting the **X**-vector entirely. 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.

The objective of the model is to identify shifts in the relationship between teacher quality and exit over time, embodied by δ_3. We also report estimates of δ_1 to give a sense of how the overall teacher exit rate changes over time. To the extent that the change can be attributed to policy implementation, it is indicative of impact on the *extensive margin*. Of course, it is difficult to rule out that other time-varying factors are at play when estimating these simple differences. Thus, we emphasize estimates of δ_3, which is the coefficient on the interaction between the post-reform indicator and teacher quality. This parameter provides an indication of the policy impact on the *intensive margin* – that is, on a per-exit basis it provides a measure of how workforce quality is changing due to push and pull factors associated with the reform. Since a primary goal of the policy was to increase exit of ineffective teachers and increase retention of effective teachers, we would expect to find a negative coefficient on the interaction.

A necessary condition for identifying the policy impact on relative exit rates is that pre-policy trends in exit rates between teachers of different levels of quality are the same. To explore the validity of the design, as well as to provide evidence on the time pattern of any responses, we also estimate event time models. For these time-disaggregated models, we include a full set of year effects and year effects interacted with teacher quality.

Beyond estimating average impacts, we also study heterogeneity across schools to shed light on distributional effects. For these models, we add interactions between the time and quality

variables with indicators for schools that are low (bottom-quintile) and high (top-quintile) achieving based on pre-period achievement. Improving teacher quality at schools serving high-need students was 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 (Bates, 2016).

5. Effects of the reform on turnover

5.1 Descriptive analysis

We begin by visually documenting trends in exit and turnover rates in Figure 1. The figure shows district-wide trends for our 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 turnover by all three measures began to rise at the conclusion of the 2010-11 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. This six-year period spans the Great Recession. Unemployment peaked in 2009 and gradually declined over the next several years. In Appendix C, using panel data on teachers from neighboring and other large Texas districts, we show a consistent U-shaped pattern in turnover over this period, with turnover returning to initial levels by the final year. Thus, much of the post-reform increase appears to be unique to HISD, suggesting the policy played at least some

role.

Figure 2 provides similar information to Figure 1 but divides teachers into groups based on our measure of quality. Teachers are assigned to the following groups based on their placement in the quality distribution: bottom quintile (least effective), middle quintiles (quintiles 2-3-4), and top quintile (most effective). The three panels report the rates of school exit, district exit, and school transfer for the three teacher quality groups. 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 by district exits. 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.

One might worry that the economic recovery also confounds our ability to attribute the differential changes in exit by quality to the reform. That is, more effective teachers might respond differently to secular changes in outside options. Here, the existing literature does not offer much guidance. The only paper that we are aware of that considers the role of the economy on teachers 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, 2015). It is difficult to extrapolate this finding to the exit decision, and unfortunately we do not have access to teacher quality measures for other Texas districts to offer a counterfactual. It is reassuring, though, that turnover was trending similarly across teacher quality groups for the three years prior to the reform despite the changing economic conditions (Figure 2). In the empirical analyses, we more formally address this point by testing for differential pre-trends and for sensitivity to a teacher's level of experience.

5.2 Estimation results

We estimate the models described in Section 4 to assess the significance and robustness of the patterns illustrated in Figures 1 and 2. First, Table 2 shows results from equation (2) where we enter the teacher-quality measure into the regressions linearly. We report results for the full specification, which 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. For each turnover outcome, we present models that aggregate the pre- and post-policy time periods and models that fully disaggregate years. All control variables except for the post-period indicator (or year indicators) are mean-centered in the regressions, including the school indicator variables, so that the intercept can be interpreted as the exit rate at the mean values of all covariates.[19]

The general patterns from Figures 1 and 2 are reflected in the model estimates and confirmed to be statistically significant in Table 2. For example, in the model of district exits, our estimate in column (2a) implies that a teacher who is one standard deviation above average in the quality distribution is an additional 6.3 percentage points less likely to exit the district during the post-policy relative to the pre-policy period. The estimated impact is attenuated for the inclusive school exit outcome in column (1a) since, as suggested by Figure 2 and shown in column (3a), there is no change in the relationship between school switching and teacher quality. Though we do not emphasize the pre-policy patterns in exit by quality, we find that less effective teachers were more likely to exit the district and less likely to transfer to a new school within the

[19] 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).

district prior to policy implementation.[20]

The event time models in the table are useful for two reasons. First, they document that there were no pre-trends in turnover by quality (i.e., the coefficients on the interactions between teacher quality and the 2009 and 2010 indicators, which are estimated relative to the holdout year 2008, are small and statistically insignificant). Moreover, though the causal impact on overall turnover rates is not well identified by our model, the pre-period trend is in the opposite direction of what we see in the post period. A second benefit of the disaggregated models is that, in principle, they allow us to test how impacts evolve over time. We do tend to find the largest and most statistically significant impacts in the final year, but are unfortunately not sufficiently powered to differentiate these from the estimates for the other reform years.

Next, in Table 3 we show results from models where we replace the linear quality measures in equation (2) with indicators for teachers' quality-quintile groups. The indicator identifying teachers in the middle quintiles (2-3-4) is omitted for comparison. We do not show the intercept coefficients and interactions to preserve space. Consistent with what we show in Figure 2, Table 3 confirms that the post-policy period is marked by a large and statistically significant increase in the likelihood of school and district exit for low-performing teachers relative to middle and high performing teachers. Between these two latter groups there is no divergence in exit rates.

Table 4 reports on the robustness of our 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

[20] 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 within HISD.

in the data, we look forward two years to determine whether the exiting teacher remained either (a) out of the school or (b) out of the district. When we make this definitional change, we are no longer able to examine outcomes for the 2013 teacher cohort. Thus, we report results from models covering the 2008-2012 cohorts using the one-year and two-year exit definitions, which are otherwise comparable to the results we show in Table 2. Although the overall levels of exit are slightly lower with the two-year definition, the patterns in our estimates are very similar across the two definitions.

In the right panel of Table 4 we return to using our full dataset and single-year exit measures and replicate the analysis in Table 2 (columns 1a, 2a, and 3a) after restricting the models to include only schools that did not experience a principal change. Changes in leadership are one mechanism by which the new evaluation system could influence the workforce. Approximately 38 percent of schools in our analytic sample retained the same principal over the course of the full data panel (this number is in line with data on principal tenure in Texas as reported by Branch, Hanushek, and Rivkin, 2012). The results are generally similar to what we report in Table 3, which suggests that principal changes are not a critical mediator of our findings.

Table 5 shows results from models run separately for teachers by experience level. We divide teachers into three groups based on experience: ≤ 5 years, 6-20 years, and more than 20 years, and run the model for each turnover outcome separately for each group. Overall exit rates increased for all experience groups and although noisily estimated, the pre-post change in the relationship between quality and exit/transfer is similar across experience groups. The consistency by experience is surely facilitated in part by the absence of tenure at HISD and the fact that teachers are primarily on 1-year contracts. This result would be unlikely to generalize to

districts with strong tenure protections (Sartain and Steinberg, 2016).

Finally, in Table 6 we estimate models that are otherwise the same as those shown in Table 2, but we replace our jackknifed teacher quality measures with single-year EVAAS® scores. The results clearly show that turnover in the post period aligns much more strongly with our jackknifed quality measures – which are unobserved by principals and district officials, at least directly – than with the noisier current-year EVAAS® scores, which are observed. This result is consistent with the interpretation that personnel decisions under the reform are being made based on more comprehensive evidence than the current-year scores.

6. Heterogeneity in effects across schools

6.1 Descriptive analysis

In Figure 3, we replicate the information shown in Figure 1, but separately for each school type. Recall that 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). The figure shows that while exit rates increased across all three groups in the post-policy period, there has been a disproportionate increase at the lowest-achieving schools.[21]

Figure 4 further divides teachers by quality within the same school groups. Reading across a row in Figure 4 holds the school-achievement group fixed, and reading down a column holds the teacher quality group fixed. Although the graphs in the figure cut the data thinly, and therefore noise is an issue, they suggest several interesting patterns. For instance, the first row of graphs shows school-exit, district-exit, and school-transfer patterns at low-achieving schools, by

[21] Appendix C shows that though the U-shaped pattern in exits across years is more pronounced for disadvantaged schools in neighboring and other large districts, the levels also return only to initial levels and do not exceed these.

teacher type. The clear bars across the first row illustrate that low-performing teachers at low-achieving schools were much more likely to exit the district in the post-policy period relative to the pre-policy period. However, when looking at rates of school exit (black bars), the pre-post change relative to other teachers at these schools shrinks because school-transfer rates (gray bars) climb for middle- and high-performing teachers. Bates (2016) suggests a potential mechanism – namely, more effective teachers under the new system have more prominent signals of their ability than had previously been the case and can leverage these signals into more desirable teaching positions. In the absence of true compensating wage differentials, as is typical in the public education context, teachers will prefer positions that are more desirable along non-pecuniary dimensions (Greenberg and McCall, 1974).[22]

Turning to middle-achieving schools, there is also a more pronounced increase in district exit rates for low-performing teachers. In this case, there is not an offsetting school-changer effect. For high-achieving schools, school changing is relatively stable across years for all teacher quality groups, and the difference in the pre-post increases in district exits between low- and high-performing teachers is muted.

