Performance Information and Personnel Decisions in the Public Sector: The Case of School Principals[1]

Julie Berry Cullen[2], Eric A. Hanushek[3], Gregory Phelan[4], and Steven G. Rivkin[5]

June 17, 2019

ABSTRACT

School accountability systems standardly report both categorical ratings and the underlying student pass rates that determine them, permitting direct investigation of how different information affects labor market outcomes of school leaders. We first use regression discontinuity design methods to investigate how outcomes vary, holding principal effectiveness constant. Results reveal large and significant impacts on principal salaries and job retention for crossing the unacceptable-acceptable boundary but not for crossing higher-ratings cutoffs. Complementary descriptive regressions corroborate these findings and also reveal that principal salary growth is higher when school pass rates are higher. Considering a composite measure of labor market success and controlling for pass rates, rewards for less readily available measures of effectiveness as well as for attaining better ratings are apparent only for the current district. The overall patterns suggest that school district administrators access and use more continuous information but also face pressure to respond to the stigma of low ratings. This apparent information breakdown could nonetheless be welfare improving if it raises the distribution of principal quality through disproportionate departures of less effective school leaders. However, the extensive overlap of achievement value-added distributions across school rating categories suggests that a categorical accountability system based on pass rates does not distinguish well among principals who lead schools with very different average achievement growth. Moreover, it almost certainly disadvantages principals in schools serving low-income populations that already have general difficulties in attracting and retaining effective educators.

[1] This work was done in conjunction with the Texas Schools Project at the University of Texas at Dallas. It was supported by grants from the Kern Family Foundation and the Laura and John Arnold Foundation. The conclusions of this research do not necessarily reflect the opinions or official position of the Texas Education Agency, the Texas Higher Education Coordinating Board, or the State of Texas.
[2] University of California, San Diego and NBER
[3] Stanford University, University of Texas at Dallas, and NBER
[4] University of Texas at Dallas
[5] University of Illinois at Chicago, University of Texas at Dallas, and NBER

1. Introduction

The lack of competitive pressures on public sector organizations has long raised concerns about low quality of service and inefficiencies in provision, with perhaps no sector receiving as much attention as public schools. Passage of the No Child Left Behind Act (NCLB) in 2001 was the culmination of many state-level efforts to measure and rate school performance with the explicit goal of elevating the quality of instruction and reducing inefficiencies. Although teacher performance under school accountability has received the most attention, it is the school leader who is the fulcrum of most school improvement efforts. The effect of accountability ratings and related performance measures on principal labor market outcomes therefore constitutes an important pathway through which accountability reforms might affect schools.

Test-based school accountability systems produce a wide variety of measures of school performance, permitting investigation of how alternative measures affect labor market outcomes of school leaders. Both categorical school ratings and the underlying pass rates that determine them are typically reported to the public, while system personnel additionally have access to more detailed and disaggregated test score data. Although school district administrators retain the authority over hiring, raises, and principal retention, they do not operate in a vacuum. Rather school superintendents report directly to school boards and likely respond to feedback from parents, politicians, and others in the community. Consequently, accountability systems may influence principal employment and salary decisions through multiple channels, and the various stakeholders likely rely on different types of information in reaching performance conclusions about principals.

In this paper, we study how different information about student body performance affects principals in Texas public elementary schools over the years 2001 to 2008. Texas offers a

noteworthy context as an early and influential mover in school accountability, as well as a state where principals are afforded substantial scope as managers and face a large labor market. Our focus on the elementary grades is motivated by the fact that achievement is the primary factor taken into account when rating schools in these early grades, which greatly simplifies characterizing information sets. Each year during our study period, elementary schools are placed into four categories – unacceptable, acceptable, recognized, and exemplary – based on overall and subgroup pass rates on statewide standardized exams.

We use regression discontinuity design (RD) methods to identify the causal impacts of reaching higher school rating categories on principals' subsequent labor market outcomes. Although there are no significant differences in the probability of principal retention or salary growth across the higher acceptable-recognized and recognized-exemplary boundaries, we find large and significant discontinuities at the lower unacceptable-acceptable boundary. A significant discontinuity at a rating threshold showing divergent treatment of equally productive principals constitutes *prima facie* evidence of an information failure.[1] One possibility is that district administrators focus on the rating despite having access to richer and more continuous information about principal performance. However, a discontinuity could emerge even when district administrators utilize more extensive information if, for example, other influential stakeholders rely heavily on the cruder and more salient ratings in drawing conclusions about whether the principal merits a raise or contract renewal. And, even if the principal is not dismissed, the environment may change in ways that make continuing in the job unattractive.

Whether the responsiveness to ratings that we uncover in the RD analysis is welfare improving depends upon the effects on the distribution of principal quality. This in turn depends

[1] Prior studies have documented that other markets react to the possible misinformation conveyed by discrete classifications. Most notably, Figlio and Lucas (2004) find that home prices and residential location respond to school grades in Florida, conditional on the variables used to construct the grades.

primarily upon the underlying motivations of district administrators and the differences in average principal effectiveness between ratings categories. For example, reluctance to fire a poor-performing principal due to psychological or political costs associated with such actions may be overcome by the public stigma surrounding a school rated as failing. Although the ineffective principal just above the threshold will retain her job, the introduction of discrete rating categories may operate below the threshold to lift the distribution of principal quality.

Answering this question requires measurement of principal effectiveness, which is not likely to be well-captured by the pass rates emphasized in the Texas system and NCLB. Since lower-income students typically receive fewer family resources that support learning and come to school less-prepared, pass rates are likely to be low at schools serving disadvantaged populations for reasons outside of the control of principals. Though inherently difficult to measure, principal effectiveness is likely to be positively correlated with proxies for value-added to achievement, such as annual school test score gains and estimates of principal fixed effects on achievement growth relative to other principals within the same connected networks. Mapping out the distributions of these proxies by ratings category illuminates the general failure of the Texas accountability system to discriminate by principal effectiveness. Although principals in the bottom quartile of effectiveness as measured by principal fixed effects are overrepresented in schools rated unacceptable, principals in these schools are also as likely to be in the top quartile as principals in schools rated more highly. The fact that even very effective principals are at risk of receiving low ratings likely amplifies staffing challenges facing low-achieving schools.

In complementary descriptive regression analyses that zoom out from the ratings thresholds, we explore how our two proxies for principal effectiveness and the other more readily observable information on school performance are associated with labor market

outcomes more broadly. These results corroborate that principal turnover is lower and salary growth is higher for principals leading campuses that escape the lowest rating, compared to other campuses within the same district and year. Salary growth is also higher when the school pass rate is higher, but we do not uncover any statistically significant effects of our best estimates of principal value-added on retention or salary growth. Interestingly, analyses that differentiate transitions to a new district from transitions within the current district do reveal suggestive evidence that higher value-added raises the probability of improved job outcomes within the current district. Though difficult to detect with the small number of principals switching districts, the absence of such a relationship with out-of-district success is consistent with the possibility that incumbent districts have greater access to and make greater use of information on productivity. Further, the home-district penalty to an unacceptable rating without a comparable receiving-district penalty underscores the possibility that pressure from interest groups enters into superintendent and/or principal decisions on continuation.[2] The overall pattern of results suggests that disseminating information that accurately measures school effectiveness could improve incentives to raise school quality and mitigate disincentives to work in schools that serve educationally disadvantaged populations.

Surprisingly, few studies have linked administrator outcomes to performance. In prior work on Texas, Cullen and Mazzeo (2008) find that first-time principals who lead schools where achievement is higher than expected given family background characteristics are more likely to move to more advantaged schools and to be promoted, realizing larger salary increases through these channels. Similarly, for Tennessee, Grissom and Bartanen (Forthcoming) find that

[2] For example, the *Tampa Bay Times* reported sudden replacement of principals when some of the Hillsborough County schools received D or F grades in Florida in 2018. Explaining that he was reacting to previous pressure, the Hillsborough superintendent of schools reported, "the State Board of Education ordered him [in 2017] to move principals out of four schools even though his own data showed they were doing a good job" (Sokol, 2018).

principals who receive high performance evaluations are more likely to leave for central office positions while those who receive poor evaluations are more likely to leave for lower-paid teaching positions. In contrast to these studies, since we have access to student level data, we are able to construct principal effectiveness measures based on individual achievement gains. For our analogous descriptive regressions, this allows us to compare market responses associated with more and less refined measures of principal effectiveness in raising student achievement. Importantly, incorporating extensive information on how the accountability ratings are assigned in our RD analysis allows us to causally identify the impact of ratings.[3]

The next section provides the relevant details on the Texas school principal labor market and school accountability system. Section 3 describes the Texas administrative data used in the analyses, while Section 4 explains the methods used to measure school performance and principal effectiveness. Sections 5 and 6 present the results of the analyses of principal labor market outcomes first pooling and then differentiating between current and destination districts. Finally, Section 7 summarizes the findings and considers implications for policy.

2. Institutional background

The principal labor market in Texas is likely to be more fluid than found in other states. Texas is one of the few states that prohibit public employees from entering into collective bargaining. School principals and teachers generally serve under term contracts, and those contracts cannot be longer than five years and are typically much shorter. Though the state does not collect data on contracts, a recent survey found that the standard contract term for principals

[3] Dizon-Ross (Forthcoming) uses a similar strategy to study how ratings affect teacher turnover in New York City, finding the surprising result that receipt of low relative to moderate ratings lowers turnover and raises the quality of entrant teachers. She speculates that this may be due to improvements in job desirability, since the effects are concentrated in schools led by principals rated as strong leaders by the teachers and teacher perceptions of this dimension of principal quality also improve.

is two years in most Texas districts.[4] Principals are required to have two years of classroom teaching experience in addition to completing a Master's degree in a principal preparation program. Although there is a state minimum salary schedule for teachers by years of experience, there are no such constraints on principal salaries, and salaries for principals are set by the superintendent, subject to school board approval.

Principals in Texas are required to be evaluated annually by central administrators. State code recommends standards for evaluating principals on specific indicators in the areas of instructional leadership, human capital development, executive leadership, school culture, and strategic operations. Importantly, academic progress of students at the school becomes a factor starting in the second year after a principal has been at a campus.

The evaluation of principals takes place within the broader system of statewide standardized testing and school accountability. The system determines not only the publicly available information on academic outcomes but also the data available to construct additional measures of principal productivity. Texas has required statewide testing since 1980 and was also an early mover on school accountability, having implemented a four-tiered school rating system starting in 1994. From that year through 2011, school ratings of unacceptable, acceptable, recognized, and exemplary were assigned by the state every year except for 2003 when there was a transition to a new standardized testing regime.[5]

In our analysis, we study elementary-school principals over the 2001 to 2008 school

[4] "HR Services' contract practices survey reveals common practices" (Texas Association of School Boards Human Resources Exchange Newsletter, February 2015, https://www.tasb.org/Services/HR-Services/Hrexchange/2015/February-2015.aspx).

[5] The Texas Assessment of Academic Skills (TAAS) was administered each spring starting in 1993 and was replaced by the Texas Assessment of Knowledge and Skills (TAKS) in 2003. Both are criterion-referenced tests that assess student mastery of grade-specific subject matter. In 2012, the testing regime changed again and a new set of rating categories were introduced in the following year. School years are referred to by the spring year (i.e., school year 2000-2001 is referred to as 2001).

years.[6] This choice of sample period and focus on elementary schools simplifies the analysis because test performance is the sole academic outcome used to construct the accountability rating.[7] The dropout rate contributes to the rating as early as grade seven, and other college readiness measures are incorporated in later grades. Elementary-school ratings depend on standardized test results in math and reading (grades 3-6), writing (grade 4), and science (grade 5). The administration of math and reading tests in consecutive grades also makes it possible to observe achievement growth in core subjects and to estimate principal value-added.

Analysis of the causal effects of ratings on labor-market outcomes requires replication of the ratings determination process, but the mapping from test scores to the campus rating is complex. First, separate pass rates for each subject based on year-specific cutoff scores for proficiency are calculated for all students and for demographic subgroups (white, black, Hispanic and economically disadvantaged) that meet minimum size requirements ranging from 30 to 50 students. Then, these pass rates are compared to thresholds that vary by rating category and year. The lowest pass rate across subjects and subgroups is the primary determinant of the accountability rating, but there are exceptions. For example, in the case of the acceptable rating, a subgroup not reaching the current statutory threshold in a subject but closing a specified percentage of the gap from the prior year can meet the alternative standard of required improvement.[8] The required improvement alternative is also available for the recognized rating, with the additional requirement that the pass rate fall no more than five percentage points below the statutory rate. The 2004 through 2008 accountability systems also include additional

[6] Other research on impacts of this Texas accountability system includes Deming et al. (2016) and Reback (2008) that study the distribution of student achievement and Craig, Imberman, and Perdue (2015) that studies budgets.
[7] Though data for 2009-2011 are available, a new measure was added to the accountability system that we were unable to successfully incorporate into our RD approach given the information available to us. The new "Texas projection measure" is based on the percent of failing students projected to pass in the next high-stakes grade given own current performance and prior year performance of all students at the school.
[8] In this case, the prior year pass rate is adjusted to account for any change in the cutoff score for passing.

exceptions provisions for campuses to be elevated to acceptable, recognized, and exemplary ratings: a specified number of subject-by-subgroups can be ignored if the pass rate falls no more than five percentage points below the statutory rate and the subject-by-subgroup did not receive an exception in the prior year.[9]

Appendix Table A1 shows the distribution of the binding subject-by-subgroup for the three boundaries between ratings categories. The patterns highlight the disproportionate share of schools for which science, which is only tested once for each cohort of students and for which the passing standard is the lowest, is the marginal subject at both the acceptable and recognized thresholds. Two factors contribute to this finding: students have more difficulty in science than in the other subjects and the much smaller number of science test-takers raises the error variance and the probability the average pass rate falls below the averages for other subjects. Importantly, it would not be apparent that science performance is often the determining factor without explicitly calculating distances from effective thresholds as we do (as described in more detail below).

