

DOES IT PAY TO GET AN A?
SCHOOL RESOURCE ALLOCATIONS IN
RESPONSE TO ACCOUNTABILITY RATINGS

Steven G. Craig¹, University of Houston

Scott A. Imberman, University of Houston and NBER

Adam Perdue, University of Houston

Abstract

This paper examines public school district budgetary responses to school accountability ratings. We estimate the effect of a “rating shock” on budgets through a major change in school accountability systems in Texas. We find a 1.5% increase in instructional budgets, mainly for teachers, as a response to a drop in rating. This increase is found to disappear within three years, suggesting temporary budget support to “learn” the new system. In contrast, using regression discontinuity estimates, we find little response to random year-to-year variation in ratings in the new system, but some evidence of rewards being provided in the old.

July, 2011

¹ Department of Economics, 204 McElhinney Hall, Houston, TX 77204-5019. The authors are grateful for helpful advice and comments from Aimee Chin, Meltem Daysal, Janet Kohlhase, Elaine Liu, Chinhui Juhn, Lori Taylor and Mary Beth Walker, as well as participants at the University of Houston/Rice University applied microeconomics seminar, the University of Florida, and the meetings of the American Economic Association, Southern Economic Association, and the American Education Finance Association. Correspondence should be sent to Scott Imberman at simberman@uh.edu or Steven Craig at scraig@uh.edu. All errors and opinions are our own. ©2011 by Steven Craig, Scott Imberman, and Adam Perdue. Short sections of text not to exceed two paragraphs may be excerpted without explicit permission provided proper citation is given.

1. Introduction

Accountability systems have been fixtures of the US public education system since the late 1990's. They evaluate schools based on student performance on statewide standardized tests, and assign simple ratings based on the test score results and sometimes other factors. The ratings are designed to be informative to parents and state legislators, and one objective of school accountability ratings appears to be to facilitate pressure from these two groups onto school administrators. While there is an extensive literature on within school responses to the tests upon which the ratings are based, there is very little exploration of whether there are resource allocation responses by school districts.² If parents and/or legislators find that ratings are informative, school administrators might respond by allocating resources between schools in response to the ratings. The difficulty for researchers has been to identify the impact of state accountability ratings independently of other causes of budgetary choices. We exploit two empirical strategies which successfully circumvent this problem, and find that schools only craft temporary budget responses to school accountability ratings.

The primary analysis method we use to investigate the school and school district response to changes in school ratings exploits the “rating shock” that arose when the state government of Texas switched accountability systems over the two year period 2002 to 2004. The new accountability system was more rigorous, thus many schools were threatened with a lower accountability grade than they had received on average from the old system. During the 2002-03

² Dee and Jacob (2009) and Neal and Schanzenbach (2010) find evidence that the Federal No Child Left Behind (NCLB) law increases achievement. Chiang (2010), Jacob (2005), Reback (2008), Hanushek and Raymond (2004, 2005), and Rockoff and Turner (2010) find test score improvements as a result of state or city based accountability regimes.

school year there was a discrete change in the school accountability system where more rigorous exams covering broader material were administered. School ratings summarizing student performance were not given, however, until one year later with the second testing cycle. This provided schools with time to adapt to the new testing regime. We exploit the information from the ‘gap’ year by using the rating shock procedure proposed by Figlio and Kenny (2009), because only after receiving the initial test results did school administrators have an opportunity to estimate their school’s new ratings and engineer a response. The value of this episode is that we can use the budgetary response to assess how administrators react to a change in (potential) ratings without a corresponding change in school quality.

We find that schools and their districts responded to the new exam by reallocating resources to schools where there was an increased likelihood of a lower accountability grade.³ Further, almost all of the incremental funds were directed to instruction, at least some of which led to increased teacher hiring. This budgetary response, which was centered amongst schools that fell to the second lowest rating category, is found to be \$75 per-pupil or 1.2% more total expenditures. As a result, instruction budgets increased by 1.5% relative to schools with no change in rating, and student teacher ratios fell by 0.3% with no corresponding change in enrollment. We find, however, that this budgetary response is temporary, as we also show that the school budgets three years from the change exhibit no difference from schools that did not experience ratings changes.

³ Unfortunately we are not able to separate the roles played by the individual schools or the school districts. Districts in Texas are given a lot of leeway in terms of how to allocate funds to schools. While most districts apportion most funds via formula, there are opportunities for funds to be apportioned in other, more subjective ways.

In contrast to the discrete change captured by the rating shock method, we also attempt a regression discontinuity (RD) design to compare the budgetary response to annual changes in rating for schools on either side of a rating boundary. The sharp discontinuity occurs because school grades are based on the percentage of students which pass the accountability exam, even one student less than the line results in a lower school grade. Due to random factors, schools that just barely receive a higher rating should be a valid comparison group for schools that just barely receive a lower rating (Lee, 2008). Our RD strategy thus tests for whether there are annual budgetary changes in response to a school's success or failure to surmount the marginal rating hurdle, again without any change in underlying school quality. Unlike the rating shock strategy, the RD method does not show significant evidence of budget reallocations due to school grades in the new accountability system, although we find evidence that successful schools received small budget increases during the early accountability system.

One reason to expect an administrative response to accountability ratings is if the ratings are important to parents. For example, Figlio and Lucas (2004) find that there is a housing market response to the information ratings provide over and above the measured learning output of schools. One caution in their results, however, is they find the housing market response seems to decline over time, possibly suggesting that the usefulness of ratings fall over time as residents learn about how accountability is related to school quality.

A second reason for expecting changes in the allocation of resources between schools is because of the range of response found by those examining within school behavior. Examples of how schools internally respond to accountability systems include Chakrabarti (2007, 2008); Carnoy and Loeb, (2003); Hanushek and Raymond, (2004, 2005); and Jacob, (2005).

Researchers have also found, however, that some gains may be due to schools “gaming” the system (Cullen and Reback, 2006; Figlio, 2006; Figlio and Getzler, 2002; Figlio and Winicki, 2005; Jacob, 2005) or focusing on marginal students (Chakrabarti, 2007; Neal and Schazenbach, 2010; Reback, 2008). This range of results suggests ratings are important to school administrators. Our research is an important extension of this work, as we provide a careful look at whether the technical response within schools to accountability is accompanied by changes in the allocation of resources both within and across schools.

Thus, while we know that schools respond to accountability in some finely detailed ways, we know only a little about the resource allocation response. Rouse, Hannaway, Goldhaber, and Figlio (2007) provide evidence from a survey of schools that identify a number of policy changes induced by low ratings, but there is very little evidence on how schools and districts re-allocate resources in response to variation in ratings. Bacolod, Dinardo and Jacobson (2009) find that schools that receive rewards for higher ratings generally put the money into teacher bonuses. Jacob (2003) looks at how school resources in Chicago adjust to the imposition of an accountability system, and finds shifts in expenditures to non-ancillary instruction amongst schools with low pre-accountability test scores but overall, he finds little change. Neither of these papers, however, looks at resource allocation responses directly as a result of ratings. One paper that considers this question is Chiang (2009) who finds evidence that schools which receive a “failing” grade in Florida increase spending on instruction and instructional tools. Nonetheless his paper only considers elementary schools for a single year. Our analyses of the resource allocation responses to accountability are considerably broader. First, we consider the response along different points of the rating distribution via our two identification strategies. Second, our

ratings shock methodology allows us to investigate how schools respond to large and unexpected changes in ratings, rather than the marginal changes in ratings analyzed by Chiang and in our regression-discontinuity analysis.

Our results suggest a nuanced response consistent with the findings of Chiang (2009), although in a much broader context. Specifically, our rating shock analysis suggests that school districts are willing to re-allocate funds towards schools that are threatened by an accountability rating reduction. These funds can only come from schools that are relatively secure since the overall district budget is essentially constant in the short-run. We also find that the resource increase is temporary, as we find it disappears within three years. Despite this finding, our RD results suggest little in the way of yearly adjustments by districts to alter the ratings, or distribution of ratings, across schools.

The combination of these findings suggest that, while schools and districts respond financially to a wholesale change in the accountability system, they are relatively insensitive to year-to-year variation in the ratings – at least for those schools which are close to a ratings margin. This combination of results suggest school districts are willing to invest in learning how to respond to accountability systems as a whole, but that annual variation in ratings is not used to incentivize schools on the margin.

2. The Texas Accountability Systems

Texas initiated one of the first education accountability systems in the US in 1993, called the Texas Assessment of Academic Skills (TAAS). Under TAAS schools were given ratings - from highest to lowest - of E (exemplary), R (recognized), A (acceptable), and L (low

performing).⁴ Accountability standards consist of clear demarcations based on the pass rates of students on a standardized exam administered by the state, along with attendance, dropout, and school completion rates. Higher performance by one student does not compensate for lower performance by another; the criteria are solely based on whether each student performs higher than a minimum.⁵ Table 1 shows the distribution of ratings by year. After a one year transition during 2002-03, the new accountability system called the Texas Assessment of Knowledge and Skills (TAKS) was implemented for the 2003-04 year. Under TAKS, the ratings distribution shifted downwards, demonstrated in the lower half of the table, as fewer schools were awarded E's and more were rated L or A.