6.2 Estimation results

We add interaction terms to equation (2) to test whether the patterns suggested by Figures 3 and 4 are statistically significant in our difference-in-differences specification. Table 7 presents results from models that use the linear quality measures, like in Table 2, but with the added interactions for school type.[23] The heterogeneity parameters of interest interact the post-policy

[22] Teachers may prefer higher-achieving schools for a variety of reasons. Survey evidence suggests that while teachers do prefer to work with higher-SES students, perhaps because this requires less effort, simple non-pecuniary benefits that are correlated with student SES, like better administrative support and shorter commute times, are more important in pushing teachers toward high-SES schools (Horng, 2009).
[23] Models that use teacher quintile groups in place of the linear quality measure, akin to what we show visually in Figure 4, yield estimates that are too imprecise to be informative.

indicator with teacher quality and school type, where bottom- and top-quintile schools are included in the model and compared to the holdout group of middle-quintile schools. The bottom two rows of Table 7 show coefficient estimates and standard errors for these triple interactions.

To interpret the findings in Table 7, first note that the post-policy effect on the relationship between each measure of turnover and teacher quality for middle-achieving schools is shown in row 4. By virtue of their omission from the model as the holdout group, the double interaction of quality and the post-period indicator is the effect for these schools. The estimates in Table 7 for middle-quintile schools are similar to the analogous global estimates shown in Table 2. For bottom quintile schools, the triple interaction coefficients in the second-to-last row indicate the quality effect at these schools relative to middle quintile schools. Although the estimates are noisy and merely suggestive, they imply that some of the benefits arising from a more negative relationship between quality and district exit (column 2) is dulled by a more positive relationship between quality and school transfer (column 3). For top-quintile schools, the triple-interaction estimates indicate that the net effect of the reform is essentially null. That is, the triple-interaction terms in the final row of Table 7 offset the baseline effects in row 4.

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 that gains in student learning would align with the change in the quality composition of exiters. However, whether or not gains are realized also depends on any impacts on the quality of teacher entrants and on effort among teachers who remain (Rothstein, 2015). Inference is further clouded by the high turnover rate among 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 gauge how the turnover aspects of the policy might be expected to affect student achievement, we estimate models of changes in school-by-grade math 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*.[24] The focal explanatory variables are the change in the level of math teacher turnover the two cohorts were exposed to, ΔTO_{sgt}, and the share of math teachers exiting between years 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. We include the unknown category to account for math teachers without jackknifed quality measures, which makes these four categories exhaustive. 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 level of teacher turnover across cohorts, which captures changes in the composition of teachers, but also for changes in exposure to turnover, which captures differential disruption.

The results are shown in Table 8. While the estimates should be viewed cautiously because they rely on realized differences in turnover across grades, they seem quite plausible.

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

The estimated coefficients on the exit shares of bottom, middle, and high quality teachers are 0.172, 0.066, and -0.169, respectively. Further supporting the validity of our quality measures, these align closely with the average jackknifed effectiveness measures for each group, which are -0.132, 0.012, and 0.175, respectively (converted to student standard deviation units). The estimated coefficient on the change in turnover implies that there is a disruption effect (Hanushek, Rivkin and Schiman, 2016; Ronfeldt, Loeb, and Wyckoff, 2013), which we estimate to be 0.088 student-level standard deviations for a school-by-grade cell that experiences 100 percent turnover. If we attribute all of the observed increase in turnover to the reform, student achievement would be predicted to fall by 0.012 standard deviations through this channel.

Setting aside the potential disruption effect, we now use the estimates from Table 3 (columns 1a and 2a) to perform a back-of-the-envelope calculation of the impact of the reform on student achievement through the changing quality of leavers. Figure 5 suitably repackages the estimates, using them to calculate the implied shares of leavers falling into each teacher quality quintile group before and after the reform. Specifically, we use the estimated coefficients on the quality and time indicators, and their interactions, to predict the exit rate for each quality group by period. We then multiply these rates by the group shares to calculate the overall exit rate and the group exit shares. Combined with our estimates of group-specific differences in teacher effectiveness, the pre-post shift toward more low-performing teacher exits shown in Figure 5 implies that the effectiveness of school (district) leavers falls by just 0.007 (0.007) student standard deviations, on average.

While small, there are three dimensions along which these back-of-the-envelope magnitudes may be viewed as conservative. First, note that any given percentage-point difference in the likelihood of exit across teacher groups (per Table 3) will map to a smaller

difference in percent shares of exits the higher is the level of turnover. Thus, the fact that turnover rose following the reform means that the amount of increased targeting that we estimate translates into smaller compositional changes than if overall turnover had remained constant. Had turnover risen by only half as much (which we simulate by cutting the coefficient on the post-policy indicator in half), the estimated achievement effects increase but remain small, at around 0.012 (0.013) student standard deviations per turnover.

Second, ETI was being phased in over our post-reform period, with student achievement formally incorporated only in the most recent year of our data panel, 2013. Impacts observed in this year may be most relevant for evaluating the policy under full implementation. If we repeat our calculations but this time use point estimates for the policy effects from 2013 in place of the pooled point estimates (columns 1b and 2b in Table 3), the expected average change in achievement rises to 0.012 and 0.009 student standard deviations per turnover for school and district exits, respectively.

The third dimension involves the longer-term effects of the policy on workforce quality overall. Following the logic of Winters and Cowen (2013) we iterate over 10 years, recognizing that our estimates of policy impacts become less informative as the workforce evolves. Since we attempt to isolate the role of the attrition channel, we hold the quality distribution of replacement teachers fixed to match the initial distribution. The district quality distribution evolves over time owing to differential attrition and we hold the exit rates by quintile-group fixed. Considering variants that make alternative assumptions about the district-level turnover rate, we find effects on average workforce quality over a 10-year horizon on the order of just 0.01 student-level standard deviations. Note that the average workforce effects are diluted because the per-turnover effects documented above are spread across all teachers.

We conclude from these calculations that while the change in the composition of turnovers induced by the policy is in the right direction, exits are not targeted well enough to induce meaningful achievement gains. 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 (Staiger and Rockoff, 2010; Winters and Cowen, 2013). One reason for the smaller on-the-ground effect size is that the policies in the simulation studies are based on value-added and, while careful to account for inaccuracies in the value-added estimates themselves, they do not account for the fact that systems that have been put into place in practice incorporate non-test-based evaluation components are less aligned with how teachers affect student achievement (Kane et al., 2011). In any evaluation of program efficacy centered on student achievement, personnel decisions that incorporate information from such measures will appear to be mis-targeted.[25] A second possible reason for the smaller impacts is that we study just the first few years under the new system at HISD, and the effectiveness of the policy may improve over time as agents gain experience with the new regime (Ahn and Vigdor, 2014).

8. Conclusion

We study the effects on workforce composition of the introduction of a new, more-rigorous teacher evaluation system in the Houston Independent School District. The new system clearly affected the composition of exiting teachers, primarily by increasing the exit rate among low performers relative to higher-performing teachers. Policy activity along this dimension has been concentrated at low-achieving schools within the district; at high-achieving schools, there is

[25] Teacher effects on outcomes other than test scores also vary substantially and are not highly correlated with teacher effects on test scores (Jackson, forthcoming), which is a rationale for the inclusion of multiple measures in teacher evaluations. However, to date there is little evidence connecting the non-test-based measures used in teacher evaluations to teacher effectiveness as measured by non-test student outcomes.

little indication that the nature of personnel decisions changed in the post-policy years.

Our analysis illustrates the potential for more rigorous teacher evaluations to improve student outcomes via better-informed personnel decisions, but also highlights a critical 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. There are also 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 – but on the dimension of selective attrition, the compositional effects have not been large enough to measurably improve student achievement.

Stepping back from our narrow policy context, a possible complementary intervention that might 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, its merit pay program has faced challenges and taken only partial steps in that direction (Brehm, Imberman, and Lovenheim, 2017; Shifrer, Turley, and Heard, 2013).

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. 2016. The Common Core Conundrum: To What Extent Should We Worry that Changes to Assessments and Standards Will Affect Test-Based Measures of Teacher Performance. CALDER Working Paper No. 152.

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. 2016. Public and Private Learning in the Market for Teachers: Evidence from the Adoption of Value-Added Measures. Unpublished manuscript.