Campus ratings are linked to both rewards and punishments. The state appropriates limited funding to provide financial awards to schools rated acceptable or above that show sustained or improved performance, as well as to schools led by principals identified as high-performing based on the same types of indicators. The highest performing campuses are also exempted from specific regulations. On the other hand, schools rated as unacceptable must work with external review teams to develop improvement plans. Receipt of an unacceptable rating in two consecutive years initiates the imposition of sanctions that become progressively more

[9] The number of subject-by-subgroup exceptions allowed is determined by campus size.

severe for each additional year the school fails to reach an acceptable rating.[10] After five years, requirements to replace staff or make other dramatic changes can directly affect principal job retention.

The detailed and summary information about academic performance and school ratings are made publicly available on the web. In evaluating principals, district administrators surely have additional information to go by, such as measures of performance on other dimensions, teacher reports, feedback from students and families, and direct observations. Yet, the extent to which these sources of information guide personnel decisions might be moderated by pressure from less informed stakeholders who focus on the more salient ratings. This motivates our primary analyses of how ratings per se impact principal labor market outcomes, as well as our secondary analyses of how other measures that differ in salience and information content are correlated with these outcomes.

3. Data on principal labor market outcomes and campus performance

To characterize labor market outcomes for elementary-school principals, we use a combination of restricted-use and publicly available data spanning the 2001 through 2008 school years. The restricted-use data we rely on are the administrative data constructed as part of the UTD Texas Schools Project.[11] Working with the Texas Education Agency, this project has combined different state data sources to create matched panels of staff and students. The personnel database provides annual information on administrator background characteristics,

[10] Starting in 2004, when the federal No Child Left Behind policy became effective, schools are also classified by whether they meet adequate yearly progress (AYP). The state aligns that determination as closely as possible to the school rating process, though federal rules require adjustments to some of the indicators, including the consideration of additional subgroups. During our sample period, only 8 percent of elementary schools designated as failing to meet AYP were also rated as unacceptable, and only 16 percent of campuses receiving an unacceptable rating failed to meet AYP. No schools progressed to a stage where repeatedly failing to meet AYP would have direct consequences for principals according to NCLB.

[11] https://www.utdallas.edu/research/tsp-erc/

total years of experience in the school system, current position, tenure, and salary. From this information, we track the careers of principals as long as they remain in Texas public schools. The student panels include demographic characteristics, instructional program participation, and achievement test scores. We merge data on campus characteristics and performance from the publicly available Texas Academic Excellence Indicator System. These comprehensive annual reports include accountability ratings, pass rates for subsets and for all students, and a broad range of contextual measures.

A significant advantage of studying Texas is the large number of principals and schools. Over our period, there are 3,942 elementary schools serving an average of 569 students in grades K-6 each year. Further, the typical school experiences a principal transition every 5 years.

Our main analytic sample includes principals with fewer than 25 years of total experience in the Texas Public Schools who have been in their current positions for at least two years. The exclusion of principals with high levels of experience reduces the incidence of exit via retirement. The exclusion of the first year in a school recognizes the limited initial control over staff composition (implying the persistence of predecessor decisions in the short run) and the introduction of academic progress for evaluation beginning in the second year.[12] Table 1 shows the effects of these sample restrictions. Starting from the full sample of campus-by-year observations, successively excluding highly experienced and new-to-campus principals hardly alters average school characteristics. Highly experienced principals are a bit more likely to have advanced education and enjoy slightly higher pay, while new-to-campus principals are quite typical. After making these exclusions, we observe 4,222 principals and 11,351 principal-by-year labor market transitions across 3,251 campuses.

[12] As expected, results are somewhat muted (e.g., escaping the lowest rating confers less of a boost to retention) when including first-year principals.

When constructing measures of principal effectiveness, we impose further restrictions on the sample. To account for fixed differences among schools, effectiveness is inferred from achievement gains relative to others serving the same campus. Thus, for this estimation, we eliminate any campus that is served by only one principal for at least two years over our sample period. This leaves us with 7,653 principal-by-year observations representing 3,285 unique principals. The final column in Table 1 shows that these principals and schools again appear to be typical, though there is a detectable fall in school average achievement and a corresponding increase in student disadvantage.

The primary measures of labor market outcomes are job retention, compensation, and student body composition. Since the latter two outcomes are observed only for those who remain in the Texas Public Schools, we also investigate exits from the system. Job retention and compensation are common measures of market outcomes but student composition is less standard and merits discussion. Past evidence highlights the influence of the quality of student and family inputs on the working conditions for teachers and administrators.[13] To create a summary measure of student advantage as a proxy for this aspect of working conditions, we regress average student pass rates across math and reading on the set of student characteristics from Table 1 as well as district and year fixed effects for all elementary school campuses in Texas over our sample period. We then extract the predicted values ignoring the year effects and, to simplify interpretation, standardize these to form an index with a mean of zero and standard deviation of one across campus-years. To characterize student composition at the district-year level analogously, we average the campus indices, weighting by enrollment, and then standardize this variable to have a zero mean and standard deviation of one across all district-years.

[13] Loeb, Kalogrides, and Horng (2010) and Hanushek, Kain, and Rivkin (2004) provide evidence of a desire for educators to work in higher-achieving, lower-poverty districts.

In our analysis of potential differences in the use of information between the current and destination district for movers, we construct a composite indicator of labor market success. This composite measure equals one for a principal who either retains her job or makes a "successful" move, defined as moving to another position within the school system and realizing above median salary growth or above median improvement in student composition, where the medians are defined based on all principals who remain in the system regardless of whether they switch jobs.

Timing is an important issue to consider when linking these labor market outcomes to measures of school performance. Though preliminary results on student test outcomes are available to district officials as early as May, preliminary accountability ratings are not released until August. Given that most principal hiring occurs in the spring, there is limited scope for immediate impacts on principal positions in the subsequent fall. We therefore use a two-year definition of outcomes, relating labor market transitions between academic years *t* and *t+2* to performance as measured in the spring of academic year *t*. For student composition, to avoid embedding any impacts of principals on student characteristics or outcomes, we calculate the change based on the values at time *t* at the sending and receiving campuses (or at the sending and receiving districts if the principal moves to a district-level position).[14] Thus, for those who continue in their current position (or move to a district-level position in the same district), the change is mechanically zero.

Table 2 shows summary statistics for the two-year labor market outcomes for relevant principals across the state (column 1), as well as for those in the subset at campuses served by multiple principals over our sample period (column 2). The majority (65.2 percent) of principals

[14] In rare cases, the receiving school or district was not operational in year t, so we use the working conditions index from t+1 if available, and t+2 if not.

in our main analytical sample are retained. Approximately one in five (19.9 percent) changes positions within the same district, one in ten (8.1 percent) exits the Texas Public Schools, and one in fifteen (6.9 percent) changes districts. Of those who change positions within the same district, three quarters make successful moves according to our definition, with most of these accompanied by above median salary gains. Successful moves outside the district account for a similar share of district movers and are also primarily attributable to salary. Altogether, 85.0 percent experience labor market success according to our composite measure. For principals at multi-principal campuses, the rate of success is similar (81.4 percent) but reflects less retention (56.1 percent) and more transitions within (19.0 percent) and across (6.3 percent) districts.

4. Measures of principal effectiveness

A natural way to judge principal effectiveness is by the academic performance of students at the school she leads. However, similar to the case of rating corporate CEOs, the level of performance depends on many factors that are not directly within the principal's control, including the composition of the student body, extent of parental support, decisions of the previous principal, and district policies. These influences clearly diminish the value of average pass rates as measures of principal effectiveness, though these and the resulting campus ratings are the most readily available to the public.

We produce an array of more compelling measures of effectiveness for comparison, though each has its limitations. The first is an adjusted pass rate, which is calculated as the residual from a regression of the average campus pass rate across math and reading on year fixed effects and the student characteristics shown in Table 1. This measure is case-mix adjusted for differences in the students served and is the type of adjustment that could be made by simply

comparing a school's performance to observably similar schools. It captures whether the campus is achieving at a level that is higher or lower than expected given the characteristics of its student body.

Recognizing that a limited set of characteristics is unlikely to adjust adequately for student and family differences, value-added models use prior achievement to account much better for unobserved heterogeneity. Estimates of school contributions to achievement growth rely on longitudinal data and could be computed by insiders, such as district administrators. The more parsimonious version of our achievement value-added model relates achievement (*A*) for student *i* in grade *g* in school *s* served by principal *p* in year *t* to a cubic in prior achievement ($f(A_{t-1})$), student characteristics (X), grade-peer characteristics (C), year-by-grade indicators (d_{gt}), and a vector of school-by-year fixed effects (g_{st}). Adding a random error (ε), the empirical model is:

$$A_{igspt} = a_1 f\left(A_{i,t-1}\right) + a_2 X_{it} + a_3 C_{gst} + d_{gt} + g_{st} + \varepsilon_{igspt} \qquad (1)$$

Achievement is defined to be the average of math and reading standardized test z-scores, where scores are normalized by grade and year across all students in the state. The vector *X* includes the student characteristics detailed in Table 1, while the vector *C* includes the averages of these characteristics for students in grade *g* in school *s* in year *t*. Estimates of the school-by-year effects are alternative proxies for principal effectiveness that capture whether achievement gains are higher or lower than expected. To reduce noise and measure more persistent differences across principals, we also calculate the average school-by-year fixed effect across a principal's term.

Our final measure is the most sophisticated but also the most demanding as far as data requirements and, for that reason, the least transparent. None of the measures discussed to this

point attempt to differentiate effective principals from schools that are effective for other reasons, such as being located in engaged communities or in localities with amenities that attract high quality teachers. To address this, we turn to panel-data methods, which have become standard in the broader labor economics literature to separate worker and firm productivity and have also been previously applied to estimate principal effectiveness.[15] This model replaces the vector of school-by-year fixed effects in equation (1) with vectors of school fixed effects (g_s) and principal fixed effects (q_p):

$$A_{igspt} = a_1 f\left(A_{i,t-1}\right) + a_2 X_{it} + a_3 C_{gst} + d_{gt} + g_s + q_p + \varepsilon_{igspt} \quad (2)$$

The estimates of principal effectiveness are based on the coefficients on the principal indicators.

This model identifies principal effectiveness vis-à-vis other principals in the same connected network, where schools in each network are linked by principal transitions among the schools. Since the estimates would otherwise be relative to an arbitrary omitted reference principal within each network, we demean our estimates by the average principal fixed effect in each connected network to anchor the estimates (Abowd, Creecy, and Kramarz, 2002).[16] Each school without a principal who ever leads another Texas public school is a separate network, in which case principal performance is measured relative to other principals at the same school. For our sample of principals leading elementary schools with at least two principals with at least two or more years of tenure over our sample period, three quarters of the networks (1,617 of 2,133) include a single campus. This is an important limitation of our relatively short panel, since systematic differences in principal effectiveness across campuses are identified only by switchers

[15] The general approach was pioneered by Abowd, Kramarz, and Margolis (1999). In the school context, see, for example, Branch, Hanushek, and Rivkin (2012), Coelli and Green (2012), Dhuey and Smith (2014), and Helal and Coelli (2016).

[16] The problems for the estimation of teacher value-added associated with test measurement error are far less important in the case of principals given the much larger number of test-takers in schools than classrooms, and Branch, Hanushek, and Rivkin (2012) show that Bayesian shrinkage has little effect on the variance of principal value-added estimates. Therefore, we do not make any further adjustments to account for sampling error.

across schools. Measuring effectiveness relative just to principals serving at the same schools will tend to understate variation in effectiveness to the extent that similarly effective principals lead the same schools.

To be unbiased measures of effectiveness within networks, principals must not be sorting based on match quality, and changes in school leadership must not be correlated with unobserved changes in other determinants of school quality. The influences of time-varying factors should be minimized by the conditioning on student and peer characteristics, as well as the sample restriction we impose throughout that excludes the first year of job spells. Evidence in Miller (2013) reveals a systematic decrease in school value-added in the year prior to the arrival of a new principal. Although poor performance may trigger a departure, the dip may also reflect a reduction in principal health, effort, or authority over the school or the impacts of other factors associated with the decision to leave. Achievement growth during a principal's first year might be inflated by recovery from the dip in the final year of the prior principal's tenure as well as the fact that the persistent influences of the prior principal are likely to be strongest during the first year of a spell.

Unfortunately, without a longer panel, it is difficult to validate these estimates by testing whether changes in principal effectiveness following turnover map to changes in student achievement.[17] Chiang, Lipscomb, and Gill (2016) pursue this type of exercise using a panel of similar length to ours. Their measure of productivity that is most closely related (i.e., school value-added during the principal's term relative to prior years) translates to less than one third of the expected change in achievement at the new school when one principal replaces another and is

[17] Bacher-Hicks, Kane, and Staiger (2014) and Chetty, Friedman, and Rockoff (2014, 2016) develop such tests for forecast unbiasedness of teacher value-added estimates. The logic is that if the estimates are valid and scaled appropriately, then changes in teacher effectiveness across schools and grades over time due to turnover should predict one-for-one changes in student achievement.

not significant.[18] This result echoes the finding in Grissom, Kalogrides, and Loeb (2015) that school-by-year value-added is more predictive of district evaluations of principals than measures that attempt to also control for school fixed effects. Importantly, these papers rely heavily on schools with multiple principals who serve short terms as leaders, and both include the initial and final years of principal terms in the analysis. Excluding these years, Laing et al. (2016) show that the variance of estimated principal value-added shrinks and that value-added increases monotonically with teacher ratings of the principal's effectiveness as an instructional leader. Taken as a whole, the evidence suggests that these estimates capture differences in principal effectiveness but may understate differences among principals at the same school due to the persistent effects of the prior principal on current achievement.

