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]

May 2021

ABSTRACT

In many settings, leaders are evaluated in contexts where complexities of production processes and conflicting pressures from interest groups pose challenges to performance evaluation. In education, school accountability systems assemble rich data and report both categorical rating and the underlying student pass rates that determine it, permitting direct investigation of how different information affects labor market outcomes of school leaders. Applying regression discontinuity methods that by design hold effectiveness constant, we find sizable positive impacts on Texas elementary school principal retention and salaries for crossing the unacceptable-acceptable boundary but not for crossing higher ratings cutoffs. The apparent information breakdown that leads to the unequal treatment of equals at the lowest boundary could raise the distribution of principal quality through disproportionate departures of less effective school leaders. However, there is substantial overlap in principal value-added distributions across rating categories, and failure to cross the acceptable threshold does not lead to future improvements in school performance. Supplementary analysis suggests that the labor market penalty to leading a school that receives the lowest rating is confined to the current district, where the stigma of a low rating is likely to be greatest.

[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] Julie Berry Cullen is a professor of economics at University of California, San Diego and research associate at NBER

[3] Eric A. Hanushek is Paul and Jean Hanna Senior Fellow in Education, Hoover Institution, Stanford University; senior research fellow at University of Texas at Dallas; and research associate at NBER

[4] Gregory Phelan is assistant professor of economics at Kennesaw State University

[5] Steven G. Rivkin is professor of economics at University of Illinois at Chicago, senior research fellow at University of Texas at Dallas, and research associate at NBER; sgrivkin@uic.edu

1. Introduction

Leadership quality is frequently cited as key to organizational success in both the public and private sectors, though the lack of competitive pressures on public sector organizations and their leaders has raised particular concerns. 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 quality and reducing inefficiencies. Importantly, the information collected under school accountability facilitates better measurement of educator productivity including that of leaders than is possible in most public or even private sector settings. In settings with complex production processes and competing interest groups, there are likely to be divergences between actual and measured or perceived effectiveness. The way these divergences play out in the education sector can provide lessons for leader evaluations more broadly.

In this paper, we study how different types of information about student performance affect labor market outcomes for Texas public elementary school principals. We focus specifically on the implications of a categorical rating system under which schools are placed into four separate categories: unacceptable, acceptable, recognized, or exemplary. As these types of rating systems have proliferated, at issue is how the crude ratings matter in and of themselves to school leader careers. Texas offers a noteworthy context to study these impacts since principals are afforded substantial scope as managers and face a large labor market. The structure of Texas public school governance is typical of public schools across the country and shares similarities to large private-sector corporations and not-for-profit organizations. School district superintendents function similarly to CEOs. Though they retain the authority over principal hiring, retention, and salaries, they do not operate in a vacuum. Rather they report directly to

school boards and almost certainly respond to feedback from parents and others in the community. Consequently, accountability systems may influence decisions about principal employment and compensation through multiple channels, with the various stakeholders likely relying on different types of performance information.

In Texas, the categorical school ratings and underlying student pass rates that determine them are reported to the public, while system personnel additionally have access to student-level longitudinal data that can be used to produce estimates of achievement growth as well as other information on staff performance. The multi-dimensional structure of school performance information raises three basic questions about the structure of rating systems that we investigate in this paper. First, does the crude nature of the categorical school ratings independently affect labor market outcomes for principals even when the underlying performance information on which they are based is readily available? Second, do any responses to the categorical ratings improve leadership quality and school performance? And third, do current and alternative employers appear to rely on different types of information in personnel decisions?

We use regression discontinuity design (RD) methods to identify the causal impacts of reaching higher school rating categories on principal labor market outcomes. We find no significant differences in the probability of principal job retention or salary growth for moving into the two highest rating categories, but there are large and significant discontinuities for moving out of the lowest category. Barely missing an acceptable rating is associated with a 38 percentage point decline in the likelihood of retention in the current job and a 6 percent loss in salary for principals.

The strikingly different outcomes for equally productive principals across the unacceptable-acceptable boundary are consistent with alternative underlying mechanisms. One

possibility is that a departing principal is not forcibly dismissed but that the stigma of leading a failing school or the imposition of external requirements related to the unacceptable rating make continuing in the job unattractive. An alternative though not mutually exclusive possibility supported by survey evidence is that principals' job security depends on avoiding a low rating (Toenjes and Garst 2000). This could be the case even when superintendents and other central administrators access the more detailed performance information if these administrators come under pressure from parents, the press, school board members or other interest groups that focus on the cruder ratings in forming opinions about the state of a school or effectiveness of a principal.[6] Ultimately, a voluntary departure to avoid being associated with a failing school or a district decision to remove a principal because her school receives an unacceptable rating each reflect an information failure. Were all stakeholders and potential employers to possess full information on performance, the principal whose pass rate falls just below the accountability cutoff would be viewed and treated identically as the principal whose pass rate falls at the cutoff.

Nonetheless, whether the lower retention rate for principals in schools that receive a low rating raises the future quality of leadership depends on the behaviors of district administrators and how principal effectiveness is distributed across schools. For example, administrators may be generally reluctant to fire poor-performing principals, and this reluctance may be overcome by public pressures when a school is rated as failing. In this case the removal of an ineffective principal at a school that receives the low rating might lift the distribution of principal quality even though an equally ineffective principal retains her job.

In our setting in which pass rates rather than achievement growth are the primary

[6] 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 pressure, the Hillsborough superintendent 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).

determinants, receipt of an unacceptable rating does not appear to be an instrument for raising leadership quality by triggering the replacement of ineffective principals. We explore implications for principal effectiveness using proxies based on school value-added, which we show have out-of-sample predictive power for student achievement growth. Mapping out the distributions of these measures by rating categories illuminates the general failure of the Texas accountability system to discriminate by principal effectiveness. Principals in the bottom quartile of effectiveness as measured by achievement value-added are overrepresented in schools rated unacceptable, but principals in these unacceptable schools are also as likely to be in the top quartile as principals in schools rated more highly. These patterns are consistent with the results in the RD analyses, where we find that failure to cross the acceptable threshold does not significantly increase either future school value-added or pass rates.

Supplementary analyses distinguish transitions to a new district from transitions within the current district. RD and descriptive multinomial logit results show that an unacceptable rating adversely affects the probability of labor market success within the same district but does not reduce the probability of a transition to a job with higher pay or better working conditions in another district. This pattern is primarily driven by the effect on continuing in the current job and underscores the possibility that pressure from interest groups enters into superintendent and/or principal decisions on continuation.

Our study first and foremost contributes to the literature analyzing the causal impacts of receiving a low accountability rating. In notable early research on rating effects, Figlio and Lucas (2004) raised the concern that discrete classifications convey misinformation to the public. The authors find that home prices respond to school grades after conditioning on the variables used to construct the grades, and follow-on research finds negative impacts of low ratings on private

donations to schools (Figlio and Kenny 2009). Others have since explored how receipt of a low rating affects school operations and educators. In the case of schools, Chiang (2009) and Rouse et al. (2013) find that the receipt of a low grade alters resource use and instructional practices. In terms of teachers, Feng, Figlio, and Sass (2018) find that teachers in Florida – particularly high-value-added teachers – are more likely to leave schools that receive a failing grade. On the other hand, Dizon-Ross (2020) finds the surprising result in New York City that teacher turnover falls and the quality of entrant teachers improves after a school's receipt of a low grade, which she speculates may be due to improvements in job desirability since the effects are concentrated in schools led by principals that teachers rate as strong leaders. As far as we know, ours is the first study to analyze causal rating impacts on the principal labor market.[7]

The threat of a low rating may have different impacts than receipt of a low rating, and a number of studies investigate the effects of this type of accountability pressure.[8] Particularly relevant to our work is the analysis in Deming et al. (2016) which is also set in Texas. The authors find that pressure to avoid classification as unacceptable among at-risk high schools has positive effects on student achievement and longer-term outcomes. Our findings reveal no evidence that the actual receipt of an unacceptable rating confers differential benefits to achievement. Complicating the picture, Rockoff and Turner (2010) find immediate achievement gains following receipt of a low rating under the New York City system, which directly factors learning gains into the determination of ratings. The divergent effects across channels and

[7] Surprisingly, few studies have linked school 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 (2019) find that 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.
[8] See Figlio and Loeb (2011) for an overview of the evidence on the broad range of impacts of school accountability, including those that are counterproductive.

settings illustrates the complexity of accountability effects, particularly where the basis of the ratings is weakly related to school effectiveness.