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. (forthcoming). What do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy*.
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.
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. 2015. Weak Markets, Strong Teachers: Recession at Career Start and Teacher Effectiveness. NBER Working Paper No. 21393.

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.

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.

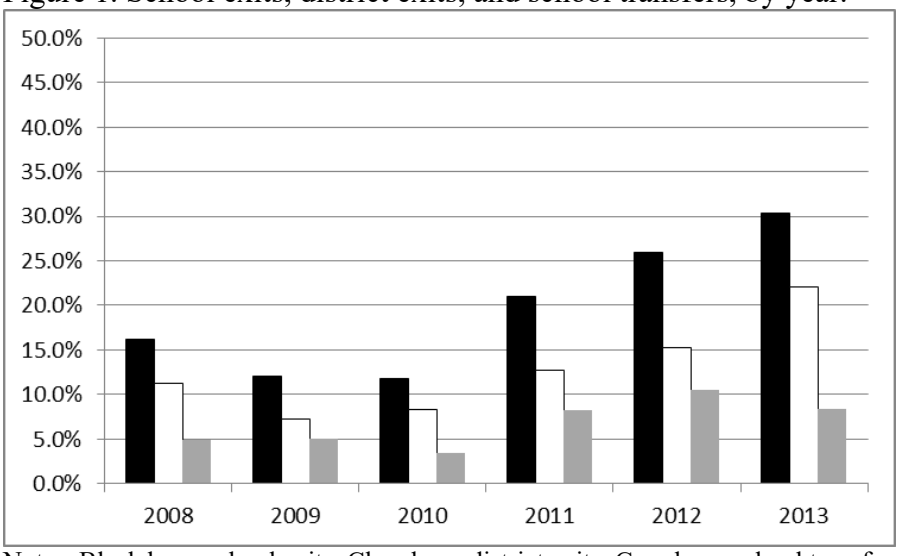
Weisberg, Daniel, Susan Sexton, Jennifer Mulhern and David Keeling. 2009. The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. New York: The New Teacher Project.

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.

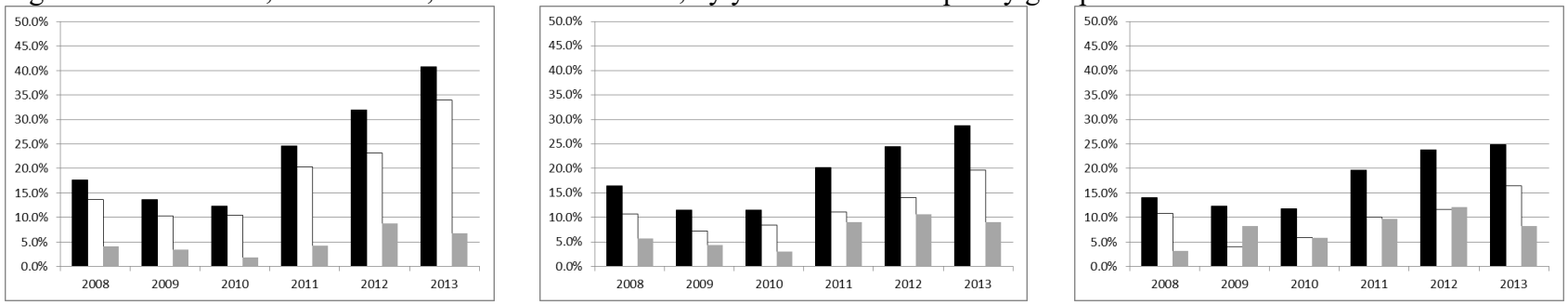
Winters, Marcus A. and Joshua M. Cowen. 2013. Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies. *Journal of Policy Analysis and Management* 32(3), 634-654.

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



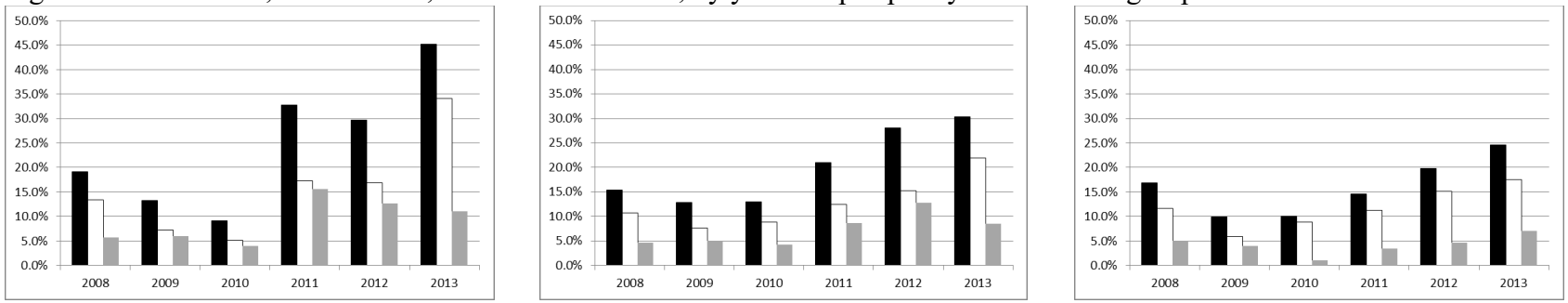
Notes: Black bars: school exits; Clear bars: district exits; Gray bars: school transfers. School exits are the sum of district exits and school transfers.

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



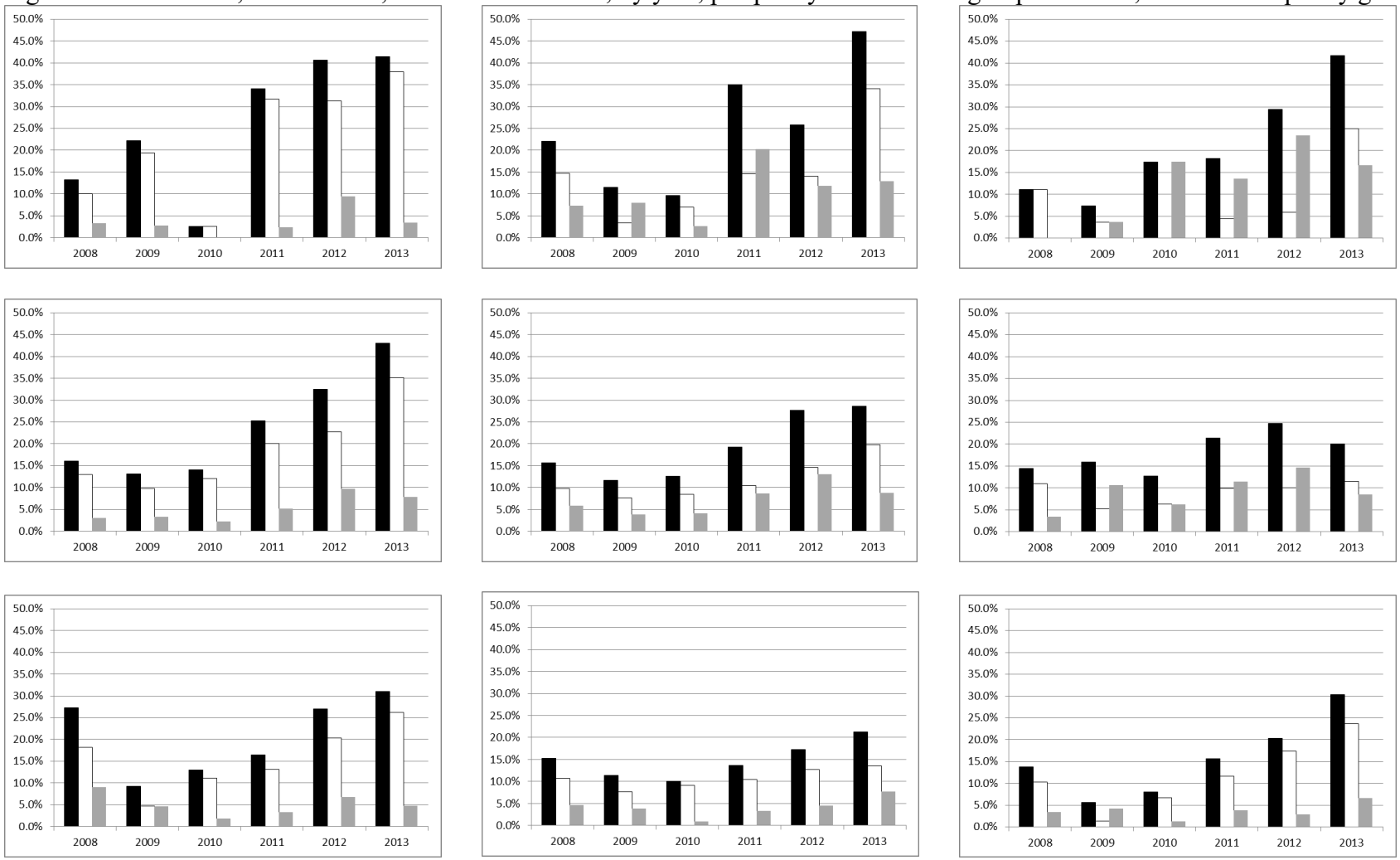
Notes: Black bars: school exits; Clear bars: district exits; Gray bars: school transfers. School exits are the sum of district exits and school transfers. From left to right, the graphs show turnovers for bottom-quintile, middle-quintiles (2-3-4), and top-quintile teachers in the quality distribution.