A school's accountability rating under TAAS is based on the share of tested students who pass the state-wide exam in multiple student groups and subjects. The groups consist of all tested students along with four student subgroups - white, African-American, Hispanic, and economically disadvantaged. The subjects are math, reading, writing, and social studies (only for 8th grade). Dropout rates and attendance could also affect the rating.⁶ Thus the system is based on test score levels rather than student gains, and the rating for the entire school is determined by the lowest performing subject-group of sufficient size. Appendix Table 1 provides a description of the requirements for achieving each rating in the years of our analysis.⁷ With the exception of

⁴ The "Low Performing" rating under the early accountability system was renamed to "Academically Unacceptable" under the later system. Additionally, both systems accommodate non-traditional schools. See footnote 22 for more detail.

⁵ While they do not factor into the accountability ratings, schools with large percentages of students scoring at a higher "commended" performance level receive additional recognition.

⁶ The attendance requirements were abandoned in 1999-2000.

⁷ More detail on the rules underlying the Texas accountability system can be found via the Texas Education Agency at <http://ritter.tea.state.tx.us/perfreport/account/>.

receiving a low rating, there are no direct punishments imposed on schools by the state, and state awards for high performance were extremely small.⁸ Thus, the rating system primarily acts as a reputation based system where schools are incentivized by the public response, unless school districts themselves develop internal penalties or rewards based on the rating. Panel A of Table 2 shows the transition matrix averaged over the years of TAAS, and shows that schools often changed ratings from year to year with only 69% of A's, 54% of R's, and 71% of E's maintaining their ratings the following year.

The 2002-03 school year served as a transition period between the old system and the new system based on the TAKS exam. Under TAKS schools faced a similar rating system with the same four ratings.⁹ Table 1 shows that the likelihood of an L rating, while still very low, triples compared to the earlier TAAS period while the likelihood of an E rating falls dramatically. Table 2 similarly shows that maintaining an E rating in TAKS is much less likely, while repeating an L rating is much more likely than under TAAS.

The TAKS system is structured similarly to TAAS but with a more difficult exam, some additional requirements for special education students and completions, and stricter passing requirements. Details on the requirements are provided in Appendix Table 2. As a result of the

⁸ Schools that received an L were subject to additional oversight and students were given the right to transfer to other public schools, although state law did not impose financial consequences. There was a risk of closure for being rated L for two or more consecutive years, although this affected very few schools as only 0.3% of all schools received an L rating two years in a row from 1998-99 to 2001-02. Schools receiving an E or R rating, along with some A schools that made large gains, were eligible for very small financial awards. In 2000-2001, which was the last year the award system was fully funded, the award was \$7.20 per enrolled student up to a maximum award of \$5000 per school, a negligible amount compared to average per-student expenditure of \$5,490.

⁹ Note that while the name of the lowest rating was changed to "Academically Unacceptable" we will continue to refer to it as L to maintain consistency throughout the paper.

increased stringency, many schools experienced a drop in their rating after the transition. Punishments for low performing schools under TAKS were strengthened, and include an option to “reconstitute” a school via mass layoffs and rehiring if a school receives an L for two consecutive years. The important element of the transition from TAAS to TAKS is that the new test was given to all students in 2003, but no school accountability grades were released. Thus our rating shock strategy is able to examine how districts and schools financially respond to the 2003 test score results, which presumably do not reflect a change in school quality since schools did not know how their students would perform on the new TAKS test.

3. Empirical Methodology

Our examination of school district behavior is through the total operating expenses in each school, which reflects the school district’s allocation strategy.¹⁰ On the other hand, we view the allocation of funds among categories of expenditure within each school as a reduced form measure reflecting the combined decisions of both the school district, and the individual school principals. Our rating shock strategy exploits the potential drop in rating facing schools as a result of the new exam in TAKS, compared to the older one in TAAS. Hence, the rating shock assesses how schools respond to an entirely new accountability system. We also estimate a model exploiting the sharp discontinuities between each rating, by estimating a regression discontinuity model comparing financial allocations to schools on either side of the rating boundary. Unlike the ratings shock model, the R-D analysis whether schools and districts respond to the marginal

¹⁰ The district’s allocation decision can be a combination of formulas, explicit exceptions, or an administrative decision process.

changes in the ratings – that is whether year-to-year perturbations in ratings matter to a school or district.

3.1. Rating Shock Strategy

Our first empirical strategy to determine the budgetary response to school accountability ratings is to examine the transition period between the first accountability system in Texas, TAAS, to the newer system, TAKS. In particular, we test how schools that faced a rating reduction in 2002-03 were either granted new resources by their school districts, and/or reallocated their budget during 2003-04, the first year in which the new TAKS system was fully implemented.¹¹ Specifically, schools did not know how their students would perform on the new TAKS exam, and they were explicitly given a year to adjust. The rating shock estimation strategy we use exploits the plausibly exogenous new testing information to produce unbiased estimates of budgetary and staffing responses.

While administrators were not provided with explicit ratings during the transition year, by this time the administrators at both the school and school district level were experienced with accountability systems, and thus were likely able to approximate their rating from the testing information. Hence, to implement our rating shock strategy we calculate what each school's rating would have been had one been assigned in the gap year based on the scores from the new exam.¹² The rating shock is therefore a strong test of the response to the ratings, because the

¹¹ This is similar to a strategy first used by Figlio and Kenny (2009).

¹² There was also uncertainty because the new criteria for translating pass rates to accountability ratings were not yet known. Nonetheless, we assume that school districts were able to approximate the new requirements prior to finalizing their expenditure decisions; hence we use

change in rating is not correlated with a change in the actual output of the school. On the other hand, the rating shock is only for a single year, and the analysis is limited because very few schools saw an increase in ratings due to the greater rigor of the new exam.

The regression using the rating shock examines the change in total expenditures, categorical expenditures, or staffing from the 2002-03 school year to the 2003-04 school year based on whether the estimated accountability rating in 2002-03 fell from its rating in 2001-02, the last year of the TAAS system.¹³ Hence we estimate:

$$\begin{aligned}
 (1) \Delta R_{i,2003-04} = & \alpha + \beta_1 Drop_{i,2002-03} \times Rating_{i,2002-03}^L + \beta_2 Drop_{i,2002-03} \\
 & \times Rating_{i,2002-03}^A + \beta_3 Drop_{i,2002-03} \times Rating_{i,2002-03}^R \\
 & + \beta_4 Increase_{i,2002-03} + \sum_{r \in L,A,R} \sum_{t=1999-2000}^{2001-02} \gamma_{jt} Rating_{i,t}^r + \mathbf{X}_i \boldsymbol{\Omega} + \varepsilon_i
 \end{aligned}$$

where ΔR is the change in resources per student in a given category from the transition year (2002-03) to the next, for either expenditures or employees in school i . \mathbf{X} is a set of school characteristics including the percent of enrollment in each grade, percent of enrollment by racial category, and the percent of enrollment identified as economically disadvantaged, LEP, gifted, special education, and vocational education. $Rating^L$, $Rating^R$, and $Rating^E$ are indicator variables for whether school i received a (calculated) rating of L, R, or E in the subscripted (transition) year. Our measure of a rating shock comes from the interaction of the predicted ratings in 2002-03 with $Drop$ - a dummy variable which equals one if the calculated rating in 2002-03 is lower

the new (2003-04) rules to estimate the 2002-03 ratings.

¹³ Due to the new testing regime, applying the “required improvement” rules to the predicted ratings would be not be sensible as the passing rates on TAKS were considerably different from those under TAAS. Thus, we ignore required improvement here.

than the last TAAS rating in 2001-02. We also control for an increase in rating, so that we compare schools whose ratings fall to those whose ratings remain the same. Finally, we include the school rating in each of the three prior years.¹⁴ Hence our estimates compare schools that had equal ratings in 2001-02 but where one received the same (imputed) rating in 2002-03 while the other received a lower rating based on the new test. Table 1 shows that while only 16% of schools experienced a ratings drop in the last year of TAAS, the calculated rating from the first year of the TAKS exam suggests roughly two-thirds of schools would experience reduced ratings moving from the old to the new systems, while less than 3% of schools were calculated to receive a higher rating.

The value of the rating shock estimation strategy, therefore, is that while the new accountability system was substantially different than the old, school administrators are nonetheless experienced in responding to the incentives inherent in any accountability system. The information school administrators did not have with the new system, however, is how well students would do on the test. Thus the gap year testing provided a baseline from which administrators could formulate policy changes to alter their students' test performance. One set of policy responses is within the school, as they learn to "teach to the test" or other strategies. The other set of potential responses is whether school districts reallocate funds between schools, or even within schools. It is this last question that motivates our work here.

The estimates from equation (1) provide information about the extent to which school districts re-allocate resources among schools as part of the policy response environment. That is, in addition to whatever actions school officials do within the school, such as adjust the

¹⁴ Our estimate of the effect of a rating drop is not sensitive to the inclusion of these past ratings.

curriculum to the demands of the new exam, an additional source of administrative behavior is illustrated by budgetary allocations. Presumably, districts will do so only if the relative performance of schools is likely to change, so that if there is a preferred distribution of accountability results, the change in resources will be the school district response to achieve that preferred distribution. The research advantage of the Texas changes is that the new system is substantially different from the old, thus making it likely that the relative distribution of ratings within a district will change.