Our study also contributes to the broader literature on leader productivity and compensation. Studies of private sector executives find substantial variation in manager effects and positive relationships between firm performance under a manager and the manager's compensation.[9] Importantly, these studies take caution not to interpret the variation across managers as necessarily reflecting differences in the causal effects of managers on firm outcomes due to the possibility of omitted variables bias. Time-varying factors raise particular concerns, and Lazear et al. (2015) are able to draw stronger inferences about supervisor productivity differences due to the availability of extensive information about other factors of production, determinants of supervisor assignments, and consideration of match effects.

Finally, our findings speak to the literature on labor mobility and wage growth and how these relate to unobserved and observed predictors of worker productivity. Under canonical models (e.g., Farber and Gibbons 1996, Schönberg 2007), the availability of detailed school achievement information reduces information asymmetries between current and outside employers.[10] In this case, just crossing a rating boundary would not be predicted to improve outcomes, and the relationship between outcomes and productivity should be similar for current and outside employers. That we find nonproductive responses to barely missing an acceptable rating and find these only in the current district underscores that district employers are not

[9] Bertrand and Schoar (2003) create proxies for CEO performance from regressions of firm outcomes on executive and firm fixed effects and total firm assets. The positive relationships between compensation and the fixed effects show that firms pay a premium for managers who are associated with better firm outcomes. Graham et al. (2011) find similar associations from models that include additional time-varying firm factors, and add the interpretative caveat that a more positive fixed effect may not only reflect higher ability that is rewarded in the labor market but also other factors including better negotiating skills.

[10] As might be expected given the limited availability of classroom-level performance metrics, Bates (2020) uncovers evidence for meaningful asymmetric information about productivity in the teacher labor market.

independent actors and must incorporate the judgements of a collection of more and less well-informed stakeholders.

2. Institutional background

The principal labor market in Texas is likely more fluid than in other states. Texas is one of the few states that prohibits public employees from entering into collective bargaining. School principals and teachers generally serve under term contracts that 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 is two years in most Texas districts (Bryant 2017). Principals are required to have two years of classroom teaching experience in addition to completing a Master's degree from 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. Salaries for principals are set by the superintendent of the school district, subject to approval of the school board.

As school leaders, principals have extensive responsibilities ranging from hiring and managing teachers to setting school budgets and policies. In Texas, principal performance is evaluated annually by district administrators. State code recommends standards for evaluating principals on specific indicators in the areas of human capital development, instructional leadership, executive leadership, school culture, and strategic operations. 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 one of the first states to employ test-based school accountability, implementing 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 testing regime.

In our analysis, we study elementary-school principals over the 2001 to 2008 school years. The 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.[11] 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). Although the administration of math and reading tests in consecutive grades makes it possible to observe achievement growth in these core subjects, the accountability system did not incorporate learning gains and remained focused until recently on achievement levels.

Over our study period, the mapping from test scores to 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)[12] 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.

[11] Although data for 2009-2011 are available, a new measure was added to the accountability system that we were unable to successfully incorporate into our regression discontinuity 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.

[12] Economically disadvantaged students are those eligible to receive free or reduced-price lunch based on family income and federal poverty guidelines.

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.[13] 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 exception provisions for campuses to be elevated to acceptable, recognized, and exemplary ratings: a specified number of subject-by-subgroups (determined by campus size) 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. As we show below, despite these efforts to build in features related to progress, it is usually the lowest performing subject-by-subgroup that is the decisive factor in the determination of the school rating.

For Texas elementary schools, 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 improvement, 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 face graduated stages of intervention. In the first year, the principal must work with an external review team to develop and implement a school improvement plan. Receipt of an unacceptable rating in two consecutive years initiates the imposition of sanctions that become progressively more severe for each additional year the school fails to reach an acceptable rating until, after five years, there are requirements to replace staff.[14] Over our sample period, it is rare for elementary

[13] In this case, the prior year pass rate is adjusted to account for any change in the cutoff score for passing.
[14] Though the state ratings are the ones that continue to be emphasized in annual school report cards, schools are also classified by whether they meet adequate yearly progress (AYP) starting in 2004 when the federal No Child Left Behind policy became effective. The federal rules require adjustments to some of the indicators and

schools to be rated unacceptable for even two consecutive years, so that there are no mechanical impacts on principal retention.

The detailed and summary information about school performance 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.

3. Data on principal labor market outcomes and school performance

We study labor market outcomes for elementary-school principals for the period 2001 through 2008, excluding 2003 since school ratings were not assigned. We rely on matched panels of staff and students from restricted-use data assembled by the University of Texas at Dallas Texas Schools Project.[15] The personnel database provides annual information on background characteristics, 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 incorporate data on school characteristics and performance from the publicly available Texas Academic Excellence Indicator System. These comprehensive annual reports include accountability ratings, pass rates for all students and

consideration of additional subgroups, leading to little overlap between the set of schools identified as failing under the state and federal systems. 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 schools 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.

[15] See https://tsp.utdallas.edu for more details on the Texas Schools Project.

subgroups, 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, schools on average experience 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 realities that 2-year contracts are the norm and that principals have limited initial control over staff composition as predecessor decisions persist in the short run. Table 1 shows the effects of these sample restrictions. Starting from the full sample of school-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 schools.

When constructing measures of effectiveness, we also omit the final year of a principal's spell because of evidence in Miller (2013) showing sizeable achievement declines in the last year. Therefore, only spells that last at least three years are included in the estimation of principal effectiveness, which includes 8,166 principal-by-year observations representing 3,248 unique principals. The final column in Table 1 shows that these longer-tenure principals and their schools again appear to be typical on other dimensions.

We use three primary measures of labor market outcomes for principals: job retention,

salary, and student case-mix. Job retention and salary are common measures of market outcomes but student case-mix is less standard and merits discussion. Past evidence highlights the influence of student and family inputs on the working conditions for teachers and administrators and the possibility for these to lead to compensating salary differentials.[16] To create a summary measure of student advantage as a proxy for this aspect of working conditions, we regress school-by-year 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 schools serving tested grades over our sample period.[17] 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 school-years. A high value for the index indicates that the student body is likely to be high-achieving. Since salary and case-mix are observed only for those principals who remain in Texas public schools, we also investigate exits from the system.

In our analysis of potential differences in the use of information between current and alternative districts, 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. A successful move is 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. In the absence of information on whether job outcomes reflect push or pull factors, it is important to acknowledge that our measure of success is subject to both type 1 and type 2 errors.

[16] 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.
[17] Online Appendix Table A1 reports the coefficient estimates for the student characteristics.

Timing is an important issue to consider when linking these labor market outcomes to measures of school performance. Though student test scores 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, we calculate the change based on the values at time *t* at the sending and receiving schools (or at the sending and receiving districts if the principal moves to a district-level position).[18] Thus, for those who continue in their current position (or move to a district-level position in the same district), the change in case-mix is mechanically zero.

Table 2 shows summary statistics for the two-year labor market outcomes for our main analytic sample in column 1, and for those in the subset with three or more years of tenure in column 2. The majority (65.2 percent) of principals in our main analytical sample are retained in their current position. Approximately one in five (19.9 percent) changes positions within the same district, one in ten (8.1 percent) exits the system, 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 improvements. Altogether, 85.0 percent of principals experience labor market success each year according to our composite measure. For principals

[18] In rare cases, the receiving school or district was not operational in year *t*, so we use the case-mix index from *t*+1 if available, and *t*+2 if not. The case-mix indices at the district-year level are enrollment-weighted averages of the school-by-year indices, standardized to have a zero mean and standard deviation of one across district-years.

with 3 or more years of tenure, the overall rate of success (84.9 percent) and its component transition rates are quite similar.

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. To address inherited differences in the achievement level of students, we use school value-added to achievement as the input in our estimate of principal effectiveness. We focus on achievement since this is the primary metric for elementary school accountability, though there is evidence that schools develop noncognitive skills as well (Jackson 2018).

Value-added models use prior achievement to account for unobserved heterogeneity, recognizing that using just a limited set of characteristics is unlikely to adjust adequately for student and family differences. Our value-added model relates achievement (A) for student i in grade g in school s 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:

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

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.