Table 3 reports correlations between our principal fixed effect measure and the other more readily calculated measures of campus performance one might use as proxies for principal effectiveness. The sample in the top panel is campus-years for campuses observed with more than one principal over our sample period, while the lower panel includes campus-years for campuses served by only one principal with two or more years of tenure. The variable in the first column is our principal fixed effect estimate from equation (2), demeaned by connected network (which is only available for the top panel of multi-principal schools). The variable in the second column is the school-by-year fixed effect based on equation (1), and the variable in column 3 averages these fixed effects across a principal's terms. Column 4 is the average campus pass rate across math and reading, and column 5 is the case-mix adjusted pass rate.

Principal value-added is positively correlated with all of the alternative campus performance measures, but the magnitude is much greater when the alternatives control for

[18] Their only statistically significant estimate relies on a simpler measure that does not adjust for prior campus performance, though the magnitude of the estimated effect is of a similar magnitude.

demographics or prior achievement. In the top panel, the correlation with the raw pass rate, which is the focus of the accountability system, equals only 0.10. This correlation rises somewhat to 0.15 when the pass rate is case-mix adjusted. The correlation with the school-by-year fixed effect starts out higher, at 0.19, and rises to 0.26 when the school-by-year effects are averaged over the term. The correlations among the alternative measures are broadly similar for the single-principal campuses in the bottom panel, supporting the evidence in Table 1 that these campuses are not systematically different from the campuses we observe with multiple principals over our period.

Reinforcing the limitations of pass rates, Figure 1 shows that the four campus accountability ratings categories do not systematically sort principals from low to high effectiveness. Though the distributions of principal fixed effects in the top panel reveal a higher concentration of ineffective principals and a lower modal effectiveness in the unacceptable category, differences are small between the acceptable and recognized ratings and virtually nonexistent between recognized and exemplary.[19] The excess mass of ineffective principals leading schools with an unacceptable rating suggests that the inclusion of required improvement provisions more closely related to student growth in the determination of the acceptable category helps to isolate schools with very low achievement gains into the unacceptable category. Yet, the distribution of principal effectiveness for those with unacceptable ratings also has a thicker right tail, illustrating that a substantial share of relatively effective principals lead schools that receive a rating of unacceptable.

In contrast, there are sharp differences in average pass rates across the rating categories, as seen in the bottom panel of Figure 1. Importantly, such differences appear even for the subset

[19] The p-values from Kolmogorov-Smirnov tests for the equality of distributions between consecutive ratings categories are all less than 0.01.

of principals who fall in the top quartile of the principal effectiveness distribution. Average pass rates for schools led by principals in the top quartile are 70 percent for schools rated unacceptable, 82 percent for schools rated acceptable, 90 percent for schools rated recognized, and 96 percent for schools rated exemplary. The rating-system reliance on pass rates rather than achievement growth clearly penalizes effective principals who work in schools serving predominantly lower achievers who struggle to earn a passing score. Finally, as shown in the middle panel, differences in the intermediate school-by-year value-added measure fall somewhere in between those for principal fixed effects and pass rates.

5. Campus rating effects on labor market outcomes

The first component of the empirical analysis investigates whether school ratings affect principal labor market outcomes. To identify the causal effects of ratings holding principal effectiveness and all else constant, we use regression discontinuity (RD) methods based on the school accountability system rules. We then provide complementary results from regressions of labor market outcomes on all measures of principal performance, recognizing that these estimates are more difficult to interpret as causal.

5.1 Regression discontinuity design approach

The RD exploits discontinuities in the probability of receiving a higher accountability rating based on the pass rate for the subgroup (i.e., demographic group-by-subject) that is likely to be binding for that campus and year. To identify this marginal subgroup for each rating boundary, we first determine the relevant pass rate threshold for each subgroup that meets applicable minimum size requirements. The threshold may be the statutory threshold, the required improvement threshold, or the exceptions threshold and is determined by the subgroup

pass rate in the prior year and whether exceptions are available. We then center subgroup pass rates around the relevant thresholds. The subgroup with the most negative (or least positive) centered pass rate is selected as the marginal subgroup for each rating category.[20] Running variable values greater than (less than) zero indicate that student performance was sufficient (not sufficient) to earn the higher rating.

We estimate our models using local linear regression with a triangular kernel on our main analytic sample. We use the structure of the accountability system and existing research to guide our choice of bandwidths. The distances between the statutory pass rates for the various ratings leads us to trim the samples to schools with running variable values within ten percentage points of the threshold in question. Virtually all schools within this range earn one of the two ratings around the threshold, while the fraction falling into a different rating category rises outside this range. We apply five alternative bandwidths to the trimmed sample—10, 7.5, 5, and 2.5 percentage points along with an optimal bandwidth described by Cattaneo and Vazquez-Bare (2016) and implemented by Calonico et al. (2017). We cluster standard errors by values of the running variable in all specifications.

Figure 2 illustrates the relationship between the probability of attaining the higher rating and the running variable for each of the school rating thresholds. Over the years 2001 to 2008, 17 percent of elementary schools were rated exemplary, 45 percent were rated recognized, 38 percent were rated acceptable, and only 1 percent received an unacceptable rating. The discontinuity is quite pronounced at all three thresholds between consecutive categories, though the bulk of the observations are at the threshold between acceptable and recognized. Though we fully incorporate the complex rules that change over time in the construction of the running

[20] Due to the required improvement provisions, Appendix Table A2 shows that the marginal student subgroup is also the lowest performing on the relevant subject only about two thirds of the time. This share fell further once the exceptions provisions were added in 2004.

variable, the presence of a small fraction (less than 2 percent) of schools whose ratings we do not correctly predict leads to a fuzzy design.[21] The corresponding first-stage estimates reported in Table 4 range from between 0.80 and 0.88 at the unacceptable-acceptable boundary, whereas they all exceed 0.96 at the recognized boundary and 0.91 at the exemplary boundary. Consequently, though we report intention-to-treat estimates for the labor market outcomes, local average treatment effect (LATE) estimates are similar in magnitude.

Any discontinuities in outcomes at the thresholds can be attributed to the receipt of the rating only if principals are unable to manipulate the running variable near the boundary and no other determinants of outcomes vary discontinuously at the boundary. Though others have shown that it is possible to manipulate pass rates by altering the test-taking pool (e.g., Cullen and Reback (2006), Figlio and Loeb (2011)), it is not feasible to do so precisely within the complex rating system. Once students sit for exams, they are scored and recorded centrally. Thus, variation in the subgroup pass rates in the neighborhood of the thresholds should be as good as random. Appendix Figure A1 shows the densities of acceptable, recognized and exemplary running variables. Formal statistical tests based on McCrary (2008) reject the null of no discontinuity for the recognized threshold, which we presume is due to chance.[22]

To explore further, we test whether there are any discontinuities in observable characteristics on either side of the ratings thresholds. We estimate a system of seemingly unrelated RD regressions using the principal and student characteristics shown in Table 1 as the dependent variables. Table 5 shows that almost none of these exhibits statistically significant

[21] One source of discrepancy is special accommodations that may be made in particular circumstances that are not elucidated in accountability manuals. Another is that it is possible for superintendents to appeal ratings, such as based on a consequential change in the coding of a student's race/ethnicity from prior years. Importantly, the underlying data reports are never altered even if an appeal is granted.

[22] The discontinuity estimates and associated standard errors for the optimal bandwidths from the first stages are 0.899 (0.543), 0.976 (0.197), and -0.019 (0.143) at the acceptable, recognized, and exemplary boundaries, respectively.

discontinuities at the ratings boundaries, and we fail to reject the null hypotheses that all coefficients are jointly equal to zero for the acceptable and exemplary boundaries, though we do reject for the recognized boundary. Similarly, for the multi-principal campuses, there are no statistically significant discontinuities in our estimates of principal effectiveness at the other two boundaries, though principals at campuses just meeting the recognized threshold are found to be more effective for the smallest two bandwidths.[23] Taken together, the validity tests suggest that the recognized boundary could be problematic, though in our context it is hard to imagine that this is due to manipulation of the running variable. Regardless, in the results that follow, we do not find any evidence of career impacts at this boundary, nor is this conclusion sensitive to the inclusion of covariates.

5.2 Regression discontinuity estimates of ratings impacts

We present results for three labor market outcomes: continuing as principal in the same school, salary growth, and changes in student composition. All three measures are defined based on positions held in year t+2 with ratings based on achievement in the spring of year t. Our measure of student composition is the normalized predicted pass rate, which weights demographic characteristics based on the relationship with the probability of passing. Since salary and student composition is observed only if the principal remains in the Texas Public Schools, we also examine ratings effects on sample attrition. The results for principal retention are depicted graphically in Figure 3, which plots the relationship between the running variable and the probability of retention around each of the ratings boundaries. A sharp contrast emerges between the sizable discontinuity at the unacceptable-acceptable boundary in the top panel and little if any jump at the two other thresholds in the bottom panels. The corresponding RD estimates reported in Table 6 confirm what is evident in the graphs. The estimates of

[23] See Appendix Table A3 for the estimates of discontinuities in principal value-added.

discontinuities associated with moving into the two higher rating categories are small and insignificant for all bandwidths, while the estimates show significant increases in retention for moving into the acceptable rating. For the optimal bandwidth, the estimate is 42.5 percentage points, which is a doubling relative to the baseline rate of retention for those campuses that do not escape the unacceptable rating. Accounting for the fuzziness of the design, the implied LATE estimate is about 20 percent larger.

An important issue concerns the channels that underlie the ratings effect on retention. The regulatory link between state-imposed sanctions and an unacceptable rating raises the possibility that the impetus for turnover is statutory requirements rather than administrator discretion. However, it takes two unacceptable ratings in successive years to trigger sanctions, so that schools not classified as unacceptable in the prior year are not at risk for sanctions. Table 7 first shows that only about 10 percent of campuses currently rated unacceptable were also rated unacceptable in the prior year. Second, the RD estimates for schools not previously rated unacceptable are significant and even larger than those estimated for the full sample, supporting the conclusion that school ratings provide information that influences discretionary personnel decisions.

Beyond continued employment, a principal's job can become better or worse in terms of salary and student composition. Figures 4 and 5 show the graphical evidence and Tables 8 and 9 present the estimates for the effects of school ratings on salary growth and the change in student composition, respectively. As with retention, there is no evidence of statistically or economically significant discontinuities at the recognized and exemplary boundaries. For the acceptable boundary, the pattern of estimates reveals improvements in salary on the order of 5-7 percent for the more narrow bandwidths. This naturally follows from the findings for job retention, since

many who are not retained move to lower-paying campus and district positions. But, this may also reflect larger raises for those who attain the acceptable rating. For student composition, though the point estimates are large and surprisingly negative for all bandwidths, none of the estimates is statistically significant.

Since the absence of compensation measures for principals who exit the Texas Public Schools could introduce selection bias, we also analyze the effect of ratings on the probability of exiting (see Appendix Figure A2 and Table A4). Crossing the acceptable threshold appears to be associated with an increase in the probability of exit, though the estimate is significant only for the smallest bandwidth considered and the magnitude is far smaller than for retention. Nevertheless, if the receipt of an acceptable rating provides public information that shifts the outside offer distribution to the right, the exclusion of leavers from the sample could bias downward the effects of an acceptable rating on salary.

Could these rating impacts have a rational basis? In the absence of frictions or information asymmetries in the principal labor market, no differences in outcomes for principals leading campuses just on opposite sides of the cutoff would have been expected. By the logic of the RD design, these principals are equally effective. Yet, receipt of the unacceptable rating has real consequences, perhaps due to public pressures on district administrators, pressure not brought to bear because of principals who fail to reach the recognized or exemplary thresholds. If reluctance to remove a principal introduces undesirable inertia that is overcome by the stigma of a failed rating, this unacceptable effect could constitute a second-best solution. However, the patterns in Figure 1 discussed above suggest that this is not the case. In fact, average effectiveness is if anything higher below the acceptable threshold than above it, and nearly one-quarter of principals in the vicinity of this threshold are in the top quartile of the overall

distribution of principal effectiveness.[24] Since leading a disadvantaged school puts a principal at greater career risk, the fact that the ratings largely fail to differentiate principals by quality may generate additional impediments to attracting and retaining effective principals in schools with low levels of achievement.

5.3 OLS estimates for broader sets of performance metrics

Although the accountability ratings do a poor job of differentiating according to estimates of principal effectiveness, districts may nevertheless favor more effective principals based on other achievement information made available through the system as well as from direct observations, parent reports, and other sources. While not necessarily causal, we can describe for principals at multi-principal campuses the pattern of associations between principal labor market outcomes and a range of alternative school and principal performance measures.

Tables 10-12 report OLS regression estimates for the probability of retention, change in log salary, and change in student composition. All specifications include ratings category indicators (acceptable is the excluded category), as well as district-by-year fixed effects and the principal and student characteristics shown in Table 1. What differs across the columns is the set of performance measures and whether or not school (or connected network) fixed effects are included. In each case, it is important to keep in mind that the interpretation varies along with the control set. For example, since all specifications include controls for student characteristics, the school average math and reading pass rate is implicitly case-mix adjusted. Further, when school fixed effects are included, variation in the pass rate is relative to other years at the same school, so it is adjusted for unobserved time-invariant school-level factors as well. Similarly, much of the differences between the principal and school-by-year fixed effects estimates is eroded in the

[24] See Appendix Table A5 for more details on the distribution of principal value-added by bandwidth around the acceptable threshold.

models that condition on school fixed effects.