3.2. Regression Discontinuity Strategy

Our second strategy to analyze the potential allocation effects is to utilize a regression discontinuity (RD) design to test for annual changes in resource allocations. This test will use data from 1997-98 through 2006-07, excluding the transition year of 2002-03, by examining the resource allocation response for schools which fall on one side of a rating boundary compared to those just on the other side. This strategy will yield a local average treatment effect, and will demonstrate resource responses to ratings that provide little information (e.g. the district knows that the schools are essentially the same except for random differences) as opposed to the ratings shock which potentially provides considerably more new information. School districts will re-allocate resources in response to accountability grades if districts use resources as an incentive mechanism, or if districts believe resources can alter the distribution of accountability grades across the district. If, on the other hand, schools and districts respond to the long run rating of a school, and believe that random differences in students for schools which are near the boundary

are beyond the control of the actions of the school, then the RD may show essentially no annual reallocations even if the rating shock strategy demonstrates a response.

The key to the RD strategy is to define the margin that influences the response to the accountability rating. We use the number of students for whom a change in test score could modify the school's rating. We further restrict our definition of boundary to identify schools where changes in only a single subject/student group cell could change the rating.¹⁵ To implement this strategy, we calculate the accountability rating for each subject–group cell that meets minimum size requirements. Based on this calculation, we identify the subject-group cell that defines whether a school is on a rating boundary since only the lowest rated cell is determinative. A school which is marginally below a ratings boundary will have a single cell that has a lower rating than any other subject- group cell. If the number of students that pass the exam in this single cell were to rise by a sufficient amount, the school's rating would rise. The number of students that cause the rating to be below a boundary, N_b , would therefore be the number of additional students in a single subject-group cell that would have to switch from failing to passing the exam for a school to rise one rating. For dropouts and completions we use the number of students in each subject-group cell that need to stay in school to change the rating.¹⁶

The measure of N_b is therefore:

¹⁵ Note that, due to the maxi-min structure of the rating formula, the rating for any school above a border could fall as a result of performance dropping in just one subject-group.

¹⁶ For the 1997-98 and 1998-99 school years there was an attendance requirement as well. Missing the requirement would drop a school to an A rating but could not cause an L rating. Since it is unclear how to convert attendance into a per-student measure similar to those for the other inputs, we ignore it in our calculations. The impact of this is negligible as Figure 1 shows that there were virtually no schools that were incorrectly assigned a different rating than they actually received under TAAS.

$$(2) N_b = \sum_s \sum_g 1(\text{Rating}_{sg} = \text{Rating}) \times 1(\text{Size}_{sg} \geq \text{Min}_{sg}) \times N_{sg}$$

where s is the test subject or performance measure, g is the student group (by race, disadvantaged, or total), 1 is the indicator function, Rating is the school's accountability rating, Rating_{sg} is the rating for the subject-group, Size_{sg} is the number of students in the subject-group, Min_{sg} is the minimum number of students that need to be in the group for it to count towards the rating, and N_{sg} is the number of additional students in group g who need to pass performance measure s to achieve the next higher rating. The first indicator function limits the summation to subject-groups that, if they were to be rated individually, would have the same rating as the school as a whole.¹⁷ Note that since we restrict marginality to schools where only one subject-group is below the boundary, N_b defaults to N_{sg} for that cell.

Alternatively, a school that is on the positive side of a rating may have many subject-student cells, or just one, which have a rating equal to the school's rating. If the cell specific rating in any one cell falls, the school's rating will fall. We use our calculations to identify the subject-student group cell that is closest to the boundary and use this cell to determine N_a , the number of additional students the school could afford to have fail before their rating falls. As such, the second forcing variable, N_a , can be calculated as:

$$(3) N_a = \min_{s,g} (1(\text{Rating}_{sg} = \text{Rating}) \times 1(\text{Size}_{sg} \geq \text{Min}_{sg}) \times N_{sg})$$

¹⁷ Since the overall school rating is based on the lowest rated subject-group, no group that counts towards the rating would have a rating lower than the school-wide rating. Nonetheless, under TAKS schools are given a limited number of exceptions for groups that perform poorly. Since these subject-groups do not count towards the ratings we do not include them in our calculation of N_b .

where N_{sg} refers to the number of students in a subject-group cell who would need to switch from passing to failing the performance measure to reach the cutoff for the next lower rating. As before, the two indicator functions identify subject group cells that have ratings equal to the school's rating, and that meet the minimum size standards.¹⁸ Hence if a school is precisely at the boundary while receiving the higher rating both N_a and N_b equal zero.

We therefore conduct a regression-discontinuity of schools whose N_a or N_b are close to zero for each boundary. Specifically, we conduct local-linear regressions of the form

$$(4)R_{i,t+1} = \alpha + \beta_1 N_{b,it} + \beta_2 N_{a,it} + \beta_3 Above_{it} + \varepsilon_{it}$$

for school i in year t where $R_{i,t+1}$ is the school specific resource in category i (either dollars per student or inputs per student) the year after the school receives a rating, N_a and N_b as defined by (1) and (2) above, and *Above* is a dummy variable indicating whether the school is above the rating threshold. We use bandwidths selected through leave-one-out cross validation.¹⁹ As suggested in Lee and Lemieux (2009), we use a rectangular kernel that involves limiting the

¹⁸ In some years a school could achieve a higher rating using year-on-year increases in performance measures if they do not score high enough to meet the requirements, called required improvement (RI). In these cases we calculate N_b using whichever method - RI or standard - that brings that subject-group closer to the cutoff. Similarly, we use the RI calculation for N_a if a school achieves a higher rating due to RI in the marginal group.

¹⁹ Specifically, we select the margin that minimizes the mean squared error by repeatedly estimating the model for "all but one" of the observations with a wide variety of margins. Nonetheless, we have also estimated (4) using parametric techniques with a 5-order and a 3-order polynomial, and using bandwidths one unit higher and one unit lower than the cross-validation bandwidths. In all of these cases we find qualitatively similar results suggesting that our estimates are robust to the choice of bandwidths or functional form. These results are provided in the online appendix.

sample to narrow bands around the cutoffs without re-weighting the data.²⁰ Figure 1 shows the first stage after the adjustments described above. At each boundary, there is a clear discontinuity in the rating whereby almost 100% of schools with $N_a \geq 0$ and $N_b = 0$ get a higher rating, with the near opposite occurring on the other side of the boundary.²¹

To explore whether the reactions by schools and districts may differ across the ratings, we run (4) separately for each boundary - between L and A schools (the LA boundary), between A and R schools (the AR boundary), and between R and E schools (the RE boundary). Our results are estimates for β_3 , the impact of being above the given boundary.

4. Data

Our data covers all public schools in the state of Texas and comes from three datasets provided by the Texas Education Agency (TEA). First is the Academic Excellence Indicator System (AEIS) which provides data on staffing, enrollment, and student demographics. Second are the Public Education Information Management System (PEIMS) financial reports that provide expenditure data for each school by category.²² We examine total expenditures, categorical expenditures, and student-faculty ratios. Third are the TEA accountability reports that provide the data that is used to calculate accountability ratings.²³

²⁰ Lee and Lemieux (2009) argue that more complex kernels provide only marginal improvements in efficiency.

²¹ Given these results, we rely on a “strict” RD design. Nonetheless estimates using a “fuzzy” design were nearly identical.

²² We use the actual expenditures by schools rather than budgeted expenditures.

²³ All data sources are publically available on the website for the TEA.

Expenditure categories are mutually exclusive and include instruction; leadership, curriculum and staff development; counseling and social work services; and extra-curricular activities. The AEIS data provides full-time equivalent teacher counts which we convert into student-teacher ratios using enrollment. All expenditures are divided by enrollment to provide per-student measures and inflated to 2007 dollars.

The RD regressions are pooled across years within each accountability system. Since schools appear multiple times, we cluster standard errors by school. In all of our analyses we drop alternative schools, charter schools, “paired” schools, those with special analysis, and schools under 200 students.²⁴ After these restrictions, we have 26,500 school-year observations for TAAS and 23,168 for TAKS.

Table 3 presents the means of school characteristic data for each of the two accountability regimes. In general, schools with higher ratings are wealthier and have fewer minorities and special needs students. Table 4 provides summary statistics for resources in the year after a school receives a rating. Under both accountability regimes schools with lower ratings have higher expenditures than those with higher ratings, although this gap widens in TAKS. These differences generally hold across expenditure categories.

²⁴ Alternative schools targeted to specific groups of students operate under a separate accountability system. These schools along with charter schools also have separate state aid and budgeting rules. “Paired” schools are too small to apply accountability standards, and hence are assigned the rating of another school in the district. Schools that undergo “special analysis” are also too small and hence are analyzed under a subjective rating system. Schools with under five students per cell have their data masked, and so we drop those with less than 200 students to avoid errors in calculating margins.

5. Empirical Results

5a. Ratings Shock Results

The regression results in Table 5 are our main ratings-shock results and show the change in resources from the 2002-03 transition year to the first official year of TAKS (2003-04) as a response to the reduction in the school rating, differentiated by the rating to which a school falls. To avoid any spurious correlation from marginal movements in ratings, we control for the rating in the previous three years.²⁵

The evidence from Table 5 suggests school districts allocate \$75 per student to schools that experience a drop in their rating to A. This is over 1.2% of the total school budget of the average A school. The second column of the table also shows that virtually all of the funds are allocated to instructional expenses (the coefficient of 68.4 is not significantly different from 74.8). This 1.5% increase in instructional resources is found to generate a reduction in the student/teacher ratio of almost 0.3%. In a separate analysis, available upon request, we find no significant changes in enrollment due to the ratings drop, hence we interpret the student-teacher ratio results as indicating an increase in teachers. For schools whose predicted rating falls to L we find an insignificant change in total or instructional resources, but nonetheless there is also a decrease in the student-teacher ratio. Finally, for schools that fall to an R we find no statistically significant changes in any resource.