The estimates of school-by-year fixed effects (g_{st}) from equation (1) provide the building blocks for our measure of principal value-added.[19] We construct this to be the average of the estimated school-by-year fixed effects during a principal spell at a school, excluding the first and last years of the spell.[20] Excluding these years helps to mitigate the influences of persistent principal decisions and shocks around transitions. We average value-added only over the current school to allow for principal-school match effects in addition to fixed principal quality differences.[21] Due to the data requirements, we are only able to calculate our spell value-added measure for the subset of principals in our main analysis sample that have at least three years of tenure.

To account for fixed differences in learning rates across schools, we would ideally infer effectiveness from achievement gains relative to others serving the same school, such as by adding school fixed effects to equation (1).[22] However, these measures of relative effectiveness are only comparable across schools in networks linked by principal transitions and, unfortunately, the majority of connected networks in our setting consist of single schools. In part due to these limitations, the evidence on whether measures of principal effectiveness based on school-by-year value-added are meaningful is not definitive. For example, Chiang, Lipscomb, and Gill (2016) find few statistically significant relationships between school value-added and

[19] We estimate equation (1) using all schools with tested grades from a wider sample period (1996-2011, excluding 2003) in order to benchmark statewide and to have coverage for elementary-school principals that transition to leadership positions at schools serving higher grade levels.
[20] As noted, 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.
[21] Jackson (2013) finds meaningful match effects for teachers, and Lazear et al. (2015) find small but significant match effects for supervisors.
[22] For examples that use this general approach see Coelli and Green (2012) and Dhuey and Smith (2014).

principal value-added estimated from non-overlapping years. However, when the first year following school leader transitions is excluded from the estimates of principal value-added, the point estimate for math is consistent with 51 percent of the difference in value-added between schools reflecting persistent differences in the effectiveness of their principals. Using a different metric for validity, Grissom, Kalogrides, and Loeb (2015) find that school-by-year value-added is more predictive of district evaluations of principals than measures that attempt to control also for school fixed effects. Finally, Branch et al (2020) find strong and highly significant relationships between value-added on the one hand and indexes created from student survey responses to questions about safety and academic engagement and teacher survey responses about principal leadership on the other, even in specifications that include school fixed effects.

Since the jury is still out, it is important to validate our measures of spell value-added as proxies for principal effectiveness. To do so, we borrow the approach developed by Bacher-Hicks, Kane, and Staiger (2014) and Chetty, Friedman, and Rockoff (2014, 2016) to test whether teacher value-added estimates are forecast unbiased. The logic is that if the estimates are valid, then changes in effectiveness over time due to turnover should predict one-for-one changes in achievement. To adapt the approach to our setting, we study how well the change in school-by-year value-added between years t-2 and t+1 following a leadership change in year t is predicted by the change in our proxy for principal effectiveness. When calculating the change in school-by-year value-added, we omit the last year of the outgoing principal's spell and the first year of the incoming principal's spell for the same reasons cited above. We use spell value-added for each principal at the school in question, since this allows for match quality between a principal and school. So that the windows do not overlap, we only include years t-3 and earlier for the outgoing principal and years t+2 and later for the incoming principal in the spell value-added

calculations. Since we continue to omit first and last years of spells, the combined set of restrictions mean that this analysis can only be carried out for schools with outgoing and incoming principals whose spells last at least four years.

Though the above approach is already pushing the limits of our data, there is an argument for further excluding years t-3 and t+2 from the spell value-added calculations. Random shocks to test scores will attenuate the relationship between changes in school-by-year and spell value-added if these are measured using adjacent years of data, since a shock that increases achievement in any given year will reduce value-added in the subsequent year by increasing prior scores. Thus, we also estimate specifications that omit the adjacent years when calculating spell value-added. This imposes the additional restriction that the principals each have at least five years of tenure. As an alternative strategy to address measurement error, we also report regressions that weight by enrollment.

Table 3 shows the results from these validity tests. Although the point estimates in the first cell in column 1 imply that only 12 percent of the change in principal effectiveness is reflected in school-by-year value-added, this more than doubles to 26 percent when observations are weighted by enrollment. As expected, the weighted and unweighted estimates in column 2 that exclude adjacent years to mitigate the effects of measurement error are similar to one another and to the weighted estimate in column 1. In the case of principals, it is not surprising that changes over time in principal effectiveness predict less than one-for-one changes in achievement because factors outside of the principal's control also affect achievement at the school. Overall, we take the evidence to suggest that our spell value-added estimates capture meaningful differences in principal effectiveness.

While ratings and pass rates are readily available to the public, estimates like ours of

principal contributions to achievement growth rely on longitudinal data and can only be computed by insiders, such as district administrators. The top panel of Figure 1 illustrates that the four campus accountability rating categories do not strongly sort principals from low to high effectiveness on the basis of spell value-added to achievement.[23] The distributions of spell value-added for principals reveal a consistent ordering, but differences are small and there is extensive overlap across all rating categories. In contrast, the bottom panel shows that there are striking differences in the distribution of average pass rates across the categories. Importantly, such differences appear even for the subset 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 focused on pass rates clearly penalizes effective principals who work in schools serving predominantly lower achievers who struggle to earn a passing score.

5. Campus rating effects on principal labor market outcomes

This section uses regression discontinuity design (RD) methods to identify the causal effects of school ratings on principal labor market outcomes. We then examine rating effects on future school performance to learn more about the consequences of any principal or district responses.

5.1 Regression discontinuity design approach

The RD analysis exploits discontinuities in the probability of receiving a higher

[23] As is clear from the figure, average spell value-added is positive when calculated across school-years. By design, the school-by-year fixed effect estimates that underlie this measure average to zero when weighted by enrollment, since equation (1) is estimated using the student as the unit of observation.

accountability rating based on the pass rate for the subgroup (i.e., demographic group-by-subject) that is 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 regular statutory threshold, the required improvement threshold, or the exception threshold if an exception is 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. Running variable values greater than (less than) zero indicate that student performance was sufficient (not sufficient) to earn the higher rating.

The distribution of binding subgroups reveals a disproportionate share of schools for which science, which is only tested once for each cohort of students, is the marginal subject at both the acceptable and recognized thresholds.[24] 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. This latter issue of test volatility was first raised by Kane and Staiger (2002). Importantly, it would not be apparent that science performance is often the determining factor without explicitly calculating distances from effective thresholds as we do. Our calculations also reveal that the marginal subgroup is typically the lowest performing in the relevant subject, despite the required improvement and exception provisions.[25]

We estimate our models using local linear regressions with a triangular kernel and use the structure of the accountability system and existing research to guide our choice of bandwidths.

[24] See Online Appendix Table A2.
[25] Online Appendix Table A3 shows that the marginal student subgroup was the lowest performing on the relevant subject about two thirds of the time prior to 2004, and this share fell by about 9 percentage points when the exception provisions were added.

The distances between the statutory pass rates for the various ratings lead 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. 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 Calónico et al. (2017). We cluster standard errors by district in all specifications.

Figure 2 illustrates the first-stage 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 cutoffs, though the bulk of the observations are at the threshold between acceptable and recognized. Even though we fully incorporate the complex, time-varying rules in the construction of the running 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.[26] The corresponding first-stage estimates are reported in Table 4 for the alternative bandwidths, with the optimal bandwidths estimated to be 3.82, 2.49, and 3.18. The estimated discontinuities 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

[26] One source of discrepancy is special accommodations that may be made in particular circumstances that are not explicitly covered 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.

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 Getzler 2006), it is not feasible to do so precisely. 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. Online 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, though this is not necessarily visually apparent and it is hard to imagine that this is due to manipulation of the running variable in our context.[27]

To explore further, we test whether there are any discontinuities in observable characteristics on either side of the rating thresholds. We estimate a system of seemingly unrelated RD regressions using the principal and student characteristics listed in Table 1 as the dependent variables. Online Appendix Table A4 shows that almost none of these exhibits statistically significant discontinuities at the rating boundaries using the optimal bandwidths from the first stages. We fail to reject the null hypothesis that all coefficients are jointly equal to zero for the acceptable and exemplary boundaries, though we do once again reject for the recognized boundary.[28] Importantly, there are no discontinuities in principal spell value-added either unconditionally or conditional on these characteristics at any of the boundaries (see Online Appendix Table A7). And, when estimating discontinuities in principal and school outcomes in

[27] 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.

[28] Online Appendix Tables A5 and A6 show that we do not reject the null at any of the boundaries for the 5-percentage point bandwidth and reject only at the exemplary boundary for the widest 10-percentage point bandwidth, respectively.

the results that follow, the tables show that the inclusion of student and principal controls has little effect on the estimates.