The results in Table 10 provide little evidence of a significant relationship between the probability a principal continues in her position in year t+2 and any of the continuous performance measures. All coefficients are small and insignificant. Consistent with the regression discontinuity estimates the receipt of an unacceptable rating is negatively related to the probability of continuing in all specifications, regardless of whether the other performance measures are included or not. The estimated coefficient is significant at the one percent level in specifications without school or connected network fixed effects, highlighting the fact that principals in schools with unacceptable ratings are less likely to continue than their district colleagues who receive an acceptable or higher rating. The inclusion of the fixed effects at the finer campus or connected network level reduces the significance of the coefficient, which is unsurprising given the rarity of the receipt of an unacceptable rating.

In contrast to Table 10, the average pass rate is significantly related to salary growth in all specifications (Table 11), though the inclusion of school fixed effects increases the standard errors. None of the coefficients on school-by-year or principal fixed effects are significant, and the substitution of the more appropriate connected network fixed effects in place of school fixed effects roughly halves the principal fixed-effect coefficient. In terms of the ratings effects, only the exemplary rating coefficient is significant in specifications that include the average pass rate, and even it becomes small and insignificant following the inclusion of school fixed effects.

The significant role of the average pass rate diverges from the quite flat or even slightly negative slopes in the plots of salary growth against the marginal accountability group pass rate shown in Figure 4. Although the pass rate of the marginal accountability group is correlated with the average pass rate in math and reading, the association is dampened by the fact that science so

often serves as the marginal subject. Figure 6 depicts plots of the school average math and reading pass rate (top) and the case-mix adjusted pass rate (bottom), against the running variable around the three accountability thresholds. Over the range of twenty percentage points of the marginal pass rate, the average pass rate varies slightly less than 10 percentage points, and the adjusted average pass rate most relevant to the regression analysis varies by only around 3 percentage points. This weak association is consistent with district administrator knowledge and use of the average pass rate but little or no association with the marginal accountability group pass rate that serves as the running variable in the RD analysis. The fact that the inclusion of school fixed effects does not dampen the estimated effect on salary growth further suggests that administrators recognize the importance of student demographics in the determination of achievement.

The third outcome is the change in student composition, our proxy for working conditions. Table 12 reveals no significant relationship between accountability rating, pass rate, or principal fixed effect and this change. There is a statistically significant positive relationship between the change and the school-by-year fixed effect, though statistical significance does not survive the inclusion of the pass rate. This provides weak evidence that higher school value-added raises the probability of being rewarded by a job offer at a more advantaged school.

6. Inside-outside differences in the use of performance information

The decisions of both the current district and potential alternative employers determine labor market outcomes, but the current district is likely to have access to and to make use of more detailed information on job performance not readily available to others. This information asymmetry suggests that the probability of retention and compensation growth within the district

may be more strongly related to true effectiveness than would the transition to a desirable position outside of the district. On the other side, however, the current district may also face more pressure from less-informed interest groups to take action in response to the more salient information released to the public.

To compare within-district and out-of-district transitions, we use a composite success measure. This variable takes a value of one if a principal remains in her position, if salary growth exceeds median salary growth, or if the change in student composition exceeds the median change for all principals who remain in the Texas public schools in year t + 2. Among principals who remain in the same district, retention accounts for the vast majority of successes, while most district switchers with successful outcomes realize larger than median changes in salary. Overall, as shown in Table 2, we classify 85.0 percent of principal-years in our main analysis sample as being associated with successful labor market outcomes two years later. The residual categories of principals who are identified as not being successful include principals who move to lower paying and less appealing positions as well as principals who exit the Texas Public School system. This latter group is quite heterogeneous. Individuals who exit may be switching to private schools, changing occupations, dropping out of the labor force or retiring – though we have reduced the incidence of retirement by restricting the sample to principals with no more than 25 years of total experience in the system.

Table 13 presents the RD estimates of the effects of ratings for any success (top panel) and then separately for within district success (middle panel) and new district success (bottom panel). Consistent with the retention findings, crossing the acceptable boundary significantly raises the probability of within district success. There is also weak evidence that crossing the recognized boundary improves this same outcome, though the estimates are only statistically

significant for the wider bandwidths. While imprecise, none of the estimates for new district success are statistically significantly different from zero. Importantly, lumping together failures and successes in the null category in the RD specifications with binary dependent variables complicates interpretation of the RD estimates. For example, most of those who do not enjoy new district success are actually classified as having within-district success.

Therefore, we supplement these with non-causal multinomial logit regressions that divide principals into within district successes, new district successes, and failures regardless of destination. These regressions have additional performance metrics including the principal fixed effect, and thus the sample is restricted to campuses with multiple principals in the sample.

The estimates in Table 14 show that the probability of new district success is significantly related just to the pass rate, while the probability of within district success is significantly related to the rating, pass rate, and the principal fixed effect. In terms of the use of information, the findings suggest that the publicly reported pass rate is related to the probabilities of success both within- and out-of-district. Interestingly, the ratings and the proxy for effectiveness, which fall on opposite sides of the spectrum in terms of salience and information content, only appear to matter within district. This is consistent with the district both possessing more information about true productivity and being subject to more pressure from the public when a school receives an unfavorable rating.

7. Conclusions

This analysis illustrates the effects of an accountability system that reports both detailed performance data and ratings based on that information. First, the regression discontinuity design results provide strong causal evidence that a failure to achieve an acceptable rating significantly

reduces the incumbent principal's probability of job retention and salary growth. Although principals in the bottom quartile of effectiveness as measured by principal fixed effects on achievement growth are overrepresented in schools rated unacceptable, these schools are as likely to be led by top quartile principals as schools rated more highly. Perhaps unexpectedly, receiving the lowest rating only has a negative association with outcomes in the current district, raising the possibility that the current employer is more susceptible to pressure from imperfectly informed parties such as local school boards.

Second, the estimates show that the associations between principal labor market outcomes and measures of school and principal performance are much stronger for measures aligned with the accountability system. Significant associations emerge between job continuation and salary growth on the one hand and the average pass rate and ratings on the other but not between the labor market outcomes and measures more closely related to value-added to achievement growth. The positive relationship between the probability of labor market success within the current district and the principal fixed effect is an exception and consistent with the notion that the current employer has more detailed information on job performance.

Taken as a whole, the results suggest that aligning the performance evaluation system better with student learning could improve the quality and allocation of school leaders. Because non-school factors account for a large portion of the variation in the pass rates currently reported, selection into a school serving higher-SES students may actually be more beneficial to a principal's labor market prospects than raising the quality of instruction. It might also be difficult to attract principals to a school that is likely to receive a low rating due to limited family resources, for fear of being penalized for any failure. Principals in high poverty schools, which are likely to have low baseline pass rates, may be especially disadvantaged in the principal labor

market through these channels.

Our findings for Texas are emblematic of the many other school accountability systems across the U.S. modeled after its system. Thus, the set of fundamental design issues we highlight have wide applicability to schools and districts in other states. Moreover, the increasing use of outcome-based incentives to reduce healthcare spending suggests that these concerns extend far beyond the education sector.

References

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. "High Wage Workers and High Wage Firms." *Econometrica* 67, no. 2 (March): 251-333.

Abowd, John M., Robert H. Creecy, and Francis Kramarz. 2002. "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data." Longitudinal Employer-Household Dynamics Technical Papers 2002-06, Center for Economic Studies, U.S. Census Bureau.

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. Cambridge, MA: National Bureau of Economic Research (November).

Bates, Michael. 2016. "Public and Private Learning in the Market for Teachers: Evidence from the Adoption of Value-Added Measures." (mimeo) Riverside, CA: University of California at Riverside (December 2).

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 W17803. Cambridge, MA: National Bureau of Economic Research (January).

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2017. "rdrobust: Software for regression-discontinuity designs." *Stata Journal* 17, no. 2: 372-404.

Cattaneo, Matias D., and Gonzalo Vazquez-Bare. 2016. "The Choice of Neighborhood in Regression Discontinuity Designs." *Observational Studies* 2: 134-146.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104, no. 9 (September): 2593-2632.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2016. "Using Lagged Outcomes to Evaluate Bias in Value-Added Models." *American Economic Review* 105, no. 5 (May): 393-99.

Chiang, Hanley, Stephen Lipscomb, and Brian Gill. 2016. "Is School Value Added Indicative of Principal Quality?" *Education Finance and Policy* 11, no. 3 (Summer): 283–309.

Coelli, Michael, and David A. Green. 2012. "Leadership Effects: School Principals and Student Outcomes." *Economics of Education Review* 31, no. 1 (February): 92-109.

Craig, Steven G., Scott A. Imberman, and Adam Perdue. 2015. "Do Administrators Respond to Their Accountability Ratings? The Response of School Budgets to Accountability Grades." *Economics of Education Review* 49 (December): 55-68.

Cullen, Julie B., and Michael J. Mazzeo. 2008. "Implicit Performance Awards: An Empirical Analysis of the Labor Market for Public School Administrators." University of California, San Diego (December).

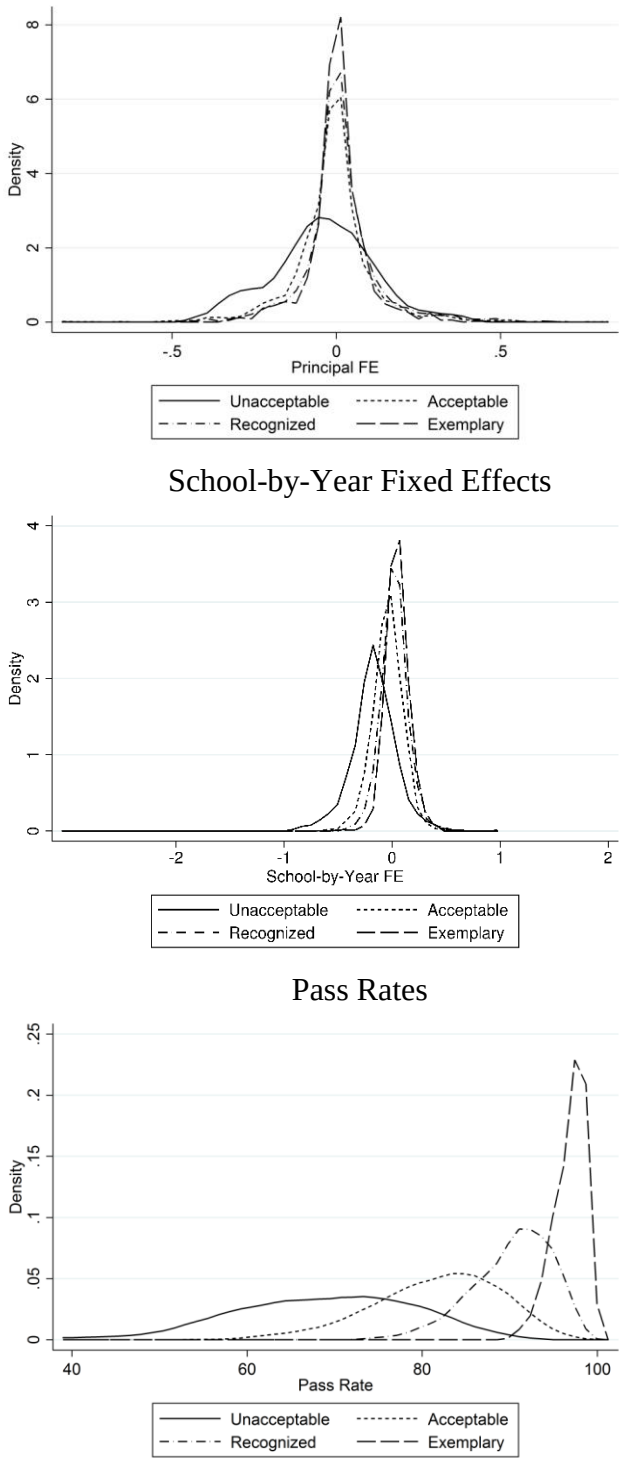
Cullen, Julie Berry , and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming under a Performance Accountability System." In *Improving School Accountability*, edited by Timothy J. Gronberg and Dennis W. Jansen: 1-34.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2016. "School Accountability, Postsecondary Attainment, and Earnings." *Review of Economics and Statistics* 98, no. 5: 848-862.

Dhuey, Elizabeth, and Justin Smith. 2014. "How Important are School Principals in the Production of Student Achievement?" *Canadian Journal of Economics/Revue canadienne d'économique* 47, no. 2 (May): 634-663.