²⁵ When the earlier ratings are omitted, the estimated budget change is 40% smaller but at identical levels of statistical significance.

To test whether the results in Table 5 are permanent or temporary budget changes, we re-estimate the rating shock regression, but on the budget change three years following the change in accountability regimes, so we look at budget changes from 2002-03 to 2005-06 as a result of the change in rating.²⁶ These results are presented in Table 6, and show that changes in school budgets for schools with lower ratings are not significantly different from schools that maintain their ratings. These results suggest, therefore, that the budget changes we observe in the year following the new rating information were temporary resources, designed to assist schools in adapting to the new regime. School districts, however, appear to leave their overall budget allocation system across schools unchanged in the longer term.

A possibly surprising aspect of our results is that we do not estimate any temporary budget allocations for schools that fall to the lowest rating, L. One potential reason is that the new TAKS accountability system contains real penalties for schools that remain L for more than one year. Thus it may be that districts believe schools have sufficient incentive to improve, and do not need temporary resources.

In order to ensure that our estimates can be interpreted causally, we conduct a “placebo test” where we look at how the estimated transition year ratings changes “affect” changes in prior resource allocations, from 2001-02 to 2002-03. None of the expenditure estimates are statistically significant, although schools that drop to an A see a small increase in student teacher ratios of 0.1. We also test whether the results for the rating shock capture underlying student characteristics. We find no quantitatively important impacts on the levels of student

²⁶ The two year estimates are about halfway between those in Tables 5 and 6.

characteristics, or on whether changes in student characteristics affect school rating changes.²⁷

Results from both of these analyses can be found in the online appendix.

5b. Regression Discontinuity

The RD results suggest that under the older TAAS accountability system districts rewarded schools for higher ratings, although the evidence is not unassailable. The results further suggest that there is some variance over exactly how the rewards were spent by schools. The empirical results also show, however, that annual budgetary allocations do not respond to ratings under the new TAKS system.

The first column of Table 7 presents the RD results for total school expenditures, divided into Panel A for the early TAAS period and Panel B for the later TAKS regime. The first column of Panel A shows how school districts change total school budget allocations for academic year $t+1$ when the accountability rating is released late in the Spring of academic year t . Schools just above a rating border are estimated to receive ‘bonus’ financial allocations on all three borders, although only the AR (acceptable/recognized) border results are statistically significant at conventional levels. However, while estimates are statistically significant, they are economically small. The estimate of \$108 per student is less than 2% of the mean total expenditure of about \$6,000 (see Table 4), and is approximately equal to the insignificant award on the RE (recognized/exemplary) border. The reward on the LA (low performing/ acceptable) border is

²⁷ Potentially the most important change we find is an increase in LEP student share of 0.3 percentage points in response to the rating shock. While the state aid formula provides additional funds for LEP students, this estimate suggests that an increase in LEP share for the average school that drops to an A would only increase spending by \$2 per-pupil. Hence, the impact on the budget is negligible.

estimated to be much larger, \$496. While this amount is over 8% of the mean expenditure of A rated schools, its estimate is not precise. Figure 2 presents these same results graphically from a fitted polynomial using local linear regression, showing the relatively sharp discontinuity on all three borders, although also illustrating the noise around the estimates. We believe these estimates are preferred compared to estimates using the change in expenditure, because the specification in changes presumes a constant base growth rate.²⁸ Nonetheless, estimates using the change in expenditures are presented in the second column of Table 7, and show no significant effects.

Table 8 shows the effects of small perturbations in the bandwidth. Specifically, the estimates of the impact of success on the LA border in columns 1 (smaller bandwidth) and 3 (larger bandwidth) are found to be larger than the base estimates and are statistically significant (column 2, identical to column. 1 in Table 7). The estimates for the AR border are similar, in that both the base case and the larger bandwidth are found to be statistically significant, although not for the smaller bandwidth. The RE estimates are similar for all three tested bandwidths, although no effects are significant at conventional levels.

Table 9 explores the issue of timing. The accountability results are known to schools and districts in the spring of the academic year. It is possible that school districts could make budgetary changes in the same year (year t) rather than in the subsequent year ($t+1$) as assumed

²⁸ To control for allocations by formula, we also tried using the residuals from a regression of total expenditures on school characteristics and district fixed-effects as the endogenous variable for the RD, but with qualitatively identical effects. The regressions include share of enrollment in each grade, bilingual education, limited English proficiency, economically disadvantaged, gifted, special education, and year and district fixed effects.

in the initial specification.²⁹ We find for Panel A that the results are similar when tested against year t as opposed to year $t+1$ budgets. Nonetheless, there is no significant relationship with the prior year's budgets as shown in column 3. This provides support that our R-D results are picking up actual budget changes as a result of ratings and are not spuriously correlated with general budgeting patterns.

The results in column 2 suggest the possibility that budgetary rewards are granted in some districts in year t , and in other districts in year $t+1$, in which case it is possible our coefficient estimate from Table 7 only represents on average about half of the rewards. We test this idea in column 4, by showing the estimates for the sum of the budget over two years (t and $t+1$). This result suggests the awards for success on the LA border could be equal to about 15% of the total budget. The pattern of the results is similar for the AR and RE borders, in that the combined change is about twice the change in the baseline result of Table 5. As before, we find the AR border effects are statistically significant, while the RE results are not.

We conclude from the sensitivity analysis that at least some school districts in Texas used the TAAS accountability system to reward success at reaching certain ratings. If we accept the statistical results at face value, the LA border is much more important to administrators than the AR or RE border as the rewards are about five times larger.

Table 10 presents results for some of the categories of expenditures. These results are less clearly the behavior of the school districts, as presumably school officials, and especially

²⁹ School districts in Texas have the option of closing their fiscal years on July 1 or on September 1. Districts receive testing results in May while complete accountability results are revealed in June under TAAS and July under TAKS. This provides districts an opportunity to distribute funds prior to the new fiscal year.

principals, have a significant effect on allocation as well. Further, the estimated effects will be evident only to the extent that school districts throughout Texas behave similarly. Nonetheless, we find that a large share of the estimated reward (83%) for schools just over the LA border spend their supplemental resources on instruction, with most of the remainder on extra-curricular activities. In contrast, schools on the AR and RE border instead spend their much smaller extra resources on counseling, plus administration and training.

While we therefore have observed some small annual budgetary response to school ratings in the early system, TAAS, we do not find any such response under the new system TAKS. The Panel B results in Table 7 show no statistical significance on any border, and in fact the point estimates in column 1 are negative for both the LA and AR borders. Figure 3 graphically presents the same information, that the discontinuities for budget reallocations in TAAS are not evident on the TAKS boundaries. The other column of Table 7 shows no appreciable change when we use the change in expenditures as the endogenous variable. Table 8 shows little sensitivity of these results to the choice of bandwidth. Table 9 likewise shows that timing relative to the rating year does not alter the results. We believe it unlikely that the stricter environment in TAKS is responsible for the change in budgetary allocations, in that while L schools faced a larger administrative penalty (for not being A), other schools did not. Thus it seems more likely that school districts abandoned the implicit reward structure because they found it was not successful, although our estimates cannot distinguish the actual cause of this change in behavior.

We perform a series of diagnostic sensitivity analysis of the above RD results, and find that our results hold up relatively well. We test for discontinuities in the demographic variables,

and mostly find only minor perturbations, to which we respond by including student characteristics in the RD regressions and find qualitatively similar results. One exception is the A/R boundary for TAKS where there are a number of significant discontinuities; hence this estimate should be interpreted with caution. Nonetheless, the A/R estimates are consistent with the other TAKS boundaries, suggesting that this is not a substantial issue. Looking at the distributions around the boundaries, we see some evidence of potential manipulation within a very small range – only one student - around the border. Part of this discontinuity is generated by exceptions and required improvement rules in the ratings formulas as is shown in the online appendix figures. Nonetheless, we check if our estimated results are sensitive to excluding observations very close to the boundaries and find that they are not, except for R/E in TAKS which becomes positive and significant. All of these analyses are reported in an online appendix.

6. Summary and Conclusion

Our objective in this research has been to expand the literature on accountability by understanding the extent to which school districts, and schools, allocate resources in response to the school accountability ratings. If districts respond to accountability by reallocating resources, then it suggests that accountability and its associated testing had either lowered the cost of information, or provided new information to parents and policy makers. Our results suggest that the implementation of a new accountability rating system provides a temporary shock to the education system, but schools and districts respond little in terms of budget allocations in response to year-to-year changes in ratings.

Using data from the state of Texas and the exogenous shock from the transition between accountability systems (the rating shock), we find that extra resources were directed towards schools which had projected drops in their rating under the new system. In particular, schools whose ratings fall to the second lowest grade during the transition between the old and new systems saw an increase of their budget of 1.2%, most of which went to instruction. Nonetheless, these resources were temporary, so that by the third year the budget allocations across schools had reverted to their original pattern. Thus school districts, already experienced with operating under an accountability regime, apparently perceive that temporary resources are useful to schools, which suggests that learning to respond to a new accountability system requires resources. That these resources are only provided temporarily suggests that once learned, the new system can be integrated with existing resources.

The RD analysis of schools on the rating boundary, using the annual panel data, (weakly) identified limited rewards for achieving higher ratings in the early accountability system (TAAS).³¹ Like the rating shock estimates, we find most of these resources were directed toward instruction. Under the later TAKS regime, however, we find that annual budgetary allocations do not respond to accountability ratings.