5.2 Regression discontinuity estimates of labor market effects

We present results for the three labor market outcomes for principals: continuing as principal in the same school, salary growth, and change in student composition. All three measures relate ratings based on student achievement in the spring of year *t* to positions held in year *t*+2. Our measure of student composition is the normalized predicted pass rate that weights demographic characteristics based on the relationship with the probability of passing. We also examine the effect of ratings on exits because salary and student composition are observed only if a principal remains in the Texas public schools. The RD estimates for the optimal bandwidths from the first-stage regressions are presented in Table 5 for specifications that alternately exclude and include the additional controls.[29]

Figure 3 plots the relationship between the running variable and the probability of retention around each of the rating boundaries. The sizable discontinuity at the unacceptable-acceptable boundary in the top panel contrasts sharply with little if any jump at the two other thresholds. The corresponding RD estimates reported in Columns 1a and 1b of Table 5 confirm what is evident in the graphs: the estimates of discontinuities associated with moving into the two higher rating categories are small and insignificant, while the estimates show significant increases in retention for crossing the acceptable threshold. For the optimal bandwidth, the estimate conditional on controls is a 38.0 percentage point increase in staying in the same position, which is nearly a doubling relative to the baseline rate of retention for campuses rated unacceptable. Accounting for the fuzziness of the design, the implied LATE estimates are about

[29] Online Appendix Tables A8-A12 show analogous results for the alternative bandwidths. The patterns of results, particularly across the 5 and 2.5 bandwidths, support conclusions based on the optimal bandwidths.

20 percent larger.

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 or voluntary departures. However, it takes two unacceptable ratings in successive years to trigger sanctions, meaning that schools not classified as unacceptable in the prior year are not at risk for sanctions. Less than 10 percent of schools currently rated unacceptable were also rated unacceptable in the prior year, precluding the possibility of estimating the effects of a second consecutive unacceptable rating. Online Appendix Table A9 shows that excluding these schools leads to even larger estimates, refuting the belief that mandatory sanctions drive the unacceptable rating effect on retention.

Beyond continued employment, a principal's job can become better or worse in terms of salary and student case-mix. Figures 4 and 5 show the graphical evidence and Columns 2a/2b and 3a/3b of Table 5 show the estimates for the effects of school ratings on salary growth and the change in student composition, respectively. Similar to the case of retention, Figure 4 and the second panel of Table 5 show that crossing the acceptable threshold significantly increases salary growth by a sizable 6 percent but that there is no significant discontinuity (and point estimates are less than one percent) at the two higher thresholds. The salary gains at the acceptable boundary combine any positive impacts on raises offered to principals at schools that achieve the higher rating and the higher probability of transitioning to lower-paying positions for those in schools below the cutoff. Figure 5 and the third panel of Table 5 reveal no evidence of positive effects on student case-mix of crossing any of the thresholds which, if present, might have muted salary responses.[30]

[30] Online Appendix Table A11 shows that the statistically significant negative estimated effect at the recognized boundary is quite sensitive to bandwidth, becoming small and insignificant at bandwidths of 5 or larger.

The absence of compensation measures for principals who exit the system could introduce selection bias in the salary growth and student case-mix specifications, but this does not appear to be driving the adverse labor market impacts of an unacceptable rating. Crossing the acceptable threshold actually appears to be associated with an increase in the probability of exit, though the estimates are not significant at even the ten percent level in the final two columns of Table 5.[31] 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 would bias downward our estimated effects of an acceptable rating on compensation.

5.3 Rating effects on future school performance

The much lower probability of continuing as principal for those just below the acceptable threshold suggests the presence of an information failure, perhaps because some influential stakeholders focus on the cruder ratings rather than the more detailed information on school performance. Nevertheless, the divergent treatment of principals on opposite sides of the boundary could still be part of a second-best solution if the stigma of an unacceptable rating helps to overcome inertia resulting from a reluctance to remove ineffective principals or resistance of ineffective principals to leave voluntarily. Any benefits depend on the effects of an unacceptable rating on the performance of incumbent principals who remain in their positions and the quality of principals who replace those who do not continue. Because of the endogeneity of continuation, we focus on the reduced-form effects of crossing the acceptable threshold on future school value-added. We recognize that the unacceptable rating may precipitate district interventions, but as long as these do not have adverse impacts on the school, then any benefits to inducing principal turnover at the barely unacceptable schools will be overstated. We also

[31] See also Online Appendix Figure A2.

estimate effects on future pass rates, though pass rates are more likely to also reflect changes in the student body due to family school choice responses to an unacceptable rating.

Table 6 shows RD estimates of the effects of crossing rating boundaries on subsequent school value-added and pass rates in years t+2 and t+3. There is no evidence that schools just below the acceptable threshold have more effective school leadership or higher pass rates either two or three years following receipt of the unacceptable rating. Future value-added is not statistically different on either side of the boundary, and the point estimates are negative only in the t+3 specifications. The future pass rate is in fact higher in schools that barely reached the acceptable rating, and (marginally) statistically significantly so in t+2. As might be expected given the null results for labor market outcomes, there is little evidence of performance impacts for crossing the recognized or exemplary boundaries.[32]

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

Decisions of both the current district and potential alternative employers affect labor market outcomes. The current district is likely to have access to and to make use of detailed information on job performance not readily available to others. The 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. Nonetheless, even if outside employers are more reliant on publicly available school performance information when drawing inferences about principal effectiveness, a low school rating may paradoxically have a larger impact on the decisions of the

[32] Online Appendix Tables A13-A16 show the school value-added and pass rate results for alternative bandwidths. The few point estimates that are statistically significant at the 10-percent level in Table 6 lose significance at wider bandwidths.

current employer if they face stronger stakeholder pressures regarding the employment of specific principals.

To compare within district and new district transitions, we use our 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 the index of student advantage exceeds the median change for all principals who remain in the system 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 obviously successful include principals who move to lower paying and less appealing positions as well as principals who exit the 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 7 presents the RD estimates of the effects of ratings for any success (left panel) and then separately for within district success (middle panel) and new district success (right panel) for the optimal bandwidths from the first stage.[33] Consistent with the retention findings, crossing the acceptable boundary significantly raises the probability of within district success. By comparison, none of the estimates for new district success are statistically significantly different from zero, and the magnitudes of the point estimates are smaller. Importantly, grouping failures

[33] Online Appendix Tables A17 and A18 show robustness to alternative bandwidths.

and successes together in the null category in the RD specifications with binary dependent variables complicates interpretation of the 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 estimates with non-causal multinomial logit regressions that divide principals into those who experience within district success, those who experience new district success, and the baseline group that experiences neither. In addition to ratings, these regressions include other performance metrics, and thus the sample is restricted to campuses led by principals who have at least three years of tenure for whom spell value-added can be calculated. The estimates in Table 8 show that the likelihood of within district success is significantly positively related to spell value-added and the pass rate and negatively related to receiving an unacceptable rating. In contrast, new district success is not significantly related to any of the performance metrics, though these estimates are noisy. Comparing the point estimates, the most notable difference is for the estimate related to receipt of an unacceptable rating. This measure – which is arguably the most salient and least informative performance measure – adversely affects the probability of within but not new district success, consistent with stakeholder pressure to take action in the case of a failing school.

7. Conclusions

Our analysis illustrates the effects of an accountability system that reports both detailed performance data and categorical ratings based on that information. The RD results provide strong causal evidence that failure to achieve an acceptable rating significantly reduces the incumbent principal's probability of job retention and salary growth. Although higher turnover for principals in schools just below the acceptable threshold could improve the quality of

leadership and achievement by overcoming inertia or reluctance to replace ineffective leaders, there is no evidence of this. Falling just below the cutoff does not lead to improvements in future school value-added or in pass rates, the latter of which are the focal metrics of the accountability system. Moreover, the limited differences in distributions of principal value-added across rating categories raise additional questions about the efficacy of an accountability system that focuses on the pass rates of the lowest performing categories of students and that does not ultimately relate ratings to school value-added or principal effectiveness.

Consideration of this evidence in combination with the findings of Deming et al. (2016) highlights the complexity of accountability effects on the quality of instruction and on student outcomes. Teachers, school administrators and districts all respond to accountability pressures, and the contradiction between the positive effects of being at risk of an unacceptable rating and the absence of school improvement following the receipt of an unacceptable rating underscores the importance of how systems are designed and implemented.