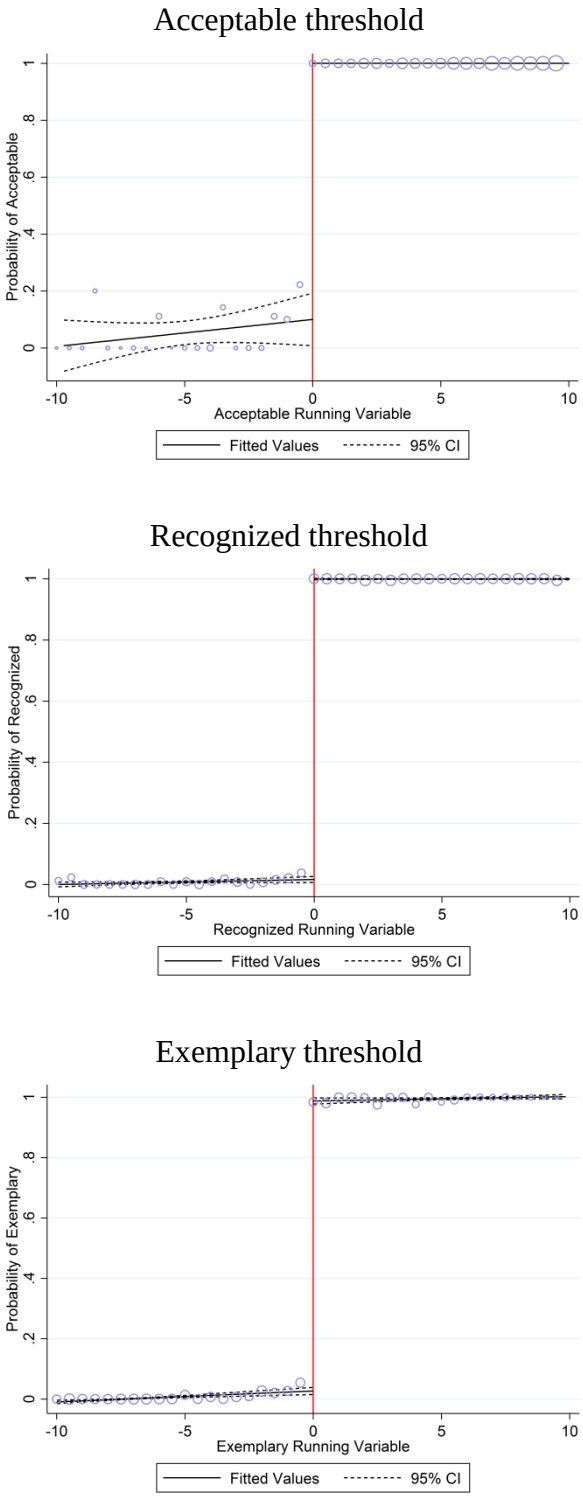
Dizon-Ross, Rebecca. Forthcoming. "How Does School Accountability Affect Teachers? Evidence from New York City" *Journal of Human Resources*.
Figlio, David N., and Maurice E. Lucas. 2004. "What's in a Grade? School Report Cards and the Housing Market" *American Economic Review* 94: 591-604.
Figlio, David N., and Susanna Loeb. 2011. "School Accountability." In *Handbook of the Economics of Education, Vol. 3*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland: 383-421.
Grissom, Jason A., and Brendan Bartanen. Forthcoming. "Principal Effectiveness and Principal Turnover." *Educational Evaluation and Policy Analysis.*
Demetra Kalogrides, and Susanna Loeb. 2015. "Using Student Test Scores to Measure Principal Performance." *Educational Evaluation and Policy Analysis* 37, no. 1 (March): 3-28.
Grissom, Jason A., Demetra Kalogrides, and Susanna Loeb. 2015. "Using Student Test Scores to Measure Principal Performance." *Educational Evaluation and Policy Analysis* 37, no. 1 (March): 3-28.
Hanushek, Eric A., John F. Kain, and Steve G. Rivkin. 2004. "Why Public Schools Lose Teachers." *Journal of Human Resources* 39, no. 2 (Spring): 326-354.
Helal, Mike, and Michael Bernard Coelli. 2016. "How Principals Affect Schools." Melbourne Institute Working Paper No. 18/16. Melbourne: University of Melbourne (June 1).
Laing, Derek, Steven G. Rivkin, Jeffrey C. Schiman, and Jason Ward. 2016. "Decentralized Governance and the Quality of School Leadership." NBER Wroking Paper No. 22061. Cambridge, MA: National Bureau of Economic Research (March).
Loeb, Susanna, Demetra Kalogrides, and Eileen Lai Horng. 2010. "Principal Preferences and the Uneven Distribution of Principals Across Schools." *Educational Evaluation and Policy Analysis* Vol. 32, no. 2 (June): 205–229.
McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142, no. 2: 698-714.
Miller, Ashley. 2013. "Principal Turnover and Student Achievement." *Economics of Education Review* 36(October): 60-72.
Reback, Randall. 2008. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics* 92, no. 5-6: 1394-1415.
Sokol, Marlene. 2018. State Grades Push Hillsborough into an Unexpected Wave of Principal Transfers. *Tampa Bay Times*, July 5.

Figure 1. Principal fixed effect and school pass rate densities, by accountability rating



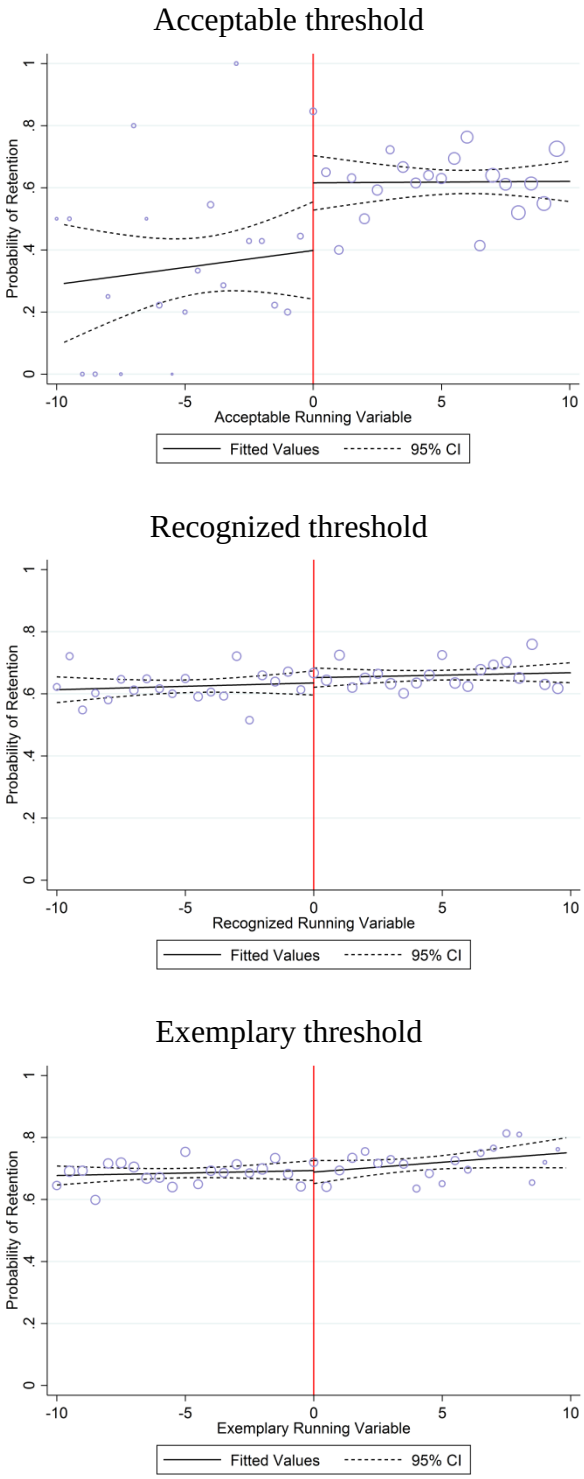
Notes: In all three panels, the sample of campuses is restricted to those served by multiple principals over our sample period, and the unit of observation is a campus-by-year. The principal and school-by-year fixed effects are estimated from models of student achievement gains, as described in the text. The pass rate is the average across math and reading by campus and year.

Figure 2. First stage probability of attaining the higher rating, by accountability rating threshold



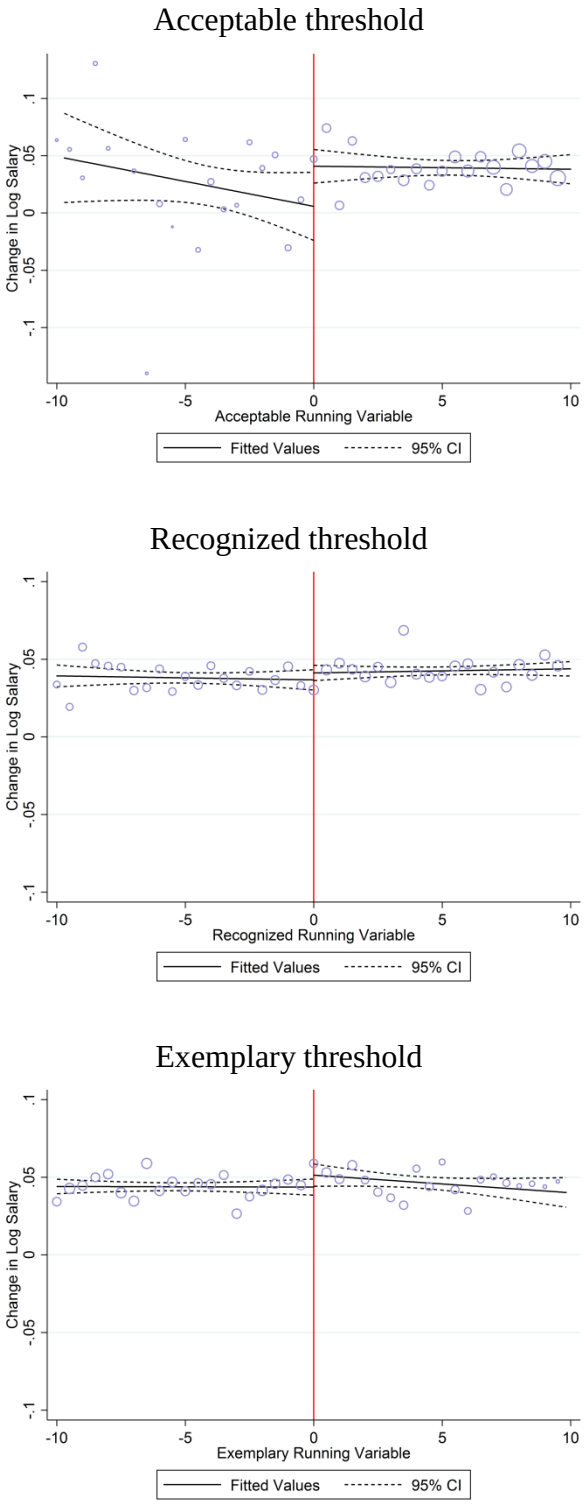
Notes: In each panel, the running variable is the difference between the pass rate for the marginal student subgroup and the relevant pass rate threshold. The bin width is 0.5 percentage points. Points are weighted by bin size (i.e., number of campus-by-year observations) and are comparable within ratings categories but not across.

Figure 3. Probability of retention, by accountability rating threshold



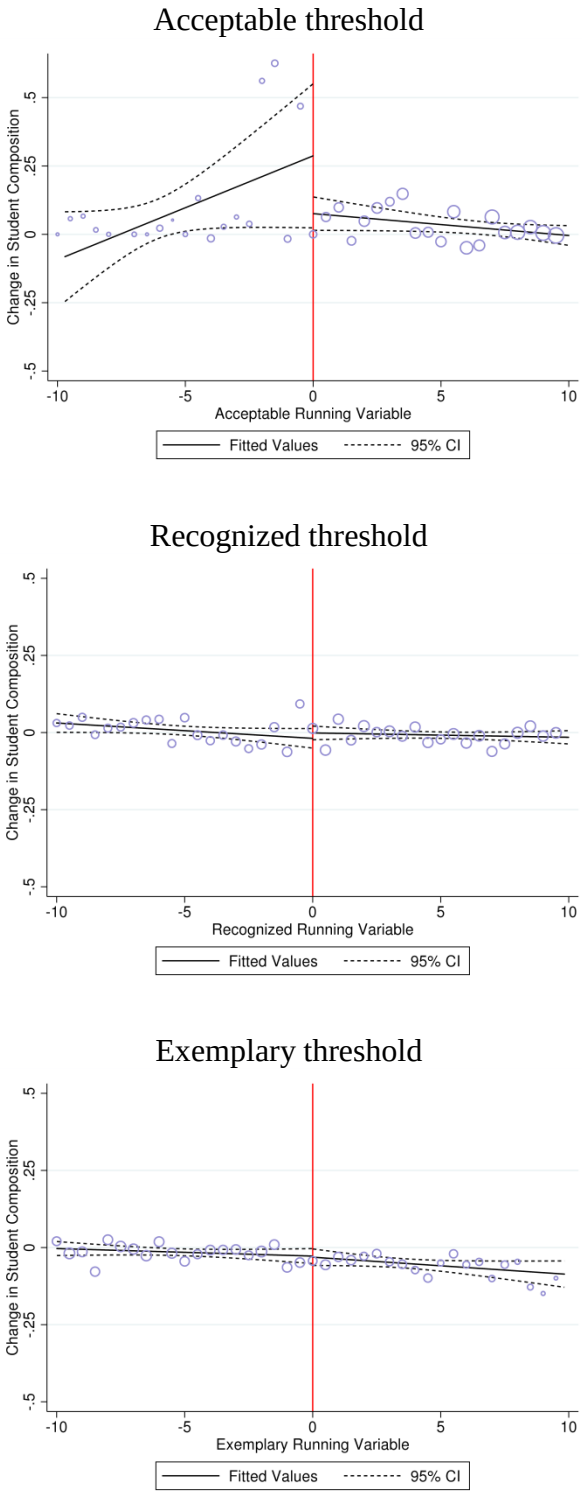
Notes: Retention is defined as continuing in the same principal position in academic year t+2, with the campus rating realized at the end of academic year t. For other details, see notes to Figure 2.