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



Notes: Black bars: school exits; Clear bars: district exits; Gray bars: school transfers. School exits are the sum of district exits and school transfers. From left to right, the graphs show turnovers at bottom-quintile, middle-quintiles (2-3-4), and top-quintile schools in the pre-policy (average of reading and math) achievement distribution.

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

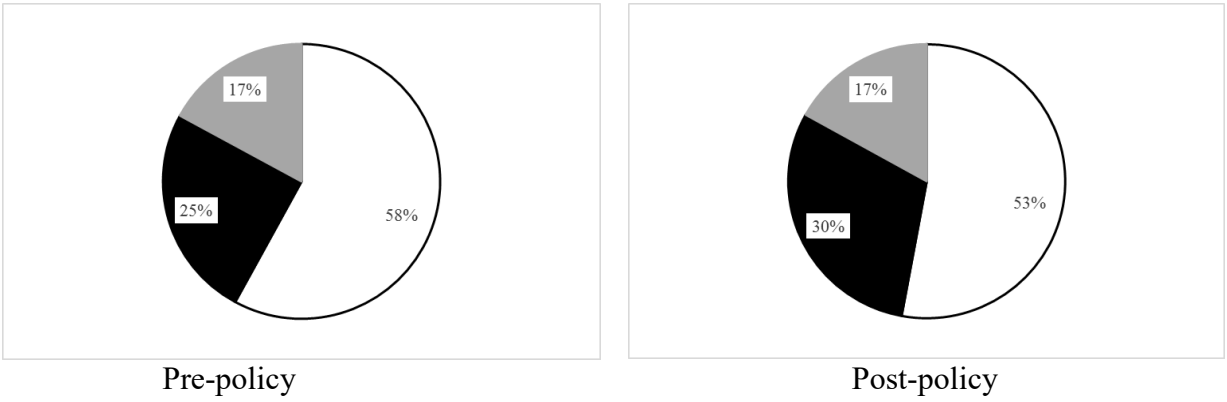


Notes: Black bars: school exits; Clear bars: district exits; Gray bars: school transfers. School exits are the sum of district exits and school transfers. Row 1: bottom-quintile schools; Row 2: middle-quintiles schools; Row 3: top-quintile schools; Column 1: bottom-quintile teachers; Column 2: middle-quintiles teachers; Column 3: top-quintile teachers.

Figure 5. School and district exits, by teacher-quality group in the pre- and post-policy periods.

School exits:

District exits:



Notes: Black: bottom-quintile teachers; Clear: middle-quintiles teachers (quintiles 2-3-4); Gray: top-quintile teachers. The figures are derived from model-based estimates of the proportions of exiting teachers by quality-quintile group in the pre- and post-policy periods, for school and district exiters, taking the pre- and post-policy total exit rates as given.

Table 1. Summary statistics for pre-policy years 2007-08 through 2009-10.

	Schools by achievement level		
	Low	Middle	High
Student characteristics			
Average achievement z-scores	-0.295	-0.061	0.484
Percent free lunch	55.3%	57.6%	33.4%
Percent reduced price lunch	8.8%	11.0%	11.0%
Percent black	40.8%	20.7%	20.1%
Percent Hispanic	56.9%	75.3%	47.2%
Percent ESL	16.7%	8.1%	4.8%
Number of grade 3-8 students tested	25,125	86,628	44,560
Math teacher characteristics			
Jackknifed quality measure	-0.012	0.047	0.098
Percent bottom quintile	21.5%	21.4%	14.5%
Percent top quintile	14.1%	21.2%	21.0%
Years of experience	9.9	10.7	12.0
Percent with 1 to 5 years experience	34.6%	33.5%	30.6%
Percent with master's degree	34.6%	30.5%	28.5%
Percent with doctoral degree	2.5%	1.4%	1.3%
Number of teacher-years	483	2508	970
Math teacher turnover			
Exited the school in t+1	13.7%	13.7%	12.1%
Exited the district in t+1	8.5%	9.0%	8.8%
Transferred to another school in t+1	5.2%	4.7%	3.3%

Notes: The columns present summary statistics for analysis schools divided into three groups based on pre-policy achievement levels, averaged across reading and math: bottom quintile, middle three quintiles, and top quintile. The bottom-quintile sample is smaller because treated schools in Fryer's 2014 study are omitted and also because low-achieving schools have lower enrollment on average. School exits are the sum of district exits and school transfers. As described in the text, our analytic dataset excludes first-year teachers and correspondingly these teachers are also excluded from the teacher summary statistics.

Table 2. Impacts of the new evaluation system on turnover, by linear math teacher quality.

	Dependent variable is an indicator for:					
	School exit		District exit		School transfer	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Intercept	0.125	0.148	0.085	0.105	0.040	0.043
	(0.006)**	(0.011)**	(0.004)**	(0.009)**	(0.004)**	(0.007)**
Intercept*POST	0.137		0.083		0.054	
	(0.012)**		(0.008)**		(0.008)**	
Intercept*2009		-0.031		-0.035		0.003
		(0.015)**		(0.012)**		(0.011)
Intercept*2010		-0.035		-0.026		-0.009
		(0.015)**		(0.012)**		(0.009)
Intercept*2011		0.062		0.025		0.037
		(0.019)**		(0.014)*		(0.013)**
Intercept*2012		0.119		0.050		0.069
		(0.018)**		(0.013)**		(0.013)**
Intercept*2013		0.172		0.123		0.049
		(0.019)**		(0.017)**		(0.012)**
Teacher quality	-0.008	-0.025	-0.034	-0.029	0.026	0.004
	(0.015)	(0.032)	(0.012)**	(0.027)	(0.009)**	(0.017)
Teacher quality*POST	-0.061		-0.063		0.002	
	(0.021)**		(0.019)**		(0.016)	
Teacher quality*2009		0.024		-0.009		0.033
		(0.036)		(0.030)		(0.024)
Teacher quality*2010		0.024		-0.002		0.027
		(0.037)		(0.031)		(0.023)
Teacher quality*2011		-0.001		-0.042		0.040
		(0.039)		(0.034)		(0.028)
Teacher quality*2012		-0.020		-0.047		0.027
		(0.044)		(0.038)		(0.032)
Teacher quality*2013		-0.074		-0.090		0.016
		(0.043)*		(0.038)**		(0.022)
Teacher characteristics	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
R-squared	0.101	0.108	0.081	0.089	0.078	0.080
N (Teacher-year)	7800	7800	7800	7800	7800	7800

Notes: ** Denotes statistical significance at the 5 percent level; * Denotes statistical significance at the 10 percent level. Standard errors clustered by school are reported in parentheses. Observations for teachers during the 2011, 2012, and 2013 school years are coded as "POST" in columns 1a, 2a, and 3a. In these columns, the parameters for the variables interacted with POST are estimated relative to the pre-period years. In columns 1b, 2b, and 3b, the year-specific parameters are estimated relative to 2008, which is the first year of our data panel. All variables in the regressions other than the post indicator (or year indicators) are mean-centered so the intercept can be interpreted as the exit rate at the mean values of all covariates. The coefficients for all variables other than the teacher quality measures and the intercept, both interacted with time, are excluded for brevity.

Table 3. Impacts of the new evaluation system on turnover, by math teacher quality quintile.