The use of a metric that is strongly associated with poverty and low socioeconomic status merits particular scrutiny. Because non-school factors account for a large portion of the variation in the pass rates, movement into a school serving more advantaged students may actually be more beneficial to a principal's labor market prospects than raising the quality of instruction. 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 relevant for the many school accountability systems across the U.S. modeled after its system. Moreover, the increasing use of outcome-based incentives to reduce healthcare spending indicates that these concerns extend far beyond the education sector. The expanded use of categorical ratings across public institutions represents an interest in more

transparent public accountability, but the impact will depend crucially on the details.

References

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. 2020. "Public and Private Employer Learning: Evidence from the Adoption of Teacher Value-Added." *Journal of Labor Economics* 38(2): 375-420.

Bertrand, Marianne, and Antoinette Schoar. 2003. "Managing with Style: The Effect of Managers on Firm Policies." *The Quarterly Journal of Economics* 4 (November): 1169-1208.

Branch, Gregory F., Eric A. Hanushek, Steven G. Rivkin, and Jeffrey C. Schiman. 2020. "How Much Does Leadership Matter? Evidence from Public Schools." unpublished manuscript.

Bryant, Troy. "Surveys on contract practices reveal common trends." Texas Association of School Boards Human Resources Exchange, February 1, 2017, (https://www.tasb.org/services/hr-services/hrx/hr-laws/surveys-on-contract-practices-reveal-common-trends.aspx).

Calónico, 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. 2009. "'How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics* 93: 1045–1057.

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.

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.

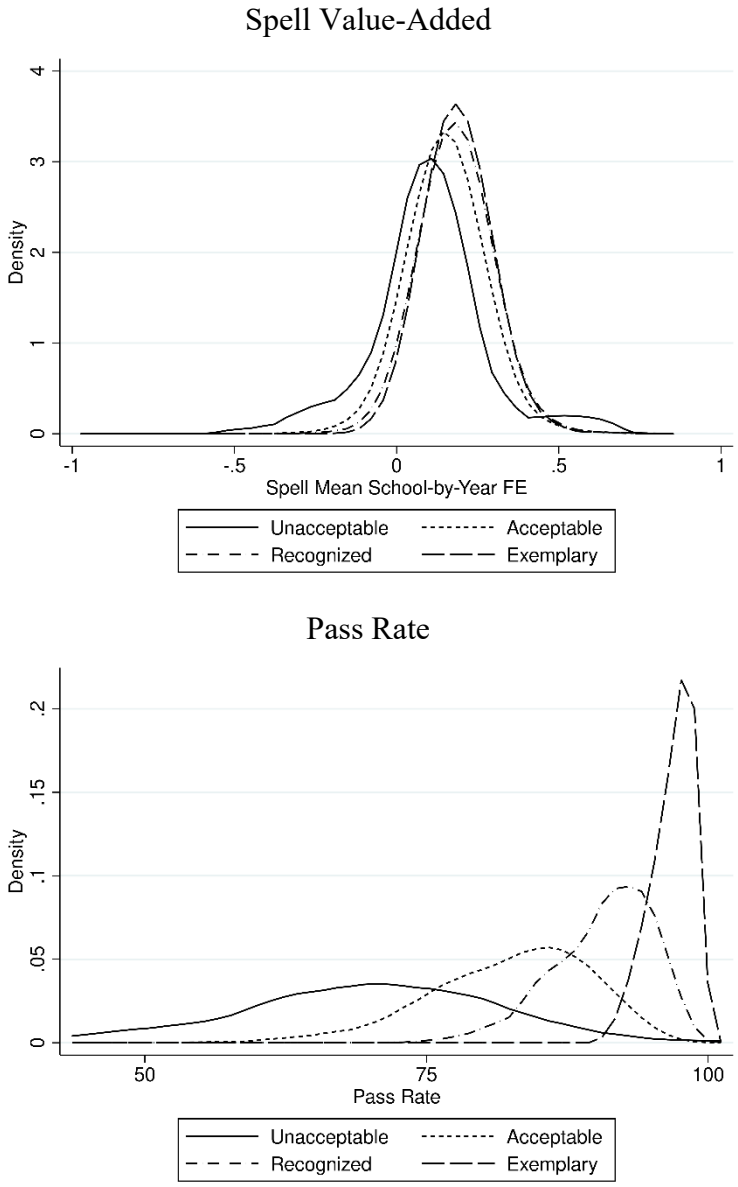
Dhuey, Elizabeth, and Justin Smith. 2014. "How Important are School Principals in the Production of Student Achievement?" *Canadian Journal of Economics/Revue canadienne d'économique* 47, no. 2 (May): 634-663.

Dizon-Ross, Rebecca. 2020. "How Does School Accountability Affect Teachers? Evidence from New York City" *Journal of Human Resources* 55, no. 1 (Winter): 76-118.

Farber, Henry S., and Robert Gibbons. 1996. "Learning and Wage Dynamics." T*he Quarterly Journal of Economics* 111(4): 1007-1047.
Feng, Li, David Figlio, and Tim Sass. 2018. "School Accountability and Teacher Mobility." *Journal of Urban Economics* 103: 1-17.
Figlio, David N., and Lawrence S. Getzler. 2006. Accountability, Ability and Disability: Gaming the System? In *Improving School Accountability*, edited by Timothy J. Gronberg and Dennis W. Jansen: 35-49.
Figlio, David N., and Lawrence Kenny. 2009. "Public Sector Performance Measurement and Stakeholder Support." *Journal of Public Economics* 93(9-10): 1069-1077.
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.
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.
Graham, John., Si Li, and Jiaping Qiu. 2012. "Managerial Attributes and Executive Compensation." *Review of Financial Studies* 25, no. 1: 144-186.
Grissom, Jason A., and Brendan Bartanen. 2019. "Principal Effectiveness and Principal Turnover." *Education Finance and Policy* 14, no. 3: 355-382.
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.
Jackson, Kirabo. 2013. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers." *Review of Economics and Statistics* 95 (October): 1096-1116.
Jackson, Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes." *Journal of Political Economy* 126, no. 5 (October): 2072-2107.
Kane, Thomas and Douglas O. Staiger. 2002. " The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives* 16, no. 2 (Fall): 91-114
Lazear, Edward P., Kathryn L. Shaw, and Christopher T. Stanton. 2015. *Journal of Labor Economics* 33(4): 823-861.
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.
Rockoff, Jonah, and Lesley J. Turner. 2010. "Short-Run Impacts of Accountability on School Quality." *American Economic Journal: Economic Policy* 2, no. 4: 119–147.
Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. "Feeling the
Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." *American Economic Journal: Economic Policy* 5(2): 251–281.
Schönberg, Uta. 2007. "Testing for Asymmetric Employer Learning." *Journal of Labor Economics* 25(4): 651-691.

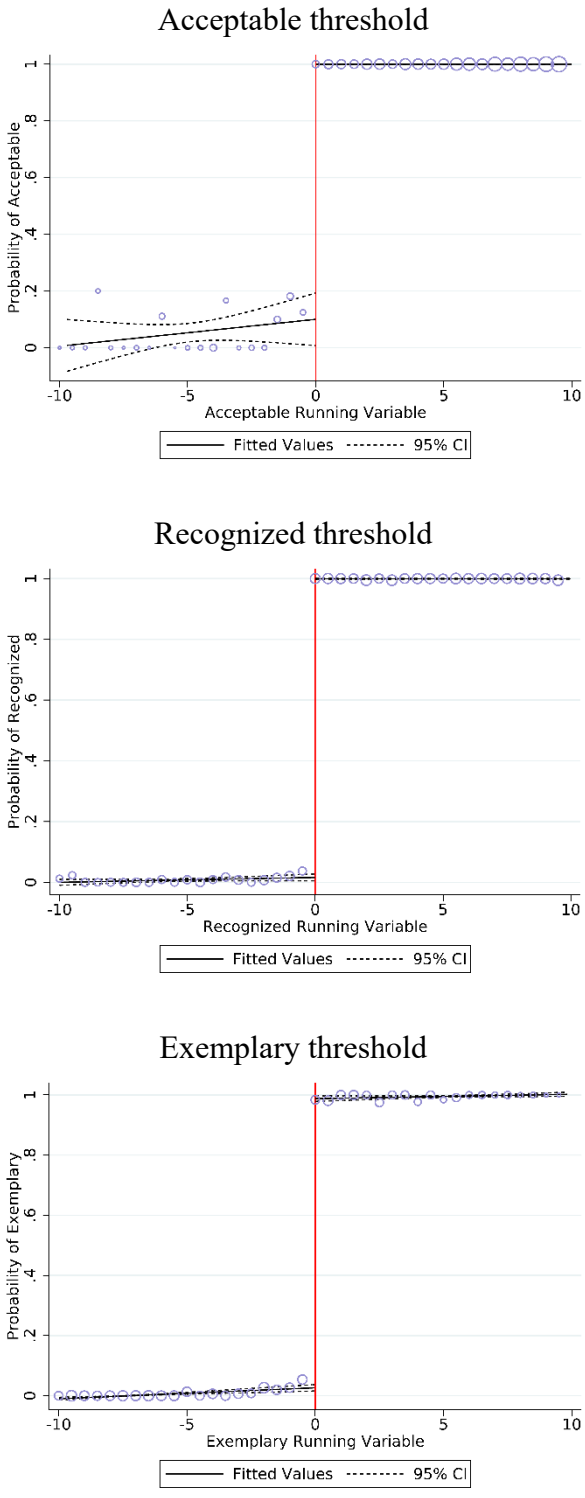
Sokol, Marlene. 2018. "State Grades Push Hillsborough into an Unexpected Wave of Principal Transfers." *Tampa Bay Times*, July 5.
Toenjes, Laurence A., and Jean E. Garst. 2000. "Identifying High Performing Texas Schools and Schools Districts and their Methods of Success." Texas Education Agency (December).