We believe the overall pattern of results suggests that schools and school districts have learned that it is not useful to respond to marginal changes in accountability ratings. That is, the

³¹ We do not believe our results are confounded with No Child Left Behind. The requirements for making AYP are different from the Texas accountability requirements. Hence our RD estimates should sufficiently target our estimates on the state rules. Second, missing AYP is mainly a concern for schools that are at risk of receiving low ratings. Our RD estimates of the marginal impact of receiving a higher rating during the TAKS regime under which NCLB was in effect, however, are generally similar across ratings.

RD analysis shows that under the initial Texas system there appears to have been some attempts to build marginal incentives to changes in ratings. The RD analysis also shows these incentives disappear under the new system. Instead, as shown by the rating shock analysis, schools and districts were willing to invest in schools that appear to have needed assistance in adapting to the new accountability system. Because we show these responses were not permanent, however, and consistent with the lack of incentives in the RD analysis, we conclude that schools and districts do not consider marginal incentives important even in an accountability regime.

References

- Bacolod, Marigee, John DiNardo, and Mireille Jacobson, "Beyond Incentives: Do Schools Use Accountability Rewards Productively?" NBER Working Paper 14775, 2009
- Carnoy, Martin and Susanna Loeb. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Education Evaluation and Policy Analysis*, 24, 2003, pp. 305 - 331.
- Cullen, Julie Berry and Randall Reback. "Tinkering Toward Accolades: School Gaming under a Performance Accountability System." NBER Working Paper #12286, June 2006.
- Chakrabarti, Rajashri. "Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs." Federal Reserve Bank of New York Staff Report no. 315, January 2008.
- . "Vouchers, Public School Response and the Role of Incentives: Evidence from Florida." Federal Reserve Bank of New York Staff Report no. 306, October 2007.
- Chiang, Hanley. "How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics*, 93, 2009, pp. 1045-1057.
- Dee, Thomas and Brian Jacob. "The Impact of No Child Left Behind on Student Achievement." NBER Working Paper #15531, 2009.
- Figlio, David and Lawrence Kenny. "Public Sector Performance Measurement and Stakeholder Support." *Journal of Public Economics*, 93, 2009, pp. 1069-1077.
- ___ and Cecelia Elana Rouse. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics*, 90, 2006, pp. 239 - 255.
- ___ "Testing, Crime, and Punishment." *Journal of Public Economics*, 90, 2006, pp. 837 - 851.
- ___ and Lawrence Getzler, ""Accountability, Ability, and Disability: Gaming the System," in *Advances in Microeconomics*, Timothy Gronberg (ed), Elsevier, 2006.
- ___ and Joshua Winicki. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics*, 89, 2005, pp. 381 - 394.
- ___ and Maurice E Lucas, "What's in a Grade? School Report Cards and the Housing Market," *American Economic Review*, 94, June, 2004, pp. 591-604.

Hanushek, Eric A. and Margaret E. Raymond. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management*, 24, 2005, pp. 297 - 327.

____ and ____ "The Effect of School Accountability Systems on the Level and Distribution of Student Achievement." *Journal of the European Economic Association*, 2, April-May 2004, pp. 406 - 415.

Jacob, Brian A. "Accountability, Incentives, and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics*, 89, 2005, pp. 761 - 796.

_____"Getting Inside Accountability: Lessons from Chicago." *Brookings-Wharton Papers on Urban Affairs*, 2003, pp. 42 - 70.

Lee, David S., "Randomized Experiments from Non-Random Selection in U.S. House Elections," *Journal of Econometrics*, 142, 2008, pp. 675-97.

Lee, David S. and Thomas Lemieux. "Regression Discontinuity Designs in Economics." NBER Working Paper #14723, February 2009.

Neal, Derek and Diane Whitmore Schanzenbach. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics*, 92, 2010, pp. 263 - 283.

Reback, Randall. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics*, 92, 2008, pp. 1394 - 1415.

Rockoff, Jonah and Lesley J. Turner. "Short-Run Impacts of Accountability on School Quality" *American Economic Journal: Economic Policy*, 2, 2010, pp.119 - 147.

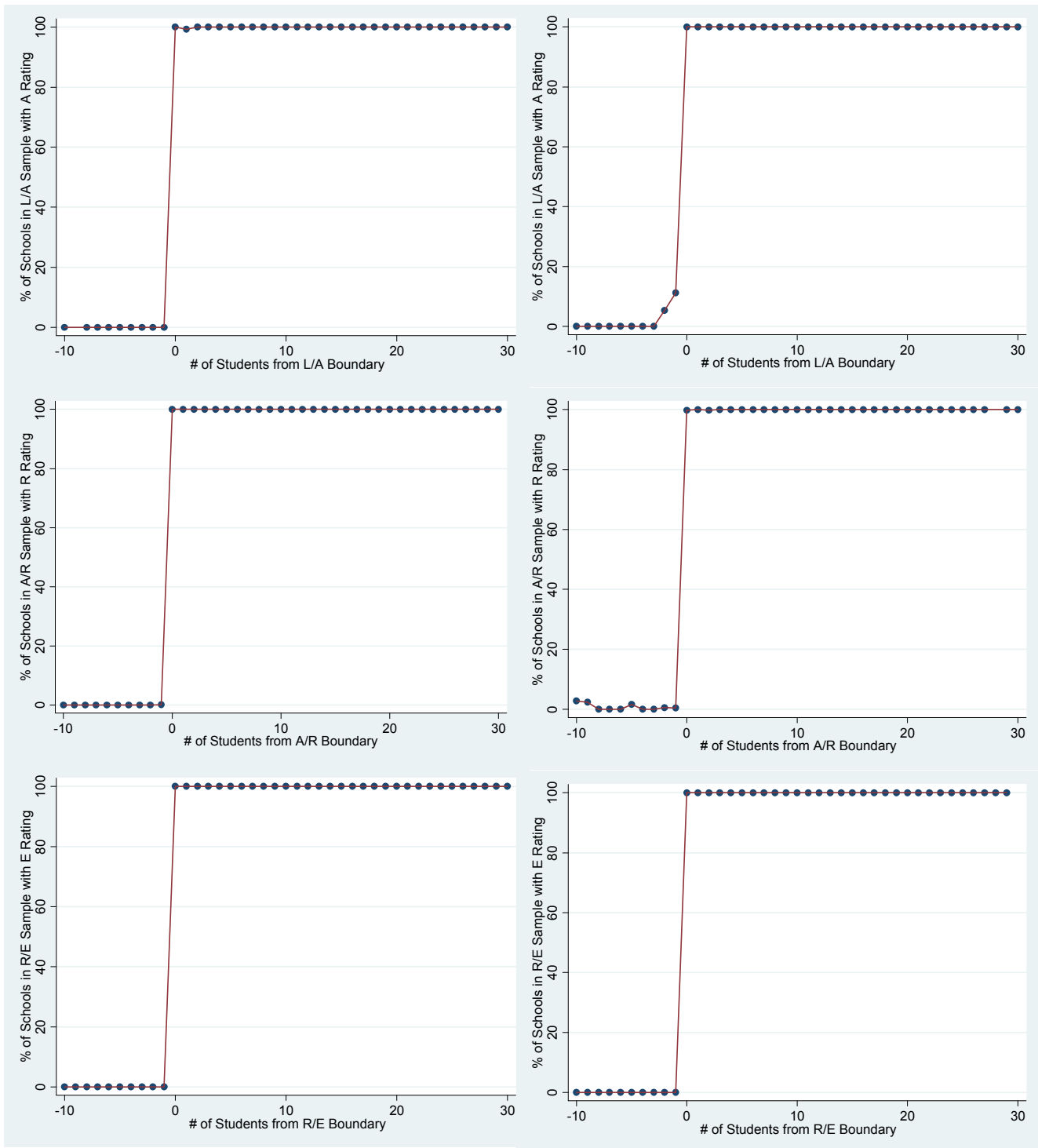
Rouse, Cecelia Elena, Jane Hannaway, Dan Goldhaber and David Figlio. "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." NBER Working Paper #13681, December 2007.

Silva, Fabio and Jon Sonstelie, "Did Serrano Cause a Decline in School Spending?", *National Tax Journal*, 48, June, 1995, 199-215.