Figure 4. Salary growth, by accountability rating threshold



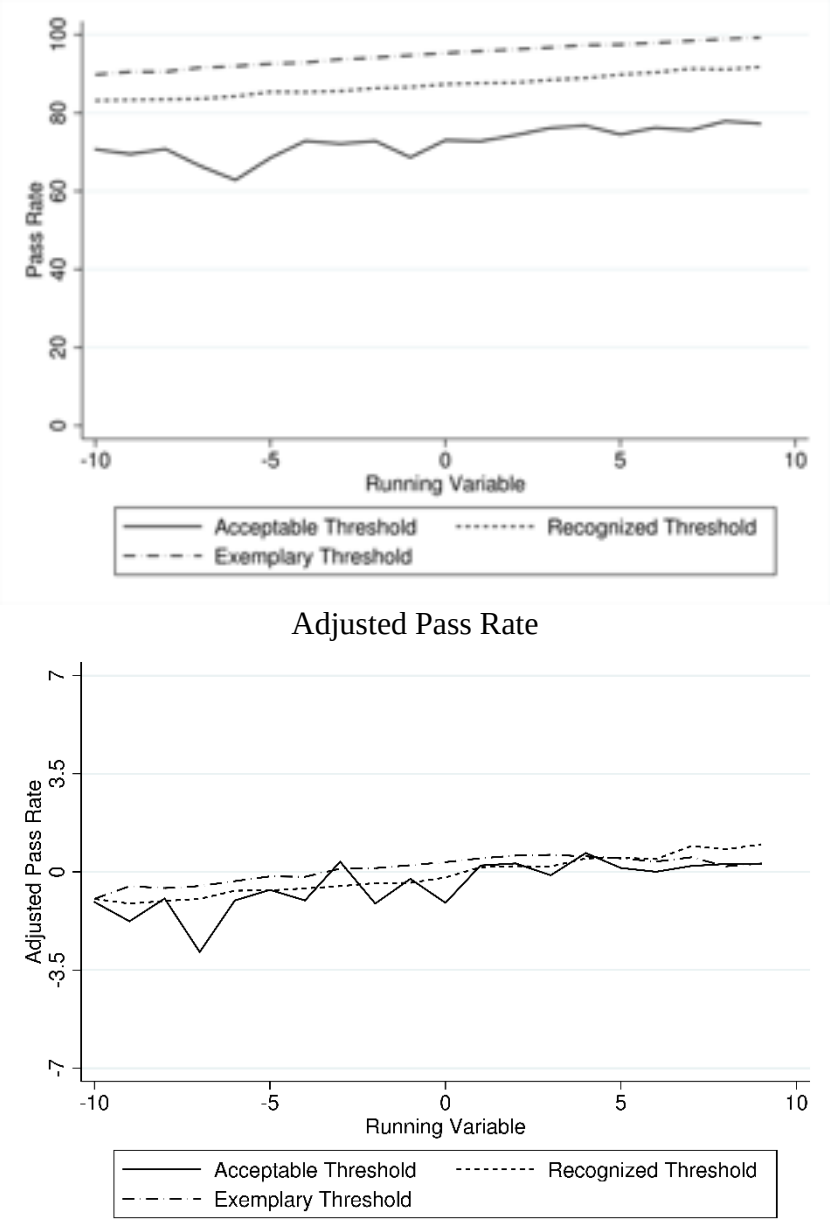
Notes: Salary growth is measured by the change in the log (real $2003) total pay between academic years t+2 and t, with the campus rating realized at the end of academic year t. For other details, see notes to Figure 2.

Figure 5. Change in student composition, by accountability rating threshold



Notes: Student composition is proxied by a predicted achievement index based on student characteristics, as described in the text. The change in student composition is between academic years t+2 and t, with the campus rating realized at the end of academic year t. For other details, see notes to Figure 2.

Figure 6. Pass rate and adjusted pass rate, by accountability rating threshold



Notes: These panels show how two campus performance measures vary across values of the running variables for the three accountability ratings thresholds. The performance measure in the top panel is the campus average pass rate across math and reading. In the bottom panel, this pass rate is adjusted for student demographics by taking residuals from OLS regressions of the pass rate on year fixed effects and the campus student characteristics shown in Table 1. The running variables are the differences between the pass rates for the marginal student subgroups and the relevant pass rate thresholds. The sample of campuses is restricted to those served by multiple principals over our sample period, and the unit of observation is a campus-by-year. The lines are generated by smoothed plots of average values by each percentage point of the running variable.

Table 1. Summary statistics for principal, campus, and student characteristics across samples

Variable	All	Experience <25 years	Tenure ≥2 years at campus	Multi-principal campuses
	(1)	(2)	(3)	(4)
Principal characteristics				
Male	0.281	0.290	0.284	0.292
Black	0.109	0.101	0.100	0.100
Hispanic	0.224	0.214	0.212	0.215
White	0.663	0.680	0.684	0.681
Other race/ethnicity	0.004	0.004	0.004	0.004
Below Master's degree	0.055	0.072	0.072	0.081
Master's degree	0.904	0.895	0.895	0.888
Doctorate degree	0.040	0.033	0.033	0.031
2 or fewer years tenure	0.272	0.329	0.274	0.327
3 years tenure	0.160	0.191	0.207	0.246
4 or more years tenure	0.568	0.479	0.519	0.427
Total years of experience	22.49	17.53	17.64	17.01
Principal salary				
Total pay (2003 dollars)	$66,478	$64,089	$64,078	$63,639
Student test performance				
Average math/reading pass rate	88.02	88.01	88.13	87.51
Math pass rate	87.07	87.03	87.17	86.47
Reading pass rate	88.85	88.87	88.96	88.42
Campus accountability rating				
Unacceptable	0.012	0.012	0.012	0.013
Acceptable	0.381	0.384	0.377	0.401
Recognized	0.438	0.441	0.446	0.439
Exemplary	0.169	0.163	0.165	0.147
Campus student characteristics				
Male	0.514	0.514	0.515	0.514
Black	0.142	0.135	0.134	0.138
Hispanic	0.466	0.459	0.459	0.475
White	0.361	0.375	0.376	0.357
Other race/ethnicity	0.031	0.031	0.031	0.030
Economically disadvantaged	0.601	0.595	0.594	0.614
Title 1 participant	0.722	0.727	0.725	0.750
Limited English proficient	0.210	0.207	0.207	0.219
Special education	0.107	0.107	0.107	0.107
Gifted and talented	0.061	0.059	0.059	0.058
Mid-year school mover	0.062	0.062	0.062	0.062
N (campus-by-year)	20,045	12,296	11,351	7,653

Notes: Summary statistics for all elementary campus-by-year observations for the years 2001 to 2008 (excluding 2003) are reported in column 1. Means for all campus-by-year observations with principals that have less than 25 years of total experience in the Texas Public Schools are reported in column 2. Means for all campus-by-year observations with principals that have less than 25 years of total experience and have been principal at the current campus for at least two years are reported in column 3. Column 4 further restricts the sample to exclude any campus led by only one principal for at least two years from 2001 to 2008.

Table 2. Summary statistics for principal labor market outcomes, by analysis sample

Variable	Experience <25 and tenure ≥2 years	Multi-principal campuses
	(1)	(2)
Outcomes for all principals		
Retained	0.652	0.561
Moved within the same district	0.199	0.252
Successful move within district	0.150	0.190
Successful move with high salary growth	0.129	0.166
Unsuccessful move within district	0.049	0.062
Moved to a new district	0.069	0.090
Successful move to a new district	0.048	0.063
Successful move with high salary growth	0.038	0.050
Unsuccessful move to a new district	0.021	0.027
Exit Texas public schools	0.081	0.097
N (school-by-year)	11,351	7,653
N (principals)	4,222	3,285
N (schools)	3,251	2,174
Outcomes for principals who remain in the system		
Salary growth	0.039	0.044
	(0.081)	(0.091)
Change in student composition	-0.012	-0.015
	(0.335)	(0.388)
N (school-by-year)	10,437	6,913
N (principals)	3,934	3,021
N (schools)	3,157	2,116

Notes: Statistics for all campus-by-year observations with principals that have less than 25 years of total experience in the Texas Public Schools and have been principal at the current campus for at least two years are reported in column 1. Column 2 further restricts the sample to exclude any campus led by only one principal for at least two years from 2001 to 2008. Standard deviations for continuous variables are shown in parentheses below the means. The outcomes are based on academic year t+2, with the campus rating realized at the end of academic year t. Retention is defined as continuing in the same principal position in academic year t+2. Successful moves are defined as realizing above median gains in log (real $2003) salary or student composition between t and t+2, relative to all principals who remain in the system. Student composition is proxied by a predicted achievement index based on student characteristics, as described in the text. Exiting Texas public schools is defined as not holding any position within the system in academic year t+2.

Table 3. Correlations between principal fixed effects and other school performance measures

	Principal FE	School-by-year FE	Mean school-by-year FE	Pass rate	Adjusted pass rate
	(1)	(2)	(3)	(4)	(5)
Multi-principal campuses					
Principal FE	1.000				
School-by-year FE	0.192	1.000			
Mean school-by-year FE	0.257	0.526	1.000		
Pass rate	0.099	0.322	0.276	1.000	
Adjusted pass rate	0.152	0.420	0.287	0.615	1.000
N			7,653		
Single-principal campuses					
Principal FE	NA				
School-by-year FE	NA	1.000			
Mean school-by-year FE	NA	0.556	1.000		
Pass rate	NA	0.307	0.285	1.000	
Adjusted pass rate	NA	0.377	0.217	0.554	1.000
N			3,698		

Notes: In the top panel, the sample is restricted to observations from campuses served by more than one principal during the course of our sample period. The sample in the bottom panel is observations from those campuses served by only one principal. In both cases, principals are required to have at least two years of tenure in their current position and 25 or fewer years of total experience in Texas public schools. The variables in the first three columns are estimates of student achievement value-added, where student achievement is defined to be the average of math and reading z-scores. Column 1 is our estimate of principal productivity from specifications following equation (2) that include principal and school fixed effects, and then demean the estimated principal fixed effects by the average within each connected network. In column 2, value-added is proxied by the concurrent school-by-year fixed effects from specifications that replace principal and school fixed effects with school-by-year fixed effects, as per equation (1). The variable in column 3 averages the school-by-year fixed effects across principals' terms. Column 4 is the current campus average pass rate across math and reading, while column 5 adjusts for student demographics by taking residuals from OLS regressions of the pass rate on year fixed effects and the campus student characteristics shown in Table 1.

Table 4. First stage probability of attaining the higher rating, by accountability rating threshold

	Bandwidth				
	10	7.5	5	2.5	Optimal
Acceptable	0.882***	0.861***	0.833***	0.796***	0.817***
	(0.058)	(0.068)	(0.082)	(0.113)	(0.093)
Mean	0.062	0.064	0.067	0.095	0.079
N	760	497	299	140	222
Recognized	0.978***	0.975***	0.972***	0.960***	0.960***
	(0.006)	(0.008)	(0.010)	(0.017)	(0.017)
Mean	0.009	0.009	0.012	0.016	0.016
N	5,613	4,252	2,879	1,458	1,457
Exemplary	0.954***	0.948***	0.936***	0.911***	0.921***
	(0.010)	(0.012)	(0.016)	(0.024)	(0.021)
Mean	0.008	0.011	0.017	0.028	0.023
N	4,935	3,925	2,690	1,419	1,767

Notes: Each cell shows the estimated discontinuity at the threshold from a separate local linear regression with a triangular kernel, with the associated standard errors clustered by values of the running variable (for which the precisions is 1/100 of a percentage point) shown in parentheses. The mean of the dependent variable is shown for the subset of principals within the bandwidth sample receiving the lower rating. The bandwidths vary across the columns as indicated by the column headers. Optimal bandwidths are estimated using the optimal MSE bandwidth selector discussed by Cattaneo and Vazquez-Bare (2016) and Calonico et al. (2017). Optimal bandwidths for Acceptable, Recognized and Exemplary thresholds are 3.82, 2.49, and 3.18, respectively. *** p<0.01, ** p<0.05, * p<0.10

Table 5. Balance tests for principal and campus student characteristics, by rating threshold

Variable	Acceptable	Recognized	Exemplary
	(1)	(2)	(3)
Principal characteristics and salary			
Male	-0.021	-0.051	-0.009
	(0.132)	(0.049)	(0.042)
Black	0.004	0.040	0.025
	(0.138)	(0.032)	(0.023)
Hispanic	0.060	-0.025	0.036
	(0.132)	(0.045)	(0.035)
Master's degree	0.059	0.034	0.014
	(0.095)	(0.032)	(0.027)
Doctorate degree	-0.059	-0.037*	-0.017
	(0.073)	(0.017)	(0.016)
Years of tenure	0.191	0.247	-0.017
	(0.556)	(0.241)	(0.208)
Total years of experience	-2.746	1.042	0.577
	(1.674)	(0.556)	(0.480)
Log total pay (2003 dollars)	-0.012	0.009	-0.006
	(0.035)	(0.013)	(0.012)
Campus student performance and characteristics			
Average math/reading pass rate	2.515	0.647	-0.034
	(2.461)	(0.548)	(0.201)
Male	-0.005	-0.002	-0.001
	(0.007)	(0.003)	(0.002)
Black	-0.007	0.027	-0.004
	(0.079)	(0.019)	(0.013)
Hispanic	0.014	-0.020	-0.017
	(0.088)	(0.033)	(0.029)
White	0.004	-0.015	0.023
	(0.055)	(0.030)	(0.028)
Economically disadvantaged	0.009	0.015	-0.009
	(0.047)	(0.025)	(0.026)
Title 1 participant	-0.010	0.055	-0.009
	(0.062)	(0.042)	(0.045)
Limited English proficient	0.019	-0.004	-0.009
	(0.071)	(0.023)	(0.017)
Special education	0.008	0.003	0.010**
	(0.009)	(0.004)	(0.004)
Gifted and talented	0.002	-0.009	0.005
	(0.012)	(0.005)	(0.006)
Mid-year school mover	-0.019	-0.006	-0.007*
	(0.012)	(0.004)	(0.004)
Chi-square statistic	17.243	33.428	25.299
p-value	0.573	0.021	0.151
N	222	1,457	1,767

Notes: Each cell reports the coefficient and standard error (in parentheses) from a separate seemingly unrelated regression discontinuity regression for the principal and student characteristics shown. The regressions are local linear regressions with triangular weights, and the bandwidths are set equal to the optimal bandwidths determined by the first stages for each threshold. Chi-square statistics and their associated p-values are reported for the test of the null hypothesis that all coefficients in the column are jointly equal to zero. *** p<0.01, ** p<0.05, * p<0.10

Table 6. Regression discontinuity estimates of the impact of attaining the higher rating on the probability of principal job retention, by rating threshold

	Bandwidth				
	10	7.5	5	2.5	Optimal
Acceptable	0.249**	0.270**	0.365***	0.467***	0.425***
	(0.097)	(0.109)	(0.129)	(0.172)	(0.142)
Mean	0.354	0.383	0.387	0.333	0.413
N	760	497	299	140	222
Recognized	0.011	0.019	0.024	0.021	0.021
	(0.026)	(0.030)	(0.036)	(0.051)	(0.051)
Mean	0.625	0.628	0.630	0.627	0.631
N	5,613	4,252	2,879	1,458	1,457
Exemplary	0.005	0.014	0.018	0.038	0.024
	(0.027)	(0.031)	(0.037)	(0.052)	(0.046)
Mean	0.685	0.689	0.694	0.689	0.693
N	4,935	3,925	2,690	1,419	1,767

Notes: Retention is defined as continuing in the same principal position in academic year t+2, with the campus rating realized at the end of academic year t. For other details, see notes to Table 4.