Figure 1. Distribution of school performance metrics, by school accountability rating



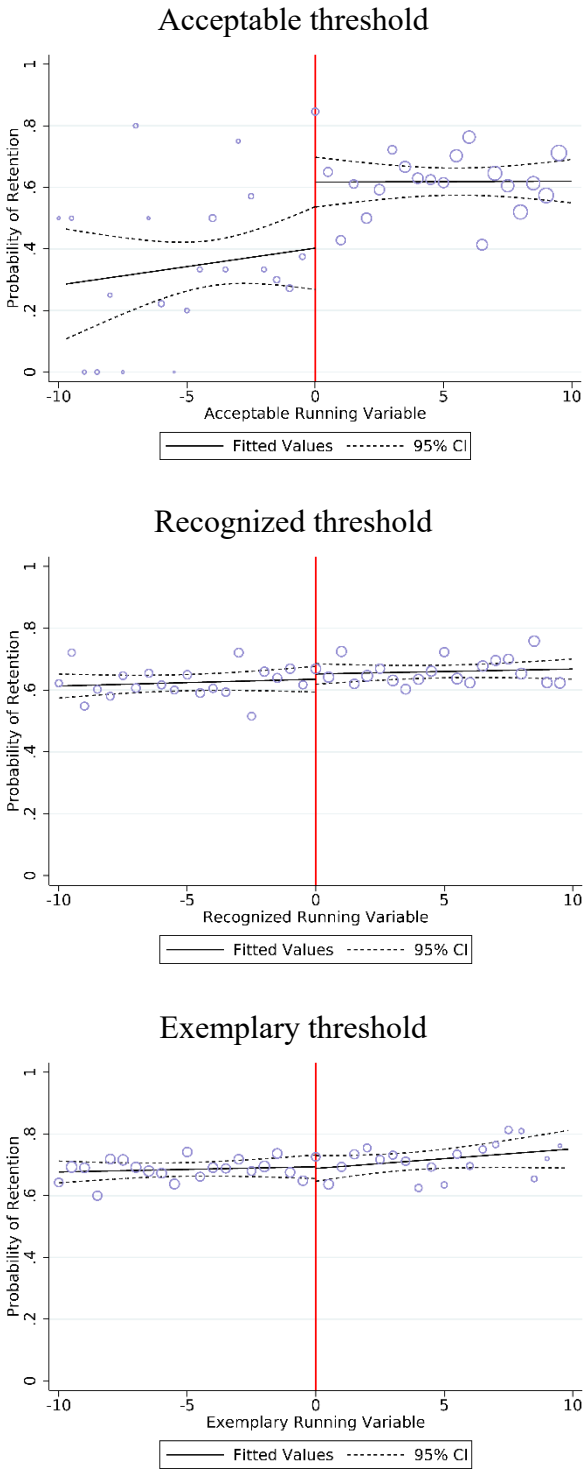
Notes: In both panels, the unit of observation is a school-by-year and the sample is restricted to school-years where the current principal serves a spell of at least three years. The restricted sample used for these figures includes 8,166 school-by-year observations and 3,248 principal spells. School-years are classified according to the campus rating earned in that year. Spell value-added in the top panel is calculated by averaging the school-by-year fixed effects estimated from the student-level achievement growth model (equation (1) in the text) across all years of a principal spell excluding the first and last year. The pass rate in the bottom panel is the average across math and reading by school and year.

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



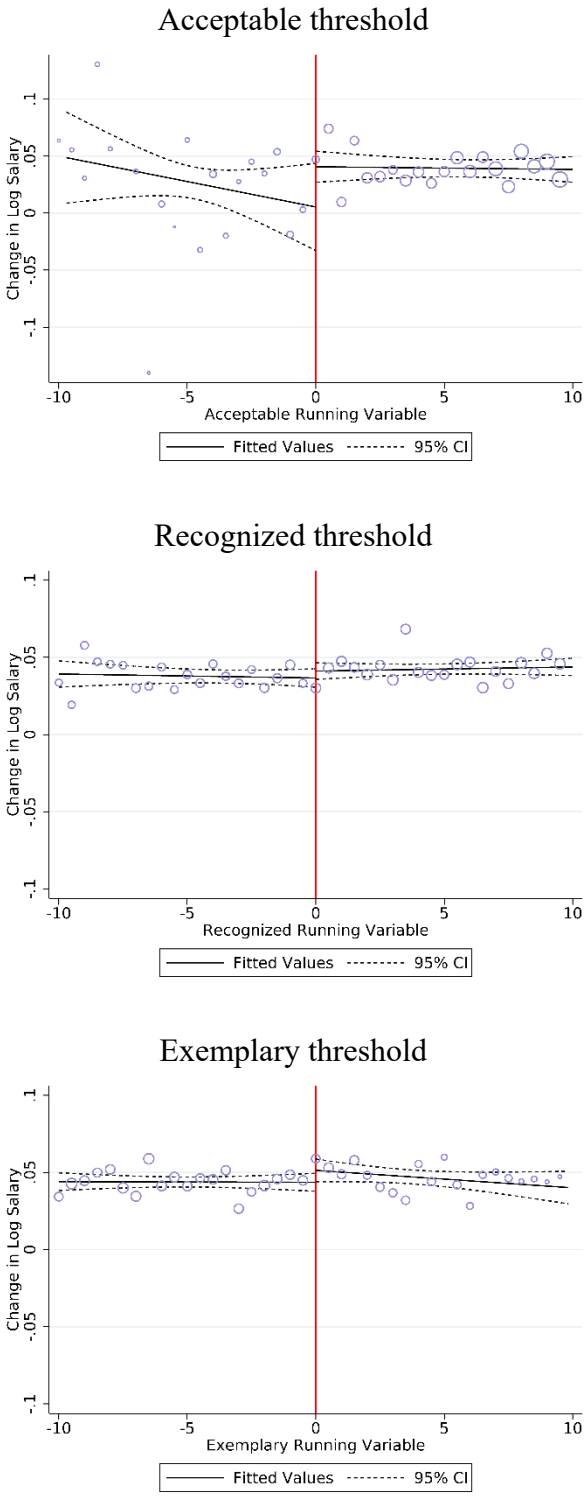
Notes: In each panel, the running variable is the difference between the pass rate for the marginal student subgroup and the relevant pass rate threshold. The bin width is 0.5 percentage points. Points are weighted by bin size (i.e., number of school-by-year observations) and are comparable within rating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