Figure 1: Changes in Accountability Rating at Rating Boundaries

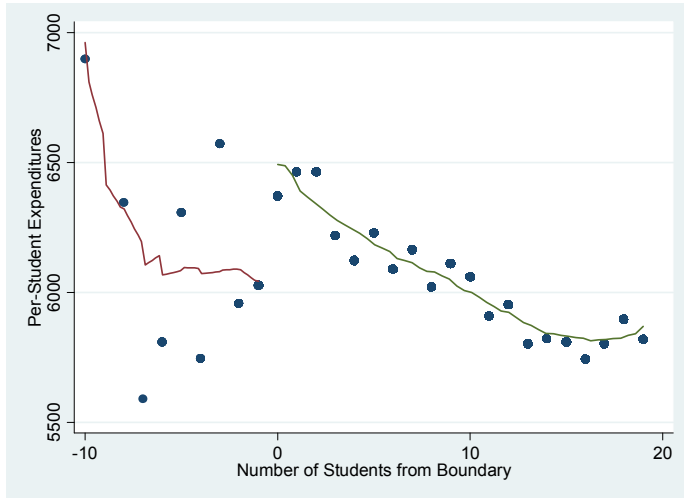
A. TAAS

B. TAKS

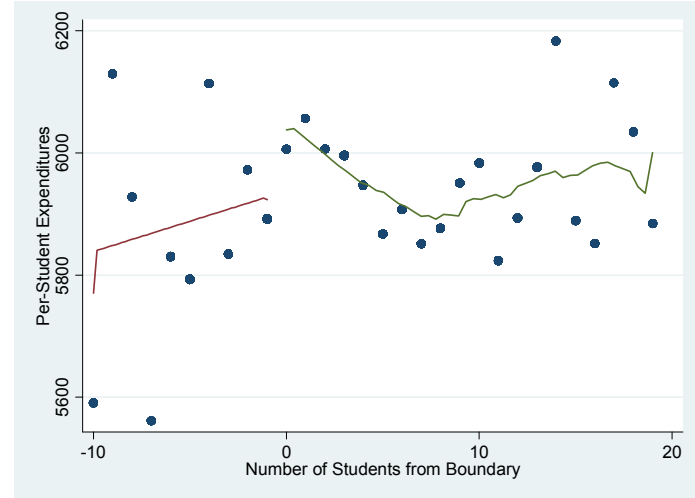


Sample is restricted to schools one rating above boundary or one rating below boundary where they miss the higher rating in only one subject-group. Schools that are paired with another school, received a rating on appeal, were identified as receiving "special analysis", or had fewer than 200 students are also excluded.

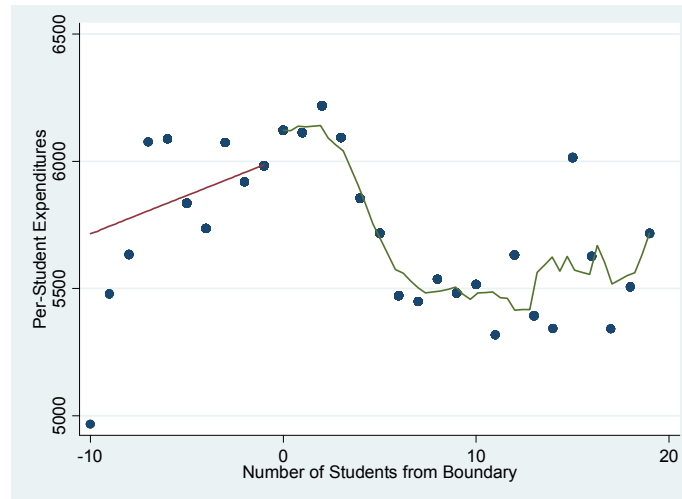
Figure 2 - Total Per-Student Expenditures in TAAS



L/A



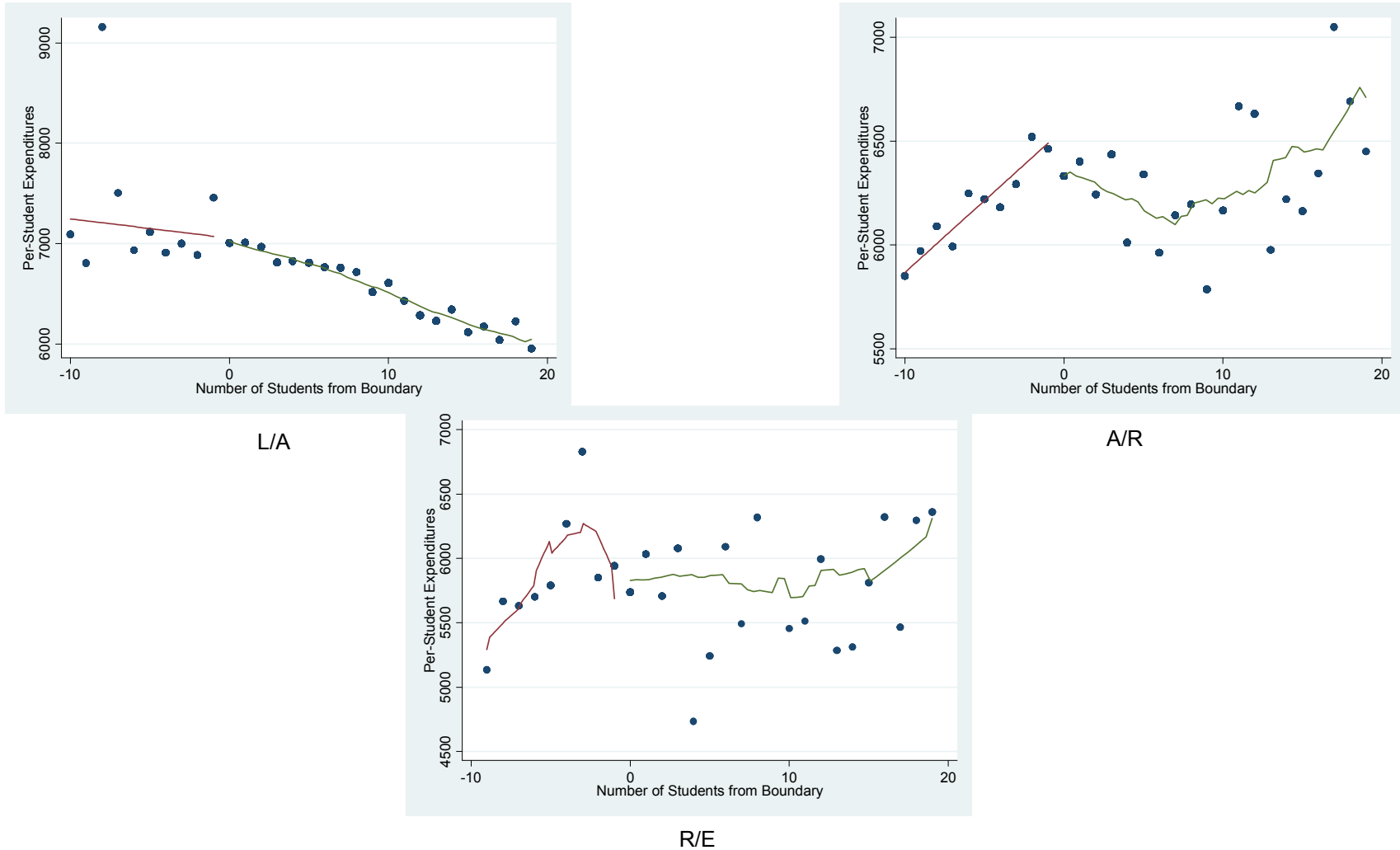
A/R



R/E

Sample is restricted to schools one rating above boundary or one rating below boundary where they miss the higher rating in only one subject-group. Schools that are paired with another school, received a rating on appeal, were identified as receiving "special analysis", or had fewer than 200 students are also excluded. All expenditures in 2007 dollars.

Figure 3 - Total Per-Student Expenditures in TAKS



Sample is restricted to schools one rating above boundary or one rating below boundary where they miss the higher rating in only one subject-group. Schools that are paired with another school, received a rating on appeal, were identified as receiving "special analysis", or had fewer than 200 students are also excluded. All expenditures in 2007 dollars.

Table 1: Distributions of Accountability Ratings

	Accountability Rating				Rating Changes		
	Low	Acceptable	Recognized	Exemplary	% Lower	% Same	% Higher
TAAS							
1997-98	0.7%	55.1%	27.1%	17.2%	-	-	-
1998-99	0.9%	51.1%	29.9%	18.2%	14.9%	66.5%	18.6%
1999-00	1.6%	45.9%	32.0%	20.5%	14.8%	64.7%	20.5%
2000-01	0.7%	38.2%	36.5%	24.7%	13.2%	61.9%	24.9%
2001-02	1.3%	31.8%	37.4%	29.5%	16.4%	58.8%	24.8%
Transition Year (2002-03)							
Estimate based on 2003-04 Rules [†]	9.6%	69.7%	18.0%	2.7%	63.1%	33.4%	3.5%
TAKS							
2003-04 [‡]	1.0%	53.4%	37.9%	7.7%	42.6%	39.8%	17.6%
2004-05	3.1%	64.3%	28.2%	4.4%	25.8%	66.9%	7.3%
2005-06	3.6%	46.6%	41.8%	8.1%	8.3%	64.2%	27.5%
2006-07	3.4%	53.6%	34.0%	9.1%	20.3%	65.0%	14.8%

[†] Only for schools that have a regular rating in 2001-02. We do not account for required improvement as the implementation of a new exam makes such a calculation inappropriate

[‡] Ratings changes are relative to 2001-02.

Sample is limited to schools that received an L, A, R, or E rating. For changes the school must have received a rating in both years. Schools that received ratings on appeal, were paired with another school, were identified as having undergone "special analysis," or have fewer than 200 students are excluded.

Table 2: Rating Transitions Under Different Accountability Regimes

A. TAAS (1997-2002)				
	Year t			
Year t - 1	L	A	R	E
L	11.3%	82.3%	6.4%	0.0%
A	2.0%	69.1%	25.5%	3.4%
R	0.1%	24.6%	53.6%	21.7%
E	0.0%	4.1%	24.6%	71.3%
B. Transition				
Last Year of TAAS (2001-02)	First Year of TAKS (2003-04)			
	L	A	R	E
L	5.6%	80.6%	12.5%	1.4%
A	1.9%	75.9%	21.3%	1.0%
R	0.2%	59.3%	38.0%	2.5%
E	0.1%	29.2%	53.8%	17.0%
C. TAKS (2004-2007)				
	Year t			
Year t - 1	L	A	R	E
L	21.8%	73.3%	4.8%	0.3%
A	4.7%	74.0%	20.9%	0.3%
R	0.2%	33.8%	57.9%	8.1%
E	0.0%	3.1%	40.5%	54.4%

Sample is limited to schools that received an L, A, R, or E rating in both years. Schools that received ratings on appeal, were paired with another school, were identified as having undergone "special analysis," or have fewer than 200 students are excluded.