	Dependent variable is an indicator for:					
	School exit		District exit		School transfer	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Bottom quintile	0.012	0.012	0.027	0.033	-0.015	-0.021
	(0.015)	0.030	(0.013)**	(0.024)	(0.008)*	(0.017)
Top quintile	0.006	-0.012	-0.009	0.010	0.015	-0.023
	(0.015)	(0.025)	(0.011)	(0.023)	(0.010)	(0.015)
Bottom quintile*POST	0.076		0.089		-0.013	
	(0.023)**		(0.022)**		(0.013)	
Top quintile*POST	-0.007		0.003		-0.010	
	(0.023)		(0.020)		(0.016)	
Bottom quintile*2009		0.000		-0.010		0.010
		(0.038)		(0.032)		(0.022)
Bottom quintile*2010		0.000		-0.007		0.007
		(0.037)		(0.032)		(0.020)
Bottom quintile*2011		0.050		0.068		-0.018
		(0.038)		(0.034)**		(0.022)
Bottom quintile*2012		0.074		0.070		0.004
		(0.045)		(0.039)*		(0.028)
Bottom quintile*2013		0.109		0.114		-0.005
		(0.051)**		(0.047)**		(0.025)
Top quintile*2009		0.032		-0.030		0.062
		(0.036)		(0.027)		(0.026)**
Top quintile*2010		0.023		-0.026		0.049
		(0.033)		(0.028)		(0.022)**
Top quintile*2011		0.023		-0.006		0.029
		(0.036)		(0.032)		(0.025)
Top quintile*2012		0.028		-0.010		0.038
		(0.036)		(0.033)		(0.027)
Top quintile*2013		-0.014		-0.032		0.018
		(0.045)		(0.041)		(0.027)
Teacher characteristics	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
R-squared	0.102	0.110	0.083	0.092	0.078	0.081
N (Teacher-year)	7800	7800	7800	7800	7800	7800

Notes: The specifications are the same as in Table 2, other than that indicators for the top and bottom quintiles replace the continuous quality variable. The omitted group includes teachers in quality quintiles 2, 3, and 4. See notes to Table 2.

Table 4. Tests for robustness of impacts of the new evaluation system on turnover, by linear math teacher quality.

	1-year exit definition 2008-2012 cohorts		2-year exit definition 2008-2012 cohorts		No principal change 1-year exit definition 2008-2013 cohorts		
	School exit (1a)	District exit (2a)	School exit (1b)	District exit (2b)	School exit (3)	District exit (4)	School transfer (5)
Intercept	0.128	0.086	0.124	0.079	0.104	0.077	0.027
	(0.005)**	(0.003)**	(0.005)**	(0.003)**	(0.006)**	(0.005)**	(0.005)**
Intercept*POST	0.113	0.058	0.112	0.051	0.102	0.064	0.038
	(0.013)**	(0.009)**	(0.013)**	(0.008)**	(0.013)**	(0.010)**	(0.010)**
Teacher quality	-0.008	-0.031	-0.011	-0.031	-0.004	-0.033	0.029
	(0.015)	(0.012)**	(0.015)	(0.012)**	(0.021)	(0.018)*	(0.011)**
Teacher quality* POST	-0.037	-0.046	-0.034	-0.041	-0.069	-0.048	-0.021
	(0.024)	(0.020)**	(0.024)	(0.019)**	(0.028)**	(0.023)**	(0.021)
Teacher characteristics	X	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X	X
R-squared	0.097	0.078	0.098	0.077	0.085	0.073	0.067
N (Teacher-year)	6656	6656	6656	6656	3054	3054	3054

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 2008-12 only, for whom we can define exits looking forward both 1 and 2 years in the data, to examine the sensitivity of our findings to the exit definition holding all else equal. In the right panel, we return to using our primary single-year exit definition and 2008-13 cohorts, but restrict the sample to include only schools in which there was not a principal change over the course of the data panel.

Table 5. Impacts of the new evaluation system on turnover by teacher experience and linear math teacher quality.

	School exit			District exit			School transfer		
	Exp ≤ 5	5< Exp≤20	Exp>20	Exp≤ 5	5<Exp≤20	Exp>20	Exp≤ 5	5< Exp≤20	Exp>20
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)	(3a)	(3b)	(3c)
Intercept	0.187	0.102	0.084	0.133	0.062	0.074	0.054	0.040	0.010
	(0.009)**	(0.007)**	(0.013)**	(0.008)**	(0.005)**	(0.012)**	(0.007)**	(0.005)**	(0.006)*
Intercept*POST	0.136	0.132	0.158	0.081	0.072	0.115	0.054	0.060	0.043
	(0.021)**	(0.013)**	(0.027)**	(0.017)**	(0.010)**	(0.025)**	(0.015)**	(0.010)**	(0.012)**
Teacher quality	-0.008	-0.012	0.045	-0.039	-0.043	0.042	0.031	0.031	0.004
	(0.030)	(0.020)	(0.034)	(0.025)	(0.014)**	(0.034)	(0.019)*	(0.014)**	(0.017)
Teacher quality*POST	-0.077	-0.054	-0.101	-0.065	-0.054	-0.109	-0.012	0.000	0.008
	(0.050)	(0.029)*	(0.056)*	(0.042)	(0.023)**	(0.055)**	(0.032)	(0.023)	(0.023)
Teacher characteristics	X	X	X	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X	X	X	X
R-squared	0.153	0.118	0.246	0.138	0.093	0.208	0.116	0.110	0.279
N (Teacher-year)	2409	4174	1217	2409	4174	1217	2409	4174	1217

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. Each column reports results from a separate regression estimated for teachers by experience group, where teachers are divided into three groups: ≤ 5 years, 6-20 years, and more than 20 years experience.

Table 6. Impacts of the new evaluation system on turnover, by linear math teacher quality and using current-year EVAAS® scores in place of jackknifed measures.

	School exit	District exit	School transfer
	(1)	(2)	(3)
Intercept	0.125	0.085	0.040
	(0.006)**	(0.004)**	(0.004)**
Intercept*POST	0.136	0.082	0.054
	(0.011)**	(0.008)**	(0.008)**
Current-year EVAAS	-0.015	-0.015	0.000
	(0.006)**	(0.005)**	(0.004)
Current-year EVAAS*POST	-0.021	-0.023	0.002
	(0.010)**	(0.008)**	(0.006)
Teacher characteristics	X	X	X
School fixed effects	X	X	X
R-squared	0.103	0.079	0.077
N (Teacher-year)	7800	7800	7800

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply. The specification is the same as in Table 2 but replaces the jackknifed quality measure with the current-year EVAAS® score.

Table 7. Heterogeneous impacts of the new evaluation system on turnover by school type and linear math teacher quality.

	School exit (1)	District exit (2)	School transfer (3)
Intercept	0.126	0.085	0.041
	(0.005)**	(0.004)**	(0.004)**
Quality	-0.006	-0.035	0.029
	(0.018)	(0.014)**	(0.011)**
Intercept*POST	0.138	0.083	0.054
	(0.011)**	(0.008)**	(0.007)**
Quality*POST	-0.081	-0.078	-0.003
	(0.026)**	(0.022)**	(0.021)
Bottom quintile school *POST	0.094	0.051	0.043
	(0.035)**	(0.026)**	(0.025)*
Top quintile school*POST	-0.070	-0.020	-0.050
	(0.023)**	(0.018)	(0.014)**
Bottom quintile school*Teacher quality	0.003	-0.014	0.017
	(0.046)	(0.033)	(0.031)
Top quintile school*Teacher quality	-0.025	0.010	-0.035
	(0.037)	(0.029)	(0.020)*
Bottom quintile school*quality* POST	-0.029	-0.076	0.047
	(0.056)	(0.066)	(0.057)
Top quintile school*quality* POST	0.127	0.094	0.034
	(0.051)**	(0.044)**	(0.033)
Teacher characteristics	X	X	X
School fixed effects	X	X	X
R-squared	0.105	0.084	0.082
N (Teacher year)	7800	7800	7800

Notes: ** Denotes statistical significance at the 5 percent level; * Denotes statistical significance at the 10 percent level. Standard errors clustered by school are reported in parentheses. The omitted school type is from the middle quintiles (2-3-4) of the achievement distribution. Observations for teachers during the 2011, 2012, and 2013 school years are coded as "POST."

Table 8. Estimated effects of turnover on student achievement.

	Dependent variable: Difference in average test scores across cohorts within a school-grade
ΔTurnover	-0.088
	(0.025)**
Share bottom-quintile exit	0.172
	(0.055)**
Share middle-quintiles exit	0.066
	(0.035)*
Share top-quintile exit	-0.169
	(0.061)**
Share unknown-quality exit	0.137
	(0.056)**
School-grade characteristics	X
School fixed effects	X
R-squared	0.093
N (School-grade-year)	2824

Notes: ** Denotes statistical significance at the 5 percent level; * Denotes statistical significance at the 10 percent level. Standard errors clustered at the school-by-grade level are reported in parentheses. The turnover shares are scaled by teachers' 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.

Appendix A. Additional background on HISD policies

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 hiring better teachers and offering individualized training to existing teachers, the hope was to shift the distribution of teacher effectiveness in the district. Figure A1 is taken from "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. It nicely illustrates the role of the three complementary levers.

Selective retention/exit has continued to be an important focus in HISD. For example, the retention rate of highly effective teachers and exit rate of ineffective teachers are among the set of key indicators of progress reported in the annual Facts and Figures brief released by HISD to the public. These indicators were first included in the 2012-13 brief and, in that year, statistics were reported for differential turnover following the 2010-11 and 2011-12 school years.