Figure 3. Probability of retention, by accountability rating threshold



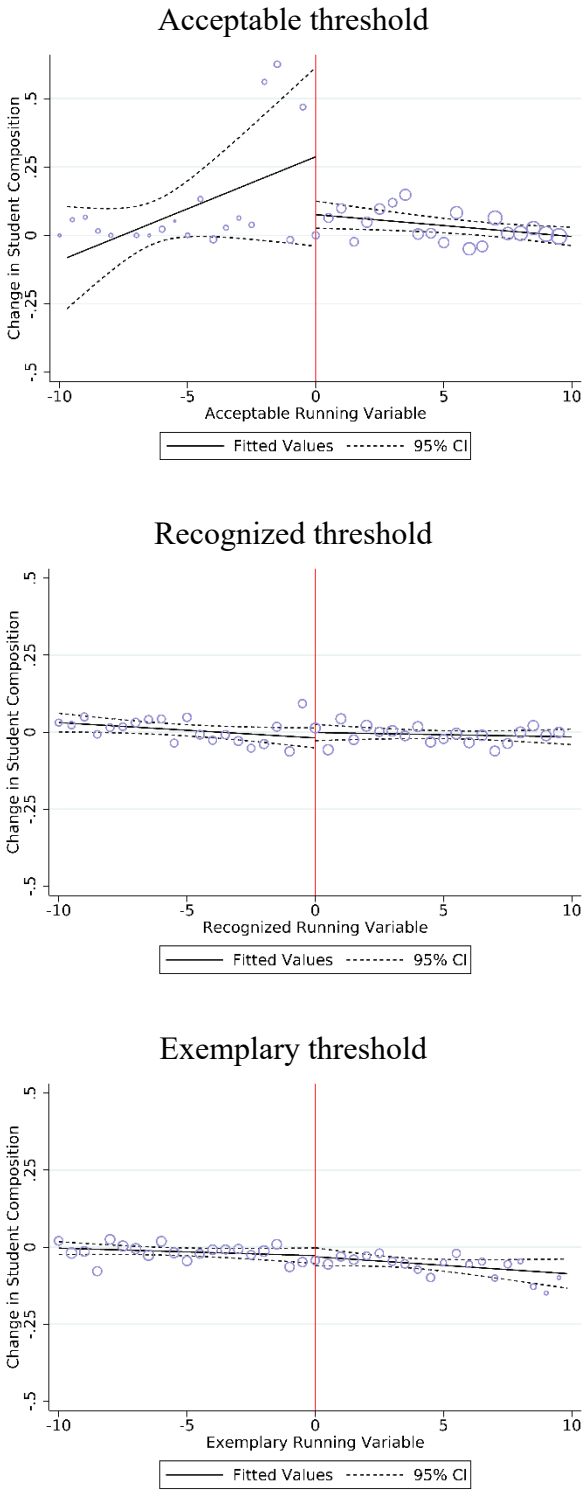
Notes: Retention is defined as continuing in the same principal position in academic year *t*+2, with the school rating realized at the end of academic year *t*. For other details, see the notes to Figure 2.

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 school rating realized at the end of academic year *t*. For other details, see the 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 school rating realized at the end of academic year t. For other details, see the notes to Figure 2.

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

Variable	All	Experience <25 years	Tenure ≥2 years at school	Tenure ≥3 years at school
	(1)	(2)	(3)	(4)
Principal characteristics				
Male	0.281	0.290	0.284	0.276
Black	0.109	0.101	0.100	0.095
Hispanic	0.224	0.214	0.212	0.207
White	0.663	0.680	0.684	0.694
Other race/ethnicity	0.004	0.004	0.004	0.004
Below Master's degree	0.055	0.072	0.072	0.063
Master's degree	0.904	0.895	0.895	0.905
Doctorate degree	0.040	0.033	0.033	0.032
2 or fewer years tenure	0.272	0.329	0.274	0.000
3 years tenure	0.160	0.191	0.207	0.279
4 or more years tenure	0.568	0.479	0.519	0.721
Total years of experience	22.49	17.53	17.64	18.42
Principal salary				
Total pay (2003 dollars)	$66,478	$64,089	$64,078	$64,718
Student test performance				
Average math/reading pass rate	88.02	88.01	88.13	88.53
Math pass rate	87.07	87.03	87.17	87.61
Reading pass rate	88.85	88.87	88.96	89.33
School accountability rating				
Unacceptable	0.012	0.012	0.012	0.009
Acceptable	0.381	0.384	0.377	0.364
Recognized	0.438	0.441	0.446	0.448
Exemplary	0.169	0.163	0.165	0.179
School student characteristics				
Male	0.514	0.514	0.515	0.515
Black	0.142	0.135	0.134	0.130
Hispanic	0.466	0.459	0.459	0.459
White	0.361	0.375	0.376	0.379
Other race/ethnicity	0.031	0.031	0.031	0.032
Economically disadvantaged	0.601	0.595	0.594	0.591
Title 1 participant	0.722	0.727	0.725	0.720
Limited English proficient	0.21	0.207	0.207	0.206
Special education	0.107	0.107	0.107	0.108
Gifted and talented	0.061	0.059	0.059	0.058
Mid-year school mover	0.062	0.062	0.062	0.061
N (school-by-year)	20,045	12,296	11,351	8,166

Notes: Means for all elementary school-by-year observations for the years 2001 to 2008 (excluding 2003) are reported in column 1. Column 2 restricts the sample to school-by-year observations with principals that have less than 25 years of total experience in Texas public schools. Columns 3 and 4 further restrict the sample to observations with principals that have led the current school for at least two years and for at least three years, respectively. The cells report proportions, other than for principal salary (in dollars), total years of experience (in years) and student pass rates (in percentages).

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

Variable	Experience <25 and tenure ≥2 years	Experience <25 and tenure ≥3 years
	(1)	(2)
Outcomes for all principals		
Retained	0.652	0.652
Moved within the same district	0.199	0.201
Successful move within district	0.150	0.152
Successful move with high salary growth	0.129	0.130
Unsuccessful move within district	0.049	0.049
Moved to a new district	0.069	0.064
Successful move to a new district	0.048	0.045
Successful move with high salary growth	0.038	0.036
Unsuccessful move to a new district	0.021	0.019
Exit Texas public schools	0.081	0.082
N (school-by-year)	11,351	8,166
N (principals)	4,222	3,248
N (schools)	3,251	2,774
Outcomes for principals who remain in the system		
Salary growth	0.039	0.039
	(0.081)	(0.080)
Change in student composition	-0.012	-0.015
	(0.335)	(0.342)
N (school-by-year)	10,437	7,494
N (principals)	3,934	3,018
N (schools)	3,157	2,657

Notes: Statistics for all school-by-year observations with principals that have less than 25 years of total experience in Texas public schools and have been principal at the current school for at least two years are reported in column 1. Column 2 further restricts the sample to principals that have been principal at the current school for at least three years. Standard deviations for continuous variables are shown in parentheses below the means. The outcomes are based on academic year t+2, with the school 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. Ordinary least squares estimates of the relationship between changes in school performance and principal effectiveness following principal transitions

Independent variable: Change in principal spell value-added after transition in *t*	Dependent variable: Change in school-by-year value-added *t*-2 to *t*+1	
	(1)	(2)
Unweighted	0.118*	0.227***
	(0.061)	(0.053)
Weighted	0.256***	0.217***
	(0.002)	(0.003)
Calculation of change in spell value-added		
Last year included for previous principal	*t*-3	*t*-4
First year included for new principal	*t*+2	*t*+3
Both principals serve at least:		
4 years	Yes	Yes
5 years	No	Yes
Number of transitions	2,758	1,272

Notes: Each cell reports results from a separate ordinary least squares regression of the change in school performance on a proxy for the change in principal effectiveness between the previous and the new principal, as well as year fixed effects. The dependent variable is the change in school-by-year value-added between the penultimate year of the previous principal (*t*-2) and the second year of the new principal (*t*+1). This change is calculated using the school-by-year fixed effects estimated from the student-level achievement growth model (equation (1) in the text). The independent variable is the difference in spell value-added between the new and previous principals. Across the two columns, different restrictions are imposed on the subset of years at the given school used for the calculation of spell value-added. Spell value-added in column 1 is calculated by averaging school-by-year value-added across years for each principal, excluding not only the first and last year of spells but also the years which are used in the calculation of the dependent variable (*t*-2 for the previous principal and *t*+1 for the new principal). In column 2, years *t*-3 and *t*+2 are also excluded. All schools for which changes in spell value-added can be calculated across principal transitions are included. Regressions are weighted by average student enrollment over the *t*-2 to *t*+1 period for the "Weighted" results. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.10