Table 3 - School Characteristics by Rating

	TAAS				TAKS			
	Low	Acceptable	Recognized	Exemplary	Low/ Unacceptable	Acceptable	Recognized	Exemplary
% Asian	1.1 (2.3)	1.8 (3.4)	2.0 (4.1)	3.2 (5.3)	0.7 (1.5)	1.9 (3.5)	3.1 (5.5)	6.4 (8.5)
% Black	29.5 (26.8)	18.6 (22.6)	11.1 (16.6)	7.1 (11.7)	30.2 (29.8)	15.6 (19.4)	11.0 (14.7)	7.3 (10.2)
% Hispanic	48.4 (30.1)	43.8 (31.7)	39.4 (32.3)	24.5 (28.6)	52.6 (32.3)	49.3 (31.5)	39.6 (30.9)	22.0 (24.7)
% White	20.8 (23.6)	35.6 (29.0)	47.1 (31.3)	64.9 (29.1)	16.4 (22.4)	32.8 (29.0)	45.9 (30.4)	63.9 (26.1)
% Economically Disadvantaged	70.9 (22.5)	59.9 (25.0)	51.0 (26.4)	31.9 (28.0)	76.9 (19.0)	62.4 (24.9)	51.6 (27.2)	28.1 (28.0)
% LEP	23.5 (23.1)	16.6 (19.4)	13.3 (17.7)	8.0 (14.3)	19.6 (22.3)	17.7 (19.8)	14.8 (18.0)	9.5 (13.9)
% Special Ed	18.3 (26.4)	14.1 (23.7)	12.0 (23.8)	13.6 (26.6)	28.8 (30.4)	19.9 (28.7)	6.5 (17.7)	1.8 (10.7)
% Gifted	21.6 (22.5)	14.9 (18.4)	12.1 (16.8)	7.2 (13.5)	18.1 (21.7)	16.4 (19.0)	13.8 (17.3)	8.7 (13.2)
Enrollment	812 (503)	749 (502)	633 (410)	601 (383)	781 (519)	783 (567)	606 (336)	595 (226)
Observations	285	12,111	8,614	5,550	593	12,537	7,985	1,537

Sample is limited to schools that received an L, A, R, or E rating. Schools that received ratings on appeal, were paired with another school, were identified as having undergone "special analysis," or have fewer than 200 students are excluded.

Table 4 - Mean School Resources by Rating

Resources in Year t + 1	A. TAAS (year t)			
	Low	Acceptable	Recognized	Exemplary
Total Operating Exp.	6,397 (2,011)	5,956 (1,238)	5,980 (1,257)	5,969 (1,505)
Instructional Exp.	4,612 (1,699)	4,376 (843)	4,461 (832)	4,470 (985)
Admin & Training Exp.	679 (308)	585 (210)	562 (240)	550 (340)
Counseling Exp.	249 (127)	230 (106)	216 (98)	210 (102)
Extra-Curricular Exp.	115 (191)	117 (210)	128 (246)	146 (282)
Student-Teacher Ratio	15.2 (2.6)	14.9 (2.5)	14.7 (2.5)	14.7 (2.5)
Resources in Year t + 1	B. TAKS (year t)			
	Low/ Unacceptable	Acceptable	Recognized	Exemplary
Total Operating Exp.	7,089 (1,567)	6,343 (1,214)	6,039 (1,232)	5,766 (982)
Instructional Exp.	4,982 (957)	4,581 (790)	4,518 (827)	4,412 (669)
Admin & Training Exp.	778 (280)	625 (179)	577 (179)	540 (149)
Counseling Exp.	288 (128)	243 (109)	217 (100)	201 (91)
Extra-Curricular Exp.	214 (315)	180 (293)	80 (188)	36 (89)
Student-Teacher Ratio	14.0 (2.4)	14.6 (2.3)	14.8 (3.1)	15.2 (1.7)

Sample is limited to schools that received an L, A, R, or E rating. Schools that received ratings on appeal, were paired with another school, were identified as having undergone "special analysis," or have fewer than 200 students are excluded. All expenditures are in 2007 dollars.

Table 5: "Ratings Shock" Estimates of Resources on Ratings Changes in Transition Year

Δ from 2002-03 to 2003-04 \rightarrow	Total	Instruction	Admin & Training	Counseling	Extra-Curricular	Student-Teacher Ratio
	(1)	(2)	(3)	(4)	(5)	(6)
Estimated Rating Falls to L	-12.5 (38.9)	16.1 (30.7)	-0.5 (5.4)	-2.4 (3.3)	3.0 (3.4)	-0.14* (0.07)
Estimated Rating Falls to A	74.8** (29.2)	68.2*** (22.7)	7.8 (4.8)	2.3 (2.7)	-4.1 (2.8)	-0.29*** (0.06)
Estimated Rating Falls to R	-7.9 (36.5)	12.3 (28.2)	-1.5 (5.7)	1.3 (3.5)	-3.1 (3.2)	-0.10 (0.08)
Estimated Rating Increases	-18.8 (74.2)	-13.3 (64.1)	0.8 (11.5)	-2.7 (6.7)	0.5 (4.2)	0.14 (0.13)
Low Rating in 2001-02	193.5 -151.1	78.1 (131.5)	77.6*** (23.9)	27.9*** (10.3)	9.8 (7.2)	-0.48** (0.22)
Acceptable Rating in 2001-02	169.7*** (40.6)	152.9*** (31.4)	22.1*** (6.7)	8.1** (4.0)	-3.6 (4.1)	-0.33*** (0.08)
Recognized Rating in 2001-02	62.6** (24.5)	47.2** (18.5)	12.4*** (3.8)	4.6** (2.3)	0.5 (2.5)	-0.05 (0.05)
Low Rating in 2000-01	-209.2 (185.8)	-255.5 (157.8)	25.3 (34.6)	6.9 (9.8)	-3.2 (8.0)	-0.23 (0.23)
Acceptable Rating in 2000-01	-28.4 (30.8)	-41.6* (23.6)	-5.0 (5.0)	-2.3 (3.0)	9.0** (3.6)	0.04 (0.06)
Recognized Rating in 2000-01	21.2 (25.2)	-3.4 (19.4)	-0.3 (3.9)	-1.6 (2.5)	9.8*** (3.0)	-0.03 (0.05)
Low Rating in 1999-00	-118.8 (109.4)	-119.7 (91.0)	3.3 (15.6)	-4.1 (7.2)	-7.8 (5.2)	0.34** (0.17)
Acceptable Rating in 1999-00	-70.8** (29.8)	-39.2* (22.6)	-8.3* (4.7)	-1.3 (3.0)	-8.7** (3.5)	0.12* (0.06)
Recognized Rating in 1999-00	-90.0*** (26.0)	-63.5*** (19.8)	-3.1 (4.0)	-2.3 (2.7)	-5.0 (3.3)	0.12** (0.05)
Observations	4958	4958	4958	4958	4958	4957

Sample is limited to schools that received an L, A, R, or E rating in 1999-00, 2000-01, 2001-02 & 2003-04. Schools that received ratings on appeal, were paired with another school, were identified as having undergone "special analysis," or have fewer than 200 students are excluded. Regressions also include controls for % of students in each grade level, % black, % Hispanic, % Asian, % Native American, % economically disadvantaged, %LEP, % special education, % gifted, and % vocational. Robust standard errors clustered in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 6: "Ratings Shock" Estimates of Long-Term (3 Years) Resouces on Ratings Changes in Transition Year

Δ from 2002-03 to 2005-06 \rightarrow	Total	Instruction	Admin & Training	Counseling	Extra-Curricular	Student-Teacher Ratio
	(1)	(2)	(3)	(4)	(5)	(6)
Estimated Rating Drops to L	49.0 (52.1)	47.8 (37.2)	18.4* (9.5)	5.0 (5.2)	-0.7 (4.7)	-0.14 (0.10)
Estimated Rating Drops to A	17.8 (41.4)	29.7 (30.6)	7.5 (7.2)	3.0 (4.3)	-5.5 (3.9)	-0.10 (0.08)
Estimated Rating Drops to R	-73.0 (50.1)	-28.5 (37.0)	-9.6 (8.5)	-2.2 (5.3)	-4.1 (4.4)	0.04 (0.10)
Estimated Rating Increases	51.7 (99.0)	9.7 (79.8)	9.3 (13.5)	1.0 (9.1)	2.7 (4.9)	0.14 (0.17)
Low Rating in 2001-02	-158.0 (201.7)	-129.4 (163.3)	46.4 (28.6)	11.5 (13.0)	-5.4 (8.8)	-0.07 (0.29)
Acceptable Rating in 2001-02	130.0** (58.4)	144.4*** (42.2)	34.1*** (10.0)	8.6 (6.0)	-10.8* (5.9)	-0.08 (0.11)
Recognized Rating in 2001-02	72.1** (35.3)	73.5*** (25.1)	12.8** (5.7)	4.4 (3.4)	-5.6 (3.6)	0.04 (0.06)
Low Rating in 2000-01	-490.2** (203.3)	-421.8** (170.4)	-32.9 (41.8)	8.4 (13.5)	-8.8 (12.1)	0.14 (0.28)
Acceptable Rating in 2000-01	-9.3 (44.3)	-25.0 (32.9)	-6.2 (7.7)	-2.2 (4.5)	3.8 (4.7)	0.02 (0.08)
Recognized Rating in 2000-01	16.7 (35.9)	-3.9 (27.1)	0.1 (6.0)	0.6 (3.7)	2.7 (3.7)	-0.06 (0.07)
Low Rating in 1999-00	-164.0 (141.3)	-110.2 (119.4)	-15.9 (19.7)	-18.1** (8.8)	4.6 (8.0)	0.32 (0.21)
Acceptable Rating in 1999-00	-119.7*** (41.8)	-71.9** (30.9)	-17.8** (7.0)	-7.8* (4.4)	-5.5 (4.8)	0.08 (0.08)
Recognized Rating in 1999-00	-99.2*** (37.4)	-73.7*** (27.6)	-7.5 (6.2)	-5.8 (4.0)	-3.5 (4.5)	0.12* (0.07)
Observations	4882	4882	4882	4882	4882	4881