Table 7. Regression discontinuity estimates of the impact of attaining the higher rating on the probability of principal job retention, acceptable rating threshold by prior year rating status

	Bandwidth				
	10	7.5	5	2.5	Optimal
Previously rated unacceptable	-0.364	-0.478	-0.247	-0.799	-0.305
	(0.323)	(0.352)	(0.513)	(0.606)	(0.529)
Mean	0.500	0.500	0.500	0.500	0.571
N	48	35	24	12	19
Not previously rated unacceptable	0.303***	0.340***	0.424***	0.559***	0.488***
	(0.104)	(0.119)	(0.139)	(0.184)	(0.153)
Mean	0.333	0.366	0.373	0.306	0.393
N	712	462	275	128	203

Notes: The top panel restricts the sample to campuses near the acceptable threshold that were rated unacceptable in the prior year, while the bottom panel only includes those that were not previously rated unacceptable. For other details, see notes to Table 4.

Table 8. Regression discontinuity estimates of the impact of attaining the higher rating on salary growth, by rating threshold

	Bandwidth				
	10	7.5	5	2.5	Optimal
Acceptable	0.033*	0.033	0.048*	0.065**	0.054**
	(0.019)	(0.022)	(0.026)	(0.033)	(0.028)
Mean	0.023	0.015	0.017	0.019	0.018
N	628	411	238	111	181
Recognized	0.005	0.002	-0.001	-0.005	-0.005
	(0.004)	(0.005)	(0.006)	(0.008)	(0.008)
Mean	0.038	0.037	0.038	0.037	0.036
N	4,970	3,760	2,546	1,285	1,284
Exemplary	0.010**	0.011**	0.011	0.007	0.007
	(0.005)	(0.006)	(0.007)	(0.010)	(0.008)
Mean	0.044	0.044	0.043	0.044	0.041
N	4,479	3,574	2,443	1,291	1,605

Notes: Salary growth is measured by the change in the log (real $2003) total pay between academic years t+2 and t, with the campus rating realized at the end of academic year t. For other details, see notes to Table 4.

Table 9. Regression discontinuity estimates of the impact of attaining the higher rating on the change in student composition, by rating threshold

	Bandwidth				
	10	7.5	5	2.5	Optimal
Acceptable	-0.268	-0.310	-0.357	-0.168	-0.319
	(0.167)	(0.192)	(0.225)	(0.306)	(0.245)
Mean	0.136	0.155	0.192	0.312	0.208
N	628	411	238	111	181
Recognized	0.017	0.009	-0.026	-0.076	-0.076
	(0.023)	(0.028)	(0.036)	(0.055)	(0.055)
Mean	0.005	0.000	-0.009	-0.012	-0.013
N	4,970	3,760	2,546	1,285	1,284
Exemplary	-0.005	-0.004	0.000	0.005	0.000
	(0.020)	(0.023)	(0.029)	(0.041)	(0.036)
Mean	-0.015	-0.016	-0.022	-0.027	-0.023
N	4,479	3,574	2,443	1,291	1,605

Notes: Student composition is proxied by an index of predicted achievement based on student characteristics, as described in the text. The change in student composition is between academic years t+2 and t, with the campus rating realized at the end of academic year t. For other details, see notes to Table 4.

Table 10. Ordinary least squares estimates of the relationships between school performance metrics and job retention

	Dependent variable: Indicator for job retention in t+2										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
School-by-year FE					-0.008	-0.096				-0.024	-0.134
					(0.056)	(0.086)				(0.066)	(0.119)
Principal FE			0.004	-0.027			-0.001	-0.058	0.004		
			(0.079)	(0.211)			(0.075)	(0.223)	(0.108)		
Pass rate	0.001	0.002					0.001	0.002	0.000	0.001	0.003
	(0.002)	(0.004)					(0.002)	(0.004)	(0.003)	(0.003)	(0.005)
Unacceptable	-0.171***	-0.108	-0.181***	-0.120	-0.182***	-0.130*	-0.171***	-0.110	-0.100	-0.171***	-0.113
	(0.066)	(0.080)	(0.064)	(0.077)	(0.066)	(0.076)	(0.066)	(0.080)	(0.097)	(0.066)	(0.079)
Recognized	0.035*	0.034	0.041**	0.040	0.042*	0.045	0.035*	0.034	0.039	0.035*	0.035
	(0.021)	(0.032)	(0.023)	(0.030)	(0.023)	(0.031)	(0.021)	(0.032)	(0.028)	(0.021)	(0.031)
Exemplary	0.021	0.029	0.031	0.038	0.032	0.047	0.021	0.029	0.028	0.022	0.033
	(0.040)	(0.071)	(0.035)	(0.067)	(0.036)	(0.069)	(0.040)	(0.071)	(0.056)	(0.040)	(0.071)
School FE	No	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes
Network FE	No	No	No	No	No	No	No	No	Yes	No	No
R-squared	0.414	0.690	0.414	0.690	0.414	0.690	0.414	0.690	0.358	0.414	0.691

Notes: Each column reports results from a separate ordinary least squares regression for the sample of 7,653 campus-by-year observations for principals at multi-principal campuses. The dependent variable is an indicator for whether or not the principal is retained. Retention is defined as continuing in the same principal position in academic year t+2, with the campus performance measures realized at the end of academic year t. The estimated coefficients on the included performance metrics are shown with standard errors clustered by district in parentheses. Acceptable is the excluded rating category. All specifications include district-by-year fixed effects and control for the principal and student characteristics shown in Table 1. Some specifications then also include either a full set of school fixed effects or of fixed effects for networks of schools that are connected by principal mobility as indicated in the bottom rows of the table. *** p<0.01, ** p<0.05, * p<0.10

Table 11. Ordinary least squares estimates of the relationships between school performance metrics and salary growth

	Dependent variable: Salary growth from t to t+2										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
School-by-year FE					0.018	0.026				-0.005	0.007
					(0.014)	(0.022)				(0.013)	(0.021)
Principal FE			0.032*	0.087			0.025	0.062	0.030		
			(0.017)	(0.053)			(0.015)	(0.051)	(0.020)		
Pass rate	0.002***	0.002**					0.002***	0.002*	0.002***	0.002***	0.002**
	(0.000)	(0.001)					(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
Unacceptable	-0.020	-0.031*	-0.038**	-0.039**	-0.036**	-0.040**	-0.020	-0.029	-0.025	-0.020	-0.030
	(0.014)	(0.018)	(0.015)	(0.019)	(0.015)	(0.019)	(0.014)	(0.019)	(0.018)	(0.014)	(0.019)
Recognized	0.003	0.001	0.014***	0.006	0.013***	0.006	0.003	0.000	0.001	0.003	0.001
	(0.003)	(0.006)	(0.004)	(0.006)	(0.004)	(0.006)	(0.003)	(0.006)	(0.005)	(0.003)	(0.006)
Exemplary	0.014**	0.007	0.030***	0.014	0.028***	0.014	0.013**	0.007	0.009	0.014**	0.007
	(0.006)	(0.010)	(0.006)	(0.009)	(0.006)	(0.009)	(0.006)	(0.010)	(0.007)	(0.006)	(0.010)
School FE	No	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes
Network FE	No	No	No	No	No	No	No	No	Yes	No	No
R-squared	0.459	0.701	0.455	0.700	0.454	0.700	0.460	0.702	0.614	0.459	0.701

Notes: Each column reports results from a separate ordinary least squares regression for the sample of 6,913 campus-by-year observations for the subset of principals at multi-principal campuses that remains in the public school system in academic year t+2. The dependent variable is salary growth. Salary growth is measured by the change in the log (real $2003) total pay between academic years t+2 and t, with the campus performance measures realized at the end of academic year t. For other details, see notes to Table 10. *** p<0.01, ** p<0.05, * p<0.10

Table 12. Ordinary least squares estimates of relationships between school performance metrics and the change in student composition

	Dependent variable: Change in student composition from t to t+2										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
School-by-year FE					0.142**	0.104				0.184**	0.146
					(0.057)	(0.090)				(0.073)	(0.099)
Principal FE			0.011	0.284			0.017	0.331	-0.000		
			(0.066)	(0.263)			(0.068)	(0.279)	(0.080)		
Pass Rate	-0.002	-0.002					-0.002	-0.003	-0.001	-0.003	-0.004
	(0.002)	(0.003)					(0.002)	(0.003)	(0.002)	(0.002)	(0.003)
Unacceptable	-0.014	-0.008	0.002	0.019	0.021	0.020	-0.014	0.000	-0.015	-0.008	-0.002
	(0.133)	(0.170)	(0.133)	(0.177)	(0.135)	(0.177)	(0.133)	(0.170)	(0.156)	(0.134)	(0.171)
Recognized	0.000	0.028	-0.009	0.017	-0.019	0.016	0.000	0.028	0.003	-0.002	0.028
	(0.018)	(0.024)	(0.018)	(0.025)	(0.018)	(0.026)	(0.018)	(0.024)	(0.022)	(0.018)	(0.024)
Exemplary	-0.011	-0.018	-0.025	-0.035	-0.043	-0.038	-0.011	-0.021	-0.057	-0.018	-0.023
	(0.030)	(0.047)	(0.028)	(0.045)	(0.030)	(0.049)	(0.030)	(0.047)	(0.043)	(0.030)	(0.048)
School FE	No	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes
Network FE	No	No	No	No	No	No	No	No	Yes	No	No
R-squared	0.459	0.701	0.455	0.700	0.454	0.700	0.460	0.702	0.614	0.459	0.701

Notes: Each column reports results from a separate ordinary least squares regression for the sample of 6,913 campus-by-year observations for the subset of principals at multi-principal campuses that remains in the public school system in academic year t+2. The dependent variable is the change in student composition between academic years t and t+2, with the campus performance measures realized at the end of academic year t. Student composition is proxied by an index of predicted achievement based on student characteristics, as described in the text. For other details, see notes to Table 10. *** p<0.01, ** p<0.05, * p<0.10

Table 13. Regression discontinuity estimates of the impact of attaining the higher rating on composite labor market success, by rating threshold and employment location

	Bandwidth				
	10	7.5	5	2.5	Optimal
Any success					
Acceptable	0.110	0.102	0.164	0.169	0.195
	(0.100)	(0.114)	(0.135)	(0.185)	(0.148)
Mean	0.584	0.585	0.600	0.643	0.667
N	760	497	299	140	222
Recognized	0.038*	0.042*	0.036	0.016	0.016
	(0.020)	(0.023)	(0.026)	(0.036)	(0.036)
Mean	0.824	0.826	0.820	0.807	0.809
N	5,613	4,252	2,879	1,458	1,457
Exemplary	0.010	0.007	-0.006	-0.023	-0.026
	(0.018)	(0.019)	(0.022)	(0.027)	(0.025)
Mean	0.877	0.878	0.873	0.881	0.874
N	4,935	3,925	2,690	1,419	1,767
Within district success					
Acceptable	0.217**	0.222*	0.289**	0.224	0.306*
	(0.105)	(0.120)	(0.142)	(0.194)	(0.158)
Mean	0.487	0.489	0.493	0.476	0.540
N	760	497	299	140	222
Recognized	0.044*	0.047*	0.048*	0.036	0.036
	(0.023)	(0.025)	(0.029)	(0.039)	(0.039)
Mean	0.770	0.770	0.768	0.757	0.760
N	5,613	4,252	2,879	1,458	1,457
Exemplary	-0.005	-0.010	-0.028	-0.035	-0.043
	(0.021)	(0.024)	(0.028)	(0.035)	(0.032)
Mean	0.834	0.837	0.832	0.841	0.834
N	4,935	3,925	2,690	1,419	1,767
New district success					
Acceptable	-0.107	-0.120	-0.125	-0.055	-0.111
	(0.068)	(0.075)	(0.081)	(0.102)	(0.086)
Mean	0.097	0.096	0.107	0.167	0.127
N	760	497	299	140	222
Recognized	-0.005	-0.005	-0.013	-0.020	-0.020
	(0.012)	(0.013)	(0.015)	(0.022)	(0.022)
Mean	0.054	0.050	0.052	0.050	0.049
N	5,613	4,252	2,879	1,458	1,457
Exemplary	0.015	0.017	0.022	0.012	0.017
	(0.012)	(0.014)	(0.017)	(0.023)	(0.021)
Mean	0.043	0.041	0.041	0.041	0.040
N	4,935	3,925	2,690	1,419	1,767