The merit pay system: ASPIRE

The merit pay system was first introduced in 2006-07. In Table A1, we show 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. The share of teachers receiving any award fell slightly in 2010-11, due to the introduction of a teacher attendance requirement and minimum threshold for student growth. There were more dramatic falls in the share in 2011-12, when performance targets jumped, and 2012-13, when low student growth

precluded teachers from receiving campus awards. Teachers experienced these changes with a lag, since awards are announced and paid the following year (regardless of whether the teacher is still an employee).

Figure A2 shows the implications for the math teachers who are the subject of our empirical analyses. 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-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 statutory changes to the generosity of the maximums (Table A1), with the cut in 2011-12 and the increase in 2012-13. Finally, the last graph shows average award amounts unconditional on receipt. Here it is clear that the changes to provisions in 2010-11 had minimal effects, while the 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 on this dimension.

Figure A1. Overview of the ETI reform.

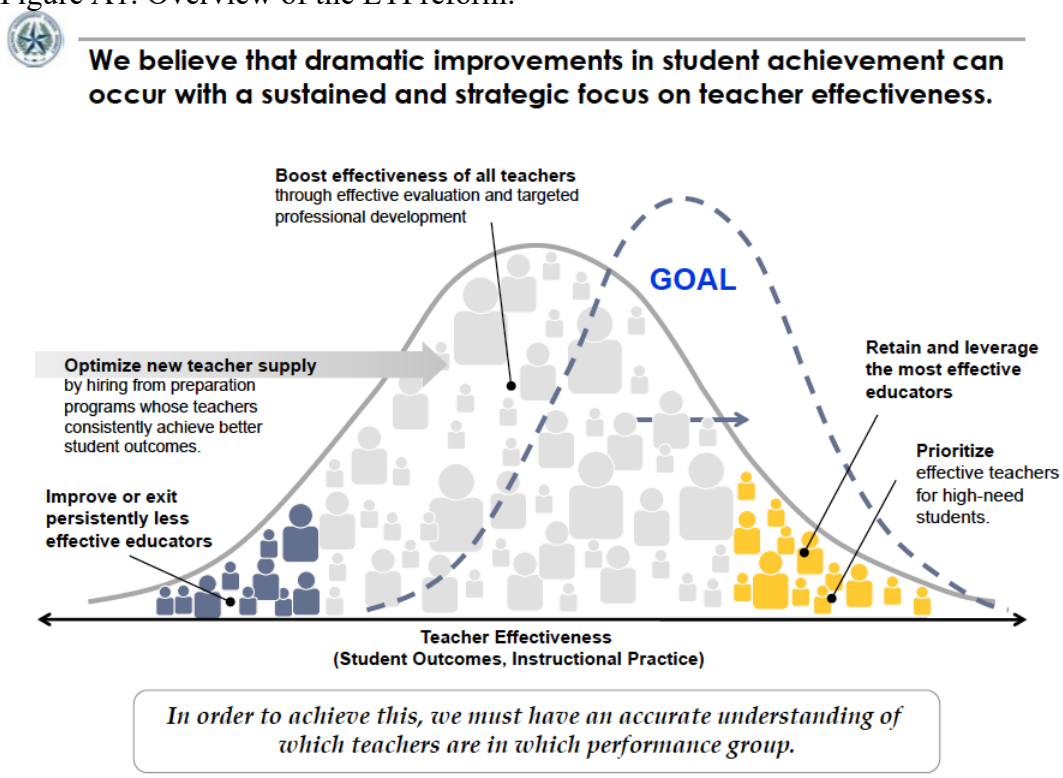
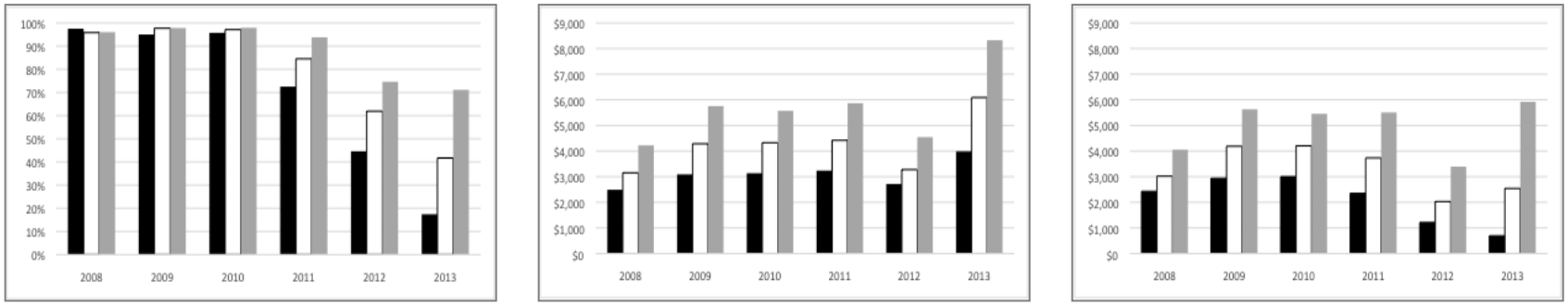


Figure A2. ASPIRE Awards, by year and teacher-quality group.



Notes: 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.

Table A1. ASPIRE program details.

	Campus Performance Awards		Individual Performance Awards				
	Distributional Targets	Award Max	Distributional Targets	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

Notes: Table 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).[26] We implement the forecasting test by estimating the following regression model, weighted by time *t* school-by-grade enrollment:

$$\Delta\overline{A}_{sgkt} = \gamma_0 + \Delta\overline{\hat{V}}'_{sgkt}\gamma_1 + \mathbf{\Delta\overline{X}_{sgt}\gamma_2} + \phi_t + \varepsilon_{sgkt} \quad \text{(B1)}$$

The dependent variable, $\Delta\overline{A}_{sgkt}$, 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 $\Delta\overline{A}_{sgkt}$. In addition to year effects, the control set includes $\Delta\overline{\hat{V}}'_{sgkt}$, which is the change in average measured teacher quality, and $\mathbf{\Delta\overline{X}_{sgt}}$, 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

[26] 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.

included in $\mathbf{V_{ikt\text{-}}}$ (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 exercise because internally valid estimates of forecasting bias can still be obtained even if survivors are over-sampled.

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 in any of the models. The larger standard errors in the reading regressions leave open the possibility of non-negligible bias and suggest more individual-level prediction errors, leading to our focus on math in the main text (though we provide comparable results for reading in Appendix D).

Table B1. Test for bias in jackknifed teacher quality measures.

	$\hat{\gamma}_1$	P-value (H_0: $\gamma_1=1$)	Number of school-grade-year cells
	(1)	(2)	(3)
Backward looking			
Grades 4-8, math	0.833	0.12	2612
	(0.106)		
Grades 4-8, reading	0.946	0.74	2630
	(0.163)		
Backward and forward looking			
Grades 4-8, math	0.999	0.99	3413
	(0.077)		
Grades 4-8, reading	0.974	0.85	3423
	(0.131)		

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

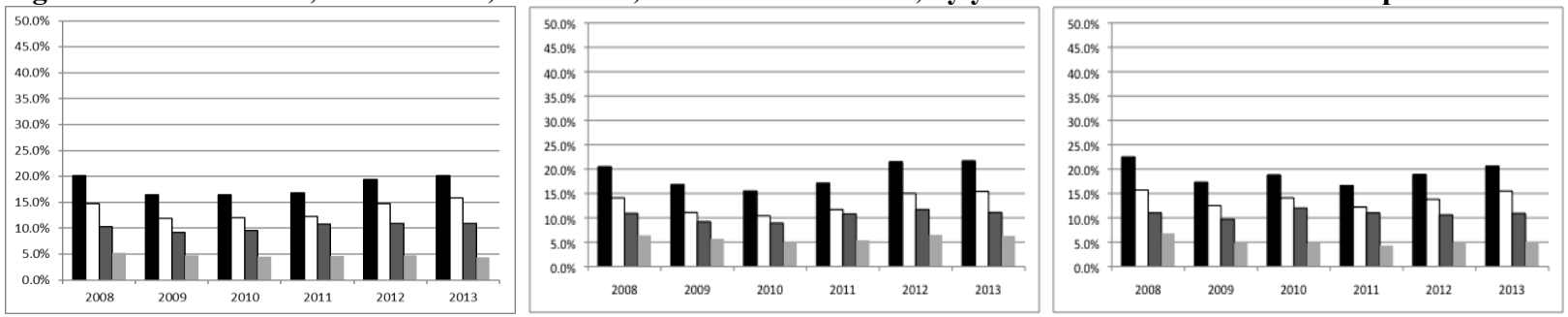
The period we study is one where economic conditions were not stable. With unemployment rates peaking in September 2009, and then steadily declining after, our post-policy period overlaps with an economic recovery. To explore how the level of teacher turnover varied with economic conditions in Texas, we compiled a statewide panel of personnel and campus data.