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***	0.862***	0.835***	0.798***	0.819***
	(0.059)	(0.069)	(0.087)	(0.135)	(0.103)
Mean	0.062	0.064	0.067	0.095	0.081
N	760	495	299	140	221
Recognized	0.978***	0.975***	0.972***	0.961***	0.961***
	(0.007)	(0.009)	(0.012)	(0.019)	(0.019)
Mean	0.009	0.009	0.012	0.016	0.016
N	5,613	4,250	2,879	1,459	1,458
Exemplary	0.954***	0.948***	0.936***	0.911***	0.921***
	(0.009)	(0.011)	(0.015)	(0.024)	(0.021)
Mean	0.008	0.011	0.017	0.028	0.023
N	4,935	3,927	2,690	1,420	1,768

Notes: Each cell shows the estimated discontinuity at the threshold from a separate local linear regression with a triangular kernel, with the associated standard error clustered by district 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 Calónico et al. (2017). The optimal bandwidths for the Acceptable, Recognized, and Exemplary thresholds are 3.82, 2.49, and 3.18 percentage points, respectively. *** p<0.01, ** p<0.05, * p<0.10

Table 5. Regression discontinuity estimates of the impact of attaining the higher rating on principal labor market outcomes, by rating threshold

	Dependent variable							
	Job retention		Salary growth		Change in student composition		Exits Texas public schools	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
Acceptable	0.419***	0.380***	0.059**	0.060**	-0.319	-0.367	0.092	0.104
	(0.138)	(0.115)	(0.025)	(0.022)	(0.286)	(0.237)	(0.069)	(0.069)
Mean	0.419		-0.017		0.208		0.145	
N	221		181		181		221	
Recognized	0.020	0.008	-0.002	0.000	-0.076	-0.072*	-0.013	-0.006
	(0.060)	(0.057)	(0.008)	(0.008)	(0.048)	(0.043)	(0.035)	(0.035)
Mean	0.631		0.001		-0.013		0.096	
N	1,458		1,284		1,284		1,458	
Exemplary	0.025	0.022	0.006	0.006	0.000	0.018	-0.002	-0.001
	(0.048)	(0.046)	(0.008)	(0.008)	(0.040)	(0.034)	(0.024)	(0.023)
Mean	0.693		0.006		-0.023		0.069	
N	1,768		1,605		1,605		1,768	
Controls	No	Yes	No	Yes	No	Yes	No	Yes

Notes: Each cell shows the estimated discontinuity at the threshold from a separate local linear regression with a triangular kernel and the optimal bandwidth from the first stage, with the associated standard error clustered by district shown in parentheses. The optimal bandwidths for the Acceptable, Recognized, and Exemplary thresholds are 3.82, 2.49, and 3.18, respectively. The results reported in the "a" columns are from specifications that do not include any additional variables in the control set, while those in the "b" columns add the year-*t* principal, school and student characteristics listed in Table 1. The means of the dependent variables, which vary across the columns, are shown for the subset of principals within the bandwidth sample receiving the lower rating. Job retention is defined as continuing in the same principal position in academic year *t*+2, with the school rating realized at the end of academic year *t*. Salary growth is measured by the change in the log (real $2003) total pay between academic years t+2 and *t*. 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*. Exiting is defined as not holding any position within the Texas public school system in academic year *t*+2. *** p<0.01, ** p<0.05, * p<0.10

Table 6. Regression discontinuity estimates of the impact of attaining the higher rating on subsequent school performance, by rating threshold

	Dependent variable			
	School-by-year value-added		Pass rate	
	(1a)	(1b)	(2a)	(2b)
	Panel A: Outcomes in t+2			
Acceptable	0.042	0.063	3.775*	3.777*
	(0.050)	(0.044)	(2.181)	(2.112)
Mean	0.147		78.29	
N	194		194	
Recognized	0.033	0.035*	0.731	0.811
	(0.021)	(0.020)	(1.069)	(0.795)
Mean	0.149		86.23	
N	1,127		1,127	
Exemplary	-0.012	-0.010	0.680	0.535
	(0.013)	(0.012)	(0.653)	(0.465)
Mean	0.206		91.99	
N	1,346		1,346	
	Panel B: Outcomes in t+3			
Acceptable	-0.041	-0.008	2.128	2.650
	(0.075)	(0.061)	(2.680)	(2.593)
Mean	0.120		79.37	
N	193		193	
Recognized	0.008	0.009	-0.318	-0.235
	(0.020)	(0.019)	(0.911)	(0.687)
Mean	0.147		85.67	
N	1,363		1,363	
Exemplary	0.015	0.015	0.634	0.660
	(0.014)	(0.012)	(0.723)	(0.532)
Mean	0.200		90.71	
N	1,629		1,629	
Controls	No	Yes	No	Yes

Notes: The dependent variables in this table are school-by-year value-added (columns 1a and 1b) and the average of the math and reading pass rate (columns 2a and 2b). School-by-year value-added is defined to be the school-by-year fixed effect estimated from the student-level achievement growth model (equation (1) in the text). The dependent variables are measured in *t*+2 in Panel A and *t*+3 in Panel B, with the school rating realized at the end of academic year *t*. The results reported in the "a" columns come from specifications that do not include any additional variables in the control set, while those in the "b" columns add the year-*t* principal and school student characteristics listed in Table 1. The samples have been restricted to school-by-year observations that fall within the optimal bandwidth from the first stage and have school-by-year value-added observed. For other details, see the notes to Table 5. *** p<0.01, ** p<0.05, * p<0.10

Table 7. Regression discontinuity estimates of the impact of attaining the higher rating on composite labor market success, by rating threshold and employment location

	Dependent variable					
	Any success		Within district success		New district success	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Acceptable	0.189*	0.165	0.288**	0.284**	-0.098	-0.119
	(0.115)	(0.115)	(0.129)	(0.118)	(0.090)	(0.080)
Mean	0.661		0.532		0.129	
N	221		221		221	
Recognized	0.016	0.011	0.035	0.026	-0.020	-0.015
	(0.042)	(0.043)	(0.048)	(0.048)	(0.025)	(0.024)
Mean	0.809		0.760		0.049	
N	1,458		1,458		1,458	
Exemplary	-0.023	-0.023	-0.041	-0.042	0.018	0.019
	(0.033)	(0.032)	(0.040)	(0.039)	(0.023)	(0.023)
Mean	0.874		0.834		0.040	
N	1,768		1,768		1,768	
Controls	No	Yes	No	Yes	No	Yes

Notes: Each cell shows the estimated discontinuity at the threshold from a separate local linear regression with a triangular kernel and the optimal bandwidth from the first stage, with the associated standard error clustered by district shown in parentheses. The optimal bandwidths for the Acceptable, Recognized, and Exemplary thresholds are 3.82, 2.49, and 3.18, respectively. The results reported in the "a" columns are from specifications that do not include any additional variables in the control set, while those in the "b" columns add the year-*t* principal and school student characteristics listed in Table 1. The means of the dependent variables, which vary across the columns, are shown for the subset of principals within the bandwidth sample receiving the lower rating. Any success is a composite principal labor outcome measure defined to include being retained at the same school or realizing above median gains in log salary or student composition between academic years *t*+2 and *t*, with the school rating realized at the end of academic year *t*. *** p<0.01, ** p<0.05, * p<0.10

Table 8. Multinomial logit estimates of relationships between school performance metrics and composite labor market success within district and out of district

	Outcomes	
	Within district success	New district success
	(1a)	(1b)
Spell value-added	1.366**	1.598
	(0.592)	(1.134)
	[0.126]	[0.016]
Pass rate	0.034***	0.020
	(0.011)	(0.020)
	[0.004]	[0.000]
Unacceptable	-1.266***	0.270
	(0.203)	(0.512)
	[-0.168]	[0.043]
Recognized	0.081	-0.074
	(0.102)	(0.206)
	[0.012]	[-0.004]
Exemplary	0.223	0.058
	(0.191)	(0.364)
	[0.026]	[-0.004]

Notes: This table reports multinomial logit estimates for the sample of principals with at least 25 years of experience that have at least three years of tenure in their current position (N=8,166), with standard errors clustered by district reported in parentheses. The three outcomes modeled are i) achieving success within the same district (column 1a), ii) achieving success in another district (column 1b), and iii) neither, where neither is the base outcome and success is defined as in Table 7. Average marginal effects (or differences in probabilities of outcomes for the binary ratings) are reported in brackets. Acceptable is the excluded rating category. The specification includes district and year fixed effects and controls for the year-*t* principal and school student characteristics listed in Table 1. *** p<0.01, ** p<0.05, * p<0.10