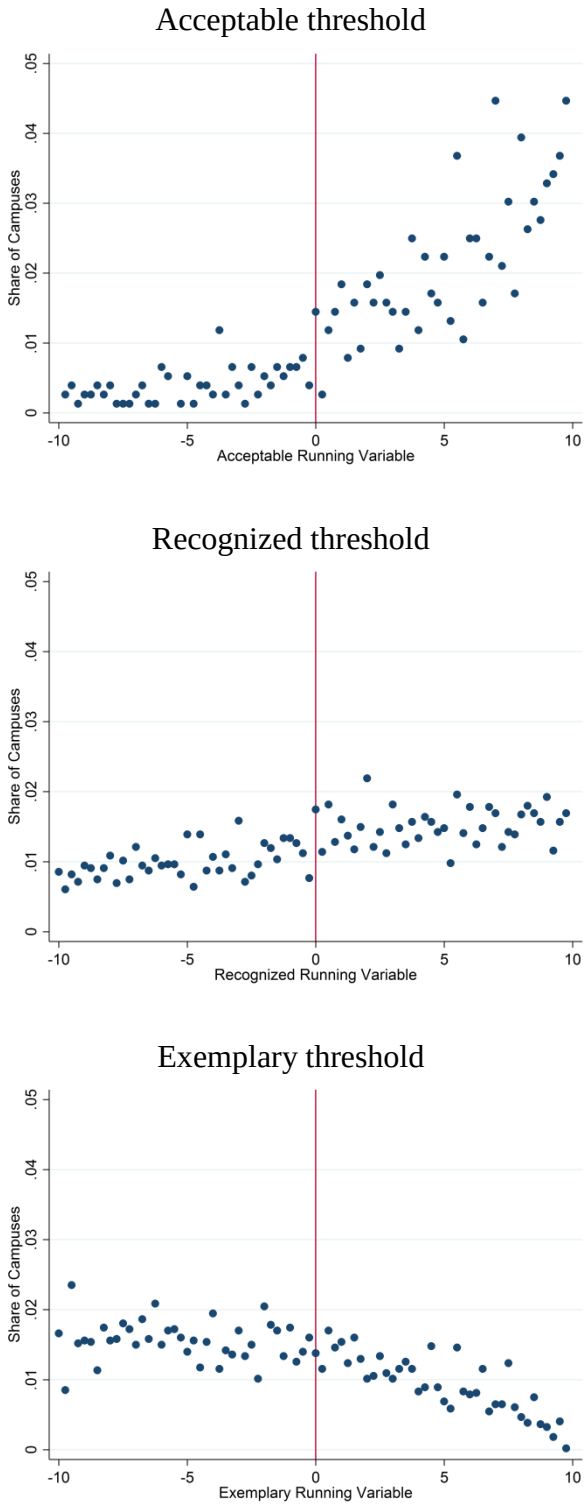
Notes: Composite principal labor market success is defined to include being retained at the same campus or realizing above median gains in log salary or student composition between academic years t+2 and t, with the campus rating realized at the end of academic year t. For other details, see notes to Table 4.

Table 14. Mulitnomial logit estimates of relationships between school performance metrics and composite labor market success within district and out of district

	(1)	(2)	(3)	(4)	(5)
Within district success					
School-by-year FE			0.504*		-0.001
			(0.275)		(0.319)
			[0.037]		[-0.024]
Principal fixed effect		0.799**		0.600*	
		(0.360)		(0.355)	
		[0.115]		[0.093]	
Pass rate	0.037***			0.035***	0.037***
	(0.008)			(0.008)	(0.009)
	[0.004]			[0.004]	[0.004]
Unacceptable	-1.017***	-1.344***	-1.303***	-1.011***	-1.015***
	(0.235)	(0.216)	(0.234)	(0.231)	(0.237)
	[-0.172]	[-0.221]	[-0.218]	[-0.171]	[-0.172]
Recognized	0.272***	0.468***	0.449***	0.267***	0.271***
	(0.094)	(0.091)	(0.095)	(0.094)	(0.095)
	[0.044]	[0.067]	[0.067]	[0.043]	[0.044]
Exemplary	0.337**	0.629***	0.591***	0.335**	0.337***
	(0.158)	(0.160)	(0.161)	(0.159)	(0.159)
	[0.047]	[0.085]	[0.083]	[0.047]	[0.048]
New district success					
School-by-year FE			1.136**		0.694
			(0.513)		(0.546)
			[0.034]		[0.031]
Principal fixed effect		0.144		-0.095	
		(0.731)		(0.733)	
		[-0.021]		[-0.025]	
Pass rate	0.039***			0.039***	0.032**
	(0.015)			(0.015)	(0.016)
	[0.000]			[0.001]	[0.000]
Unacceptable	0.261	-0.085	0.033	0.261	0.270
	(0.422)	(0.397)	(0.410)	(0.421)	(0.427)
	[0.050]	[0.042]	[0.048]	[0.049]	[0.050]
Recognized	-0.068	0.148	0.075	-0.068	-0.075
	(0.160)	(0.165)	(0.167)	(0.160)	(0.161)
	[-0.012]	[-0.009]	[-0.012]	[-0.012]	[-0.013]
Exemplary	0.159	0.466	0.334	0.160	0.137
	(0.300)	(0.299)	(0.306)	(0.300)	(0.303)
	[-0.004]	[0.000]	[-0.005]	[-0.004]	[-0.006]

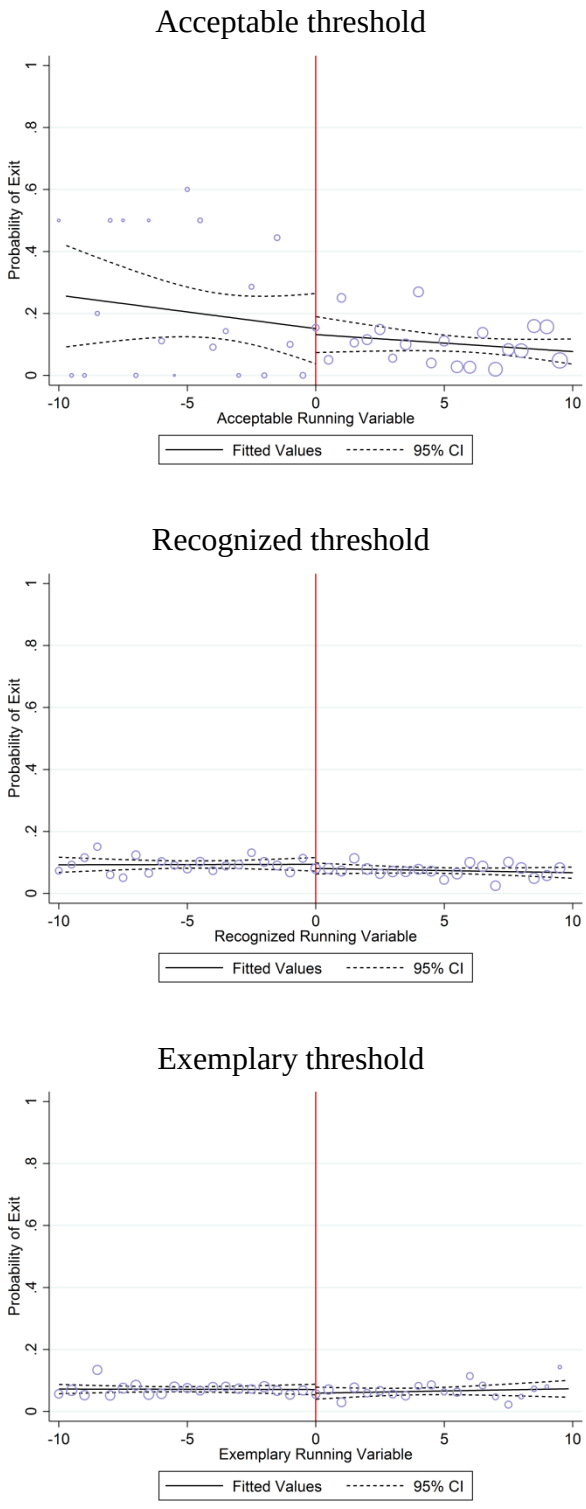
Notes: Each column reports multinomial logit coefficient estimates from a separate specification for the multi-principal sample (N=7,653). Standard errors clustered by district are reported in parentheses. Average marginal effects (or differences in probabilities of outcomes for the binary ratings) are reported in brackets. Acceptable is the excluded rating category. The three outcomes modeled are i) achieving success within the same district, ii) achieving success in another district, and iii) neither, where neither is the base outcome and success is defined as in Table 13. All specifications include district and year fixed effects and control for the principal and student characteristics shown in Table 1. District and year fixed effects are included in lieu of district-by-year fixed effects since estimation of the more saturated models fails to converge. *** p<0.01, ** p<0.05, * p<0.10

Appendix Figure A1. Running variable density, by accountability rating threshold



Notes: The bin width is 0.25 percentage points. In each case, the running variable is the difference between the required pass rate and the pass rate of the binding subgroup.

Appendix Figure A2. Probability of exiting Texas public schools, by accountability rating threshold



Notes: Exiting is defined as not holding any position within the Texas public school system in academic year t+2, while the rating is realized at the end of academic year t. For other details, see notes to Figure 2.

Appendix Table A1. Marginal student subgroup shares, by accountability rating

	Any subgroup	Marginal student subgroup				
		All students	White	Black	Hispanic	Disadv.
Acceptable						
Math	0.049	0.000	0.000	0.039	0.005	0.005
Reading	0.302	0.034	0.001	0.104	0.093	0.070
Science	0.439	0.079	0.001	0.092	0.118	0.149
Writing	0.187	0.053	0.005	0.013	0.046	0.070
Recognized						
Math	0.248	0.008	0.004	0.074	0.059	0.103
Reading	0.195	0.005	0.001	0.034	0.064	0.091
Science	0.416	0.091	0.010	0.022	0.122	0.171
Writing	0.122	0.029	0.009	0.008	0.026	0.050
Exemplary						
Math	0.275	0.021	0.020	0.045	0.070	0.119
Reading	0.296	0.017	0.014	0.031	0.091	0.143
Science	0.249	0.126	0.038	0.004	0.042	0.039
Writing	0.163	0.053	0.043	0.004	0.022	0.041

Notes: Each cell shows the share of marginal subgroups falling in a specific category for the 10 percentage point bandwidth sample around the accountability threshold indicated in the row heading. The marginal subgroup is the one that determines the running variable for the regression discontinuity analysis, and is the one with the most negative (or least positive) gap between the required pass rate and the subgroup pass rate. Not shown are the cases (about 2% for each category) where the marginal student subgroup is special education students taking alternate non-grade level assessments (SDAA and SDAA II) offered between 2004 and 2007.

Appendix Table A2. Marginal subgroup lowest performing shares, by accountability rating and time period

	Share of marginal subgroups that are also the lowest performing subgroup in the marginal subject	
	Pre-2004	Post-2004
Acceptable	0.672	0.584
Recognized	0.688	0.601
Exemplary	0.622	0.574

Notes: Each cell shows the share of marginal subgroups that are also the lowest performing in the marginal subject for the 10 percentage point bandwidth sample around the accountability threshold indicated in the row heading.

Appendix Table A3. Regression discontinuity estimates of the impact of attaining the higher rating on principal value-added, by rating threshold

	Bandwidth				
	10	7.5	5	2.5	Optimal
Acceptable	-0.010	0.010	0.042	0.129	0.078
	(0.045)	(0.055)	(0.072)	(0.111)	(0.088)
Mean	-0.032	-0.020	-0.018	-0.007	-0.015
N	579	385	227	108	164
Recognized	0.009	0.009	0.016	0.044**	0.044**
	(0.009)	(0.011)	(0.013)	(0.019)	(0.019)
Mean	-0.002	-0.001	-0.000	0.004	0.003
N	3,813	2,918	1,974	1,003	1,003
Exemplary	0.006	0.008	0.011	0.014	0.009
	(0.008)	(0.009)	(0.011)	(0.017)	(0.015)
Mean	0.013	0.013	0.014	0.010	0.009
N	3,122	2,491	1,671	872	1,088

Notes: Principal value-added is estimated from specifications following equation (2) that include principal and school fixed effects, and then demean the estimated principal fixed effects by the average within each connected network. The sample excludes any campus led by only one principal for at least two years from 2001 to 2008. For other details, see notes to Table 4.

Appendix Table A4. Regression discontinuity estimates of the impact of attaining the higher rating on the probability of exiting the Texas public schools, by rating threshold

	Bandwidth				
	10	7.5	5	2.5	Optimal
Acceptable	0.027	0.057	0.053	0.164**	0.087
	(0.065)	(0.072)	(0.077)	(0.075)	(0.075)
Mean	0.195	0.191	0.200	0.167	0.143
N	760	497	299	140	222
Recognized	-0.013	-0.012	-0.009	-0.014	-0.014
	(0.015)	(0.017)	(0.021)	(0.030)	(0.030)
Mean	0.093	0.091	0.093	0.098	0.096
N	5,613	4,252	2,879	1,458	1,457
Exemplary	-0.011	-0.013	-0.006	0.001	-0.003
	(0.014)	(0.016)	(0.018)	(0.024)	(0.022)
Mean	0.071	0.071	0.072	0.069	0.069
N	4,935	3,925	2,90	1,419	1,767

Notes: Exiting is defined as not holding any position within the Texas public school system in academic year t+2, while the rating is realized at the end of academic year t. For other details, see notes to Table 4.

Appendix Table A5. Principal fixed effect averages and quartile shares by bandwidth around the acceptable boundary

Principal fixed effect	Bandwidth		
	10	5	Optimal
Average below	-0.032	-0.018	-0.015
N	90	60	48
Average above	-0.033	-0.033	-0.034
N	489	167	116
Lowest quartile	0.417	0.423	0.427
Second quartile	0.183	0.167	0.159
Third quartile	0.169	0.176	0.165
Top quartile	0.231	0.233	0.250
N	579	227	164

Notes: Percentiles of principal fixed effects are calculated using all observations in the sample described in the last column of Table 1. The 25th, 50th, and 75th percentiles are -0.044, -0.001, and 0.041, respectively. The optimal bandwidth is 3.82 percentage points.