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 3-8. In order to classify schools by achievement levels comparably to our analysis of HISD, we combine TAKS pass rate data for three springs (2007-08, 2008-09 and 2009-10). 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.

In Figure C1 we show trends in the three measures of turnover we focus on in our analysis of HISD – school exits, district exits, and school changes (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 Texas, which we additionally include in the charts. Across all three samples, there is evidence of a U-shaped pattern in turnovers. Notably, though turnover rates rise smoothly over the last years of the period, they tend to return to the levels in the initial year. 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.

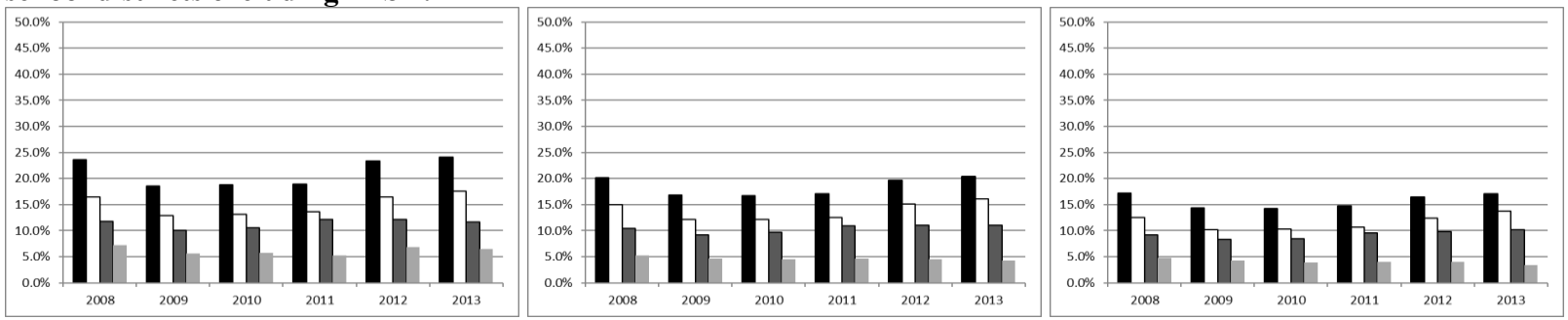
Figures C2a, C2b and C2c are comparable to Figure 3. Each figure shows turnover rates broken down by school achievement group (within district) for one of the three district samples, as indicated by the letter (a, b or c). As in HISD, the U-shape is more marked for lower achieving schools. However, once again, turnover rates return only to 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.

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



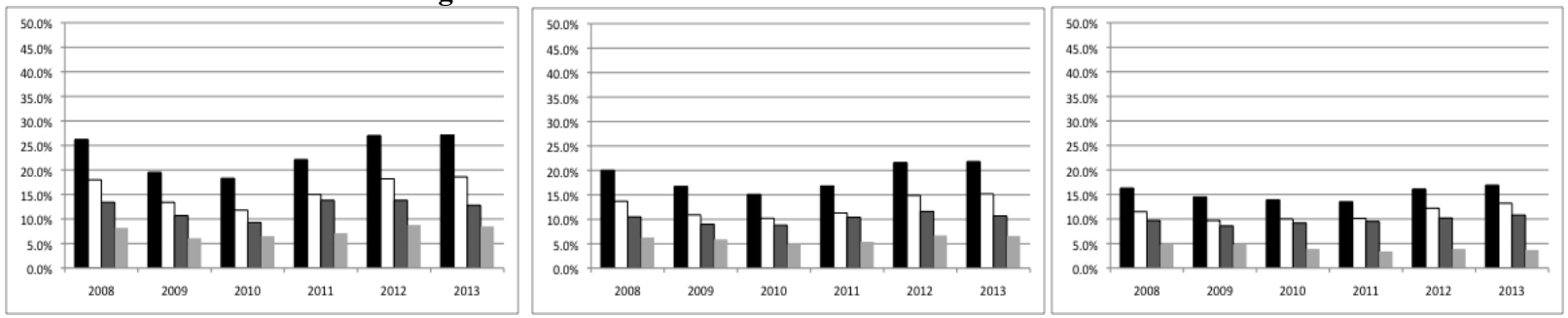
Notes: Black bars: school exits; Clear bars: district exits; Dark gray bars: state exits; Light gray bars: school transfers (within district). School exits are the sum of district exits and school transfers. From left to right, the graphs show turnovers at all Texas school districts other than HISD, at the five-largest school districts excluding HISD (Dallas, Cypress-Fairbanks, Northside, Austin, and Fort Worth), and at districts adjacent to HISD (Cypress-Fairbanks, Spring Branch, Katy, Alief, Stafford, Fort Bend, Alvin, Pearland, Pasadena, Galena Park, Sheldon, Humble, Aldine). 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 C2a. School exits, district exits, state exits, and school transfers, by year and school achievement group, for all Texas school districts excluding HISD.



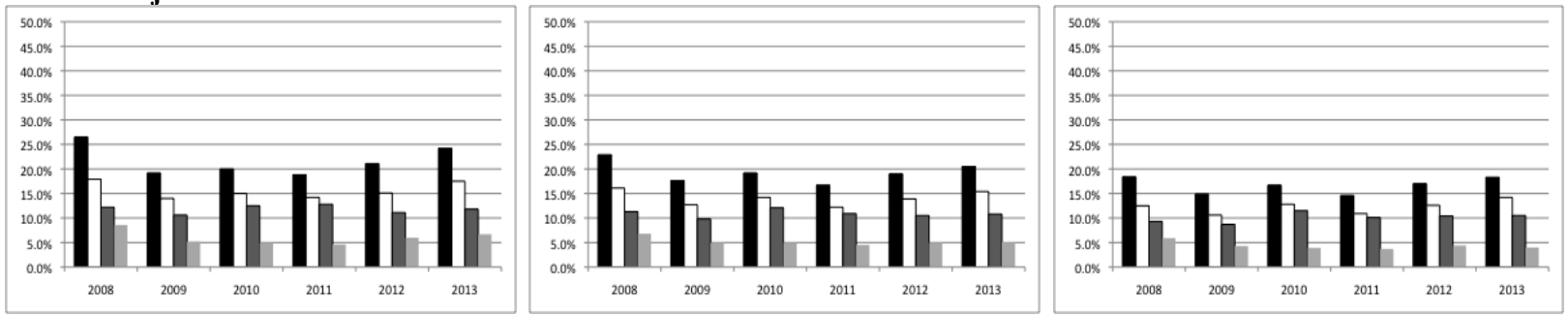
Notes: Black bars: school exits; Clear bars: district exits; Dark gray bars: state exits; Light gray bars: school transfers (within district). School exits are the sum of district exits and school transfers. From left to right, the graphs show turnovers at bottom-quintile, middle-quintiles (2-3-4), and top-quintile schools. Districts included: all Texas districts other than HISD. The turnover rates are teacher-weighted so they can be interpreted as the likelihood of exit for the typical teacher in the sample of districts.

Figure C2b. School exits, district exits, state exits, and school transfers, by year and school achievement group, five largest school districts in Texas excluding HISD.



Notes: Black bars: school exits; Clear bars: district exits; Dark gray bars: state exits; Light gray bars: school transfers (within district). School exits are the sum of district exits and school transfers. From left to right, the graphs show turnovers at bottom-quintile, middle-quintiles (2-3-4), and top-quintile schools. Districts included: Dallas, Cypress-Fairbanks, Northside, Austin, and Fort Worth. The turnover rates are teacher-weighted so they can be interpreted as the likelihood of exit for the typical teacher in the sample of districts.

Figure C2c. School exits, district exits, state exits, and school transfers, by year and school achievement group, school districts in Texas adjacent to HISD.



Notes: Black bars: school exits; Clear bars: district exits; Dark gray bars: state exits; Light gray bars: school transfers (within district). School exits are the sum of district exits and school transfers. From left to right, the graphs show turnovers at bottom-quintile, middle-quintiles (2-3-4), and top-quintile schools. Districts included: Cypress-Fairbanks, Spring Branch, Katy, Alief, Stafford, Fort Bend, Alvin, Pearland, Pasadena, Galena Park, Sheldon, Humble, and Aldine. The turnover rates are teacher-weighted so they can be interpreted as the likelihood of exit for the typical teacher in the sample of districts.

Appendix D. Supplementary tables

Because of the data limitation described in the text, for the 2013 cohort we must code moves to non-teaching positions as exits. For consistency in the analysis in the main text, we code moves to non-teaching positions as exits in all years. Table D1 shows results comparable to our main results in Table 2 but restricted to data from teacher cohorts between 2008 and 2012. For these cohorts we can estimate models of exit where (a) moves to non-teaching positions are treated as exits and (b) moves to non-teaching positions are not treated as exits.

Table D1 shows that this data limitation has no bearing on our findings, as results using both coding schemes are very similar for the 2008-2012 cohorts. This can be seen by comparing the (a) and (b) columns for each type of turnover. Note that the overall quality coefficients for the post-policy period are lower than in Table 2 because the policy impacts were highest in 2013, which is excluded from these models.

Table D1. Turnover models with linear math teacher quality. Standard versus alternative exit definitions for teachers in 2008-2012.

	Dependent variable is an indicator for:					
	School exit		District exit		School transfer	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Intercept	0.125	0.119	0.085	0.071	0.040	0.047
	(0.005)**	(0.005)**	(0.004)**	(0.003)**	(0.004)**	(0.004)**
Intercept*POST	0.113	0.110	0.058	0.050	0.055	0.063
	(0.013)**	(0.013)**	(0.009)**	(0.008)**	(0.009)**	(0.010)**
Teacher quality	-0.008	-0.005	-0.031	-0.032	0.024	0.025
	(0.015)	(0.015)	(0.012)**	(0.011)**	(0.009)**	(0.010)**
Teacher quality*POST	-0.037	-0.034	-0.046	-0.047	0.010	0.011
	(0.024)	(0.024)	(0.020)**	(0.019)**	(0.018)	(0.018)
Alternative exit definition		X		X		X
Teacher characteristics	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
R-Squared	0.097	0.099	0.078	0.084	0.089	0.088
N (Teacher-year)	6656	6656	6656	6656	6656	6656

Notes: Standard errors clustered by school are reported in parentheses. Observations for teachers during the 2011 and 2012 school years are coded as "POST". Parameters for the variables interacted with POST are estimated relative to the pre-period years. All variables in the regressions other than the post indicator are mean-centered so the intercept can be interpreted as the exit rate at the mean values of all covariates. The coefficients for all variables other than the teacher quality measures and the intercept, both interacted with time, are excluded for brevity.

Table D2 replicates our main findings from Table 2 but using teacher quality measures based on student test-score growth in reading instead of math. The results are qualitatively similar to what we show for math. Note that there is substantial overlap in the reading and math teacher samples (that is, there are many teachers who are responsible for student instruction in math and reading – e.g., self-contained elementary teachers).

Table D2. Impacts of the new evaluation system on turnover, by linear reading teacher quality.

	Dependent variable is an indicator for:					
	School exit		District exit		School transfer	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Intercept	0.124	0.146	0.087	0.108	0.040	0.037
	(0.005)**	(0.010)**	(0.004)**	(0.009)**	(0.004)**	(0.007)**
Intercept*POST	0.127		0.079		0.048	
	(0.011)**		(0.008)**		(0.007)**	
Intercept*2009		-0.015		-0.029		0.014
		(0.015)		(0.012)**		(0.010)
Intercept*2010		-0.049		-0.034		-0.015
		(0.014)**		(0.012)**		(0.008)*
Intercept*2011		0.051		0.028		0.024
		(0.018)**		(0.013)**		(0.012)*
Intercept*2012		0.106		0.038		0.068
		(0.018)**		(0.014)**		(0.013)**
Intercept*2013		0.176		0.119		0.057
		(0.020)**		(0.017)**		(0.011)**
Teacher quality	0.013	0.039	0.001	0.042	0.012	-0.003
	(0.018)	(0.042)	(0.015)	(0.032)	(0.011)	(0.023)
Teacher quality*POST	-0.080		-0.109		0.029	
	(0.030)**		(0.027)**		(0.018)	
Teacher quality*2009		-0.029		-0.059		0.030
		(0.052)		(0.037)		(0.033)
Teacher quality*2010		-0.027		-0.043		0.015
		(0.048)		(0.039)		(0.028)
Teacher quality*2011		-0.075		-0.115		0.040
		(0.055)		(0.044)**		(0.027)
Teacher quality*2012		-0.059		-0.119		0.060
		(0.063)		(0.050)**		(0.045)
Teacher quality*2013		-0.163		-0.223		0.060
		(0.063)**		(0.053)**		(0.036)*
Teacher characteristics	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
R-squared	0.096	0.106	0.076	0.088	0.080	0.084
N (Teacher-year)	7863	7863	7863	7863	7863	7863

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply.

Table D3 replicates our main results in Table 3 but using teacher quality measures based on student growth in reading instead of math. The results are qualitatively similar, although in the reading models there is some evidence that top-quintile teacher exit rates declined post-policy. Note that there is substantial overlap in the reading and math teacher samples (that is, many teachers are responsible for student instruction in math and reading – e.g., self-contained elementary teachers).

Table D3. Impacts of the new evaluation system on turnover, by reading teacher quality quintile.

	School exit		District exit		School transfer	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Bottom quintile	-0.021	-0.043	-0.018	-0.039	-0.003	-0.004
	(0.015)	(0.025)*	(0.014)	(0.021)*	(0.008)	(0.015)
Top quintile	0.014	-0.002	-0.005	-0.003	0.020	0.001
	(0.015)	(0.031)	(0.011)	(0.024)	(0.008)**	(0.014)
Bottom quintile*POST	0.054		0.073		-0.019	
	(0.026)**		(0.024)**		(0.015)	
Top quintile*POST	-0.039		-0.046		0.008	
	(0.024)		(0.019)**		(0.015)	
Bottom quintile*2009		0.043		0.042		0.001
		(0.037)		(0.030)		(0.022)
Bottom quintile*2010		0.023		0.021		0.002
		(0.030)		(0.027)		(0.015)
Bottom quintile*2011		0.060		0.075		-0.015
		(0.040)		(0.037)**		(0.021)
Bottom quintile*2012		0.072		0.108		-0.036
		(0.040)*		(0.036)**		(0.024)
Bottom quintile*2013		0.099		0.100		-0.001
		(0.045)**		(0.041)**		(0.028)
Top quintile*2009		0.045		0.014		0.031
		(0.044)		(0.031)		(0.023)
Top quintile*2010		0.006		-0.019		0.025
		(0.034)		(0.029)		(0.020)
Top quintile*2011		-0.018		-0.036		0.018
		(0.041)		(0.032)		(0.023)
Top quintile*2012		0.001		-0.020		0.021
		(0.044)		(0.035)		(0.027)
Top quintile*2013		-0.052		-0.098		0.046
		(0.049)		(0.037)**		(0.028)*
Teacher characteristics	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
R-squared	0.096	0.106	0.080	0.089	0.081	0.085
N (Teacher-year)	7863	7863	7863	7863	7863	7863

Notes: The estimates in this table are comparable to estimates in Table 3 and the notes to Table 3 apply.

Table D4 replicates our main findings from Table 2 for math but adds the schools where Fryer (2014) intervened back into the sample. The results are very similar to what we report in Table 2, which is expected because the Fryer schools make up a small fraction of district schools.

Table D4. Impacts of the new evaluation system on turnover, by linear math teacher quality, including Fryer schools.

	Dependent variable is an indicator for:					
	School exit		District exit		School transfer	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Intercept	0.127	0.144	0.085	0.102	0.043	0.042
	(0.006)**	(0.011)**	(0.004)**	(0.009)**	(0.004)**	(0.007)**
Intercept*POST	0.147		0.092		0.055	
	(0.011)**		(0.008)**		(0.008)**	
Intercept*2009		-0.029		-0.033		0.004
		(0.014)**		(0.011)**		(0.010)
Intercept*2010		-0.021		-0.019		-0.002
		(0.015)		(0.012)		(0.010)
Intercept*2011		0.076		0.034		0.042
		(0.019)**		(0.013)**		(0.013)**
Intercept*2012		0.139		0.067		0.072
		(0.019)**		(0.014)**		(0.012)**
Intercept*2013		0.186		0.132		0.054
		(0.019)**		(0.017)**		(0.012)**
Teacher quality	-0.008	-0.017	-0.031	-0.021	0.023	0.004
	(0.015)	(0.032)	(0.011)**	(0.027)	(0.009)**	(0.017)
Teacher quality*POST	-0.056		-0.070		0.014	
	(0.021)**		(0.018)**		(0.016)	
Teacher quality*2009		0.013		-0.016		0.029
		(0.035)		(0.029)		(0.023)
Teacher quality*2010		0.011		-0.009		0.020
		(0.041)		(0.032)		(0.023)
Teacher quality*2011		-0.024		-0.061		0.037
		(0.039)		(0.033)*		(0.027)
Teacher quality*2012		-0.020		-0.057		0.037
		(0.043)		(0.036)		(0.031)
Teacher quality*2013		-0.069		-0.101		0.032
		(0.042)		(0.037)**		(0.023)
Teacher characteristics	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
R-squared	0.108	0.114	0.086	0.094	0.082	0.083
N (Teacher-year)	8335	8335	8335	8335	8335	8335

Notes: The estimates in this table are comparable to estimates in Table 2 and the notes to Table 2 apply.