Sample is limited to schools that received an L, A, R, or E rating in 1999-00, 2000-01, 2001-02 & 2003-04. Schools that received ratings on appeal, were paired with another school, were identified as having undergone "special analysis," or have fewer than 200 students are excluded. Regressions also include controls for % of students in each grade level, % black, % Hispanic, % Asian, % Native American, % economically disadvantaged, %LEP, % special education, % gifted, and % vocational. Robust standard errors clustered in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 7: Regression Discontinuity: The Effect of Rating on Per-Student Total Operating Expenditures

Year t Rating ↓	A. TAAS		B. TAKS	
	Expenditures in Year $t+1$	Change in Expenditures from t to $t+1$	Expenditures in Year $t+1$	Change in Expenditures from t to $t+1$
	(1)	(2)	(3)	(4)
Low / Acceptable	496.1 (308.7)	-9.0 (94.9)	-117.1 (204.0)	-68.2 (100.5)
Obs	779	374	3086	1803
Bandwidth Below	4	19	17	19
Bandwidth Above	4	2	5	3
Acceptable / Recognized	108.2** (55.2)	20.0 (29.0)	-31.7 (50.7)	37.6 (26.9)
Obs	8731	6511	6626	7332
Bandwidth Below	9	16	17	10
Bandwidth Above	5	3	4	5
Recognized/Exemplary	113.7 (77.9)	-15.3 (32.9)	105.8 (200.4)	8.2 (41.5)
Obs	3777	5342	1796	2190
Bandwidth Below	19	15	3	18
Bandwidth Above	2	4	7	19

Estimate is for the intercept term for receiving a higher rating from a linear regression with the provided bandwidths. Slopes are permitted to vary on either side of the cutoff. Sample is limited to schools that received an L, A, R, or E rating. Marginality below the rating cutoff is defined as affecting only one subject-group. Schools that receive ratings on appeal, were paired with another school, were identified as having undergone "special analysis" are excluded, as are schools with fewer than 200 students. All expenditures are in 2007 dollars. Robust standard errors clustered by school in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 8: Regression Discontinuity: The Effect of Rating on Per-Student Total Operating Expenditures in Following Year - Sensitivity to Bandwidth

Year t Rating ↓	A. TAAS			B. TAKS		
	CV - Optimal Bandwidth Minus 1	CV - Optimal Bandwidth	CV - Optimal Bandwidth Plus 1	CV - Optimal Bandwidth Minus 1	CV - Optimal Bandwidth	CV - Optimal Bandwidth Plus 1
	(1)	(2)	(3)	(4)	(5)	(6)
Low / Acceptable	736.9** (348.0)	496.1 (308.7)	540.1* (279.1)	-116.9 (208.8)	-117.1 (204.0)	-117.5 (204.0)
Obs	511	779	1099	2391	3086	3854
Bandwidth Below	3	4	5	16	17	18
Bandwidth Above	3	4	5	4	5	6
Acceptable / Recognized	84.6 (57.8)	108.2** (55.2)	112.1** (52.8)	-24.3 (51.6)	-31.7 (50.7)	-30.7 (49.4)
Obs	7756	8731	9409	5803	6626	7407
Bandwidth Below	8	9	10	16	17	18
Bandwidth Above	4	5	6	3	4	5
Recognized/Exemplary	113.7 (77.9)	113.7 (77.9)	94.7 (74.4)	-161.3 (240.3)	105.8 (200.4)	59.4 (156.8)
Obs	2658	3777	4653	1608	1796	1905
Bandwidth Below	18	19	20	2	3	4
Bandwidth Above	1	2	3	6	7	8

Columns (2) and (5) are repeated from Table 7. Estimate is for the intercept term for receiving a higher rating from a linear regression with the provided bandwidths. Slopes are permitted to vary on either side of the cutoff. Sample is limited to schools that received an L, A, R, or E rating. Marginality below the rating cutoff is defined as affecting only one subject-group. Schools that receive ratings on appeal, were paired with another school, were identified as having undergone "special analysis" are excluded, as are schools with fewer than 200 students. All expenditures are in 2007 dollars. Robust standard errors clustered by school in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 9: Estimates of Relationship Between Prior Per-Student Total Operating Expenditures and Discontinuity

Year t Rating ↓	A. TAAS				B. TAKS			
	Year $t + 1$ (Baseline) (1)	Year t (2)	Year $t - 1$ (3)	2 Years (t & $t+1$) (4)	Year $t + 1$ (Baseline) (5)	Year t (6)	Year $t - 1$ (7)	2 Years (t & $t+1$) (8)
Low / Acceptable	496.1 (308.7)	520.1* (299.3)	98.4 (162.3)	1016.2* (594.9)	-117.1 (204.0)	-73.7 (181.9)	-127.9 (88.4)	-190.7 (373.4)
Obs	779	779	633	779	3086	3086	2574	3086
Bandwidth Below	4	4	4	4	17	17	17	17
Bandwidth Above	4	4	4	4	5	5	5	5
Acceptable / Recognized	108.2** (55.2)	83.4 (53.7)	-6.9 (29.9)	191.6* (104.9)	-31.7 (50.7)	-80.2* (48.6)	-28.9 (37.0)	-111.9 (94.4)
Obs	8731	8731	7050	8731	6626	6658	4788	6626
Bandwidth Below	9	9	9	9	17	17	17	17
Bandwidth Above	5	5	5	5	4	4	4	4
Recognized/Exemplary	113.7 (77.9)	115.3 (71.8)	-3.8 (33.6)	229.1 (145.6)	105.8 (200.4)	83.2 (184.2)	4.2 (68.6)	188.9 (379.3)
Obs	3777	3777	3144	3777	1796	1796	1331	1796
Bandwidth Below	19	19	19	19	3	3	3	3
Bandwidth Above	2	2	2	2	7	7	7	7

Columns (1) and (5) are repeated from Table 7. Estimate is for the intercept term for receiving a higher rating from a linear regression with the provided bandwidths. Slopes are permitted to vary on either side of the cutoff. Sample is limited to schools that received an L, A, R, or E rating. Marginality below the rating cutoff is defined as affecting only one subject-group. Schools that receive ratings on appeal, were paired with another school, were identified as having undergone "special analysis" are excluded, as are schools with fewer than 200 students. All expenditures are in 2007 dollars. Robust standard errors clustered by school in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 10: Regression Discontinuity Results for Effect of Rating on Expenditures and Staffing

Outcome in Year $t + 1 \rightarrow$ Year t Rating \downarrow	Instruction (1)	Admin & Training (2)	Counseling (3)	Extra- Curricular (4)	Student- Teacher Ratio (5)
A. TAAS					
Low / Acceptable	413.3* (238.5)	34.8 (34.7)	-9.9 (16.3)	114.6*** (24.7)	-0.16 (0.46)
Obs	336	1503	802	1114	791
Bandwidth Below	3	12	14	11	7
Bandwidth Above	2	6	4	5	4
Acceptable / Recognized	47.8 (35.9)	26.7*** (7.5)	9.6* (5.2)	-1.4 (10.3)	0.02 (0.11)
Obs	7797	6497	7598	8776	6481
Bandwidth Below	10	12	5	20	10
Bandwidth Above	4	3	4	5	3
Recognized/Exemplary	29.2 (49.6)	39.5*** (13.4)	17.8*** (5.2)	2.5 (17.6)	0.17 (0.14)
Obs	3777	6403	5727	4586	3747
Bandwidth Below	20	5	7	5	7
Bandwidth Above	2	8	5	3	2
B. TAKS					
Low / Acceptable	-182.8 (136.7)	-12.8 (25.3)	29.3 (21.6)	-125.5* (71.1)	0.10 (0.40)
Obs	2392	3086	1278	4667	2380
Bandwidth Below	17	13	7	7	11
Bandwidth Above	4	2	2	7	4
Acceptable / Recognized	-73.0** (33.6)	6.0 (7.6)	21.1*** (7.7)	-10.7 (10.7)	0.42*** (0.09)
Obs	6626	6586	5768	6618	6623
Bandwidth Below	17	4	3	16	17
Bandwidth Above	4	5	4	4	4
Recognized/Exemplary	42.0 (138.1)	-16.2 (31.6)	9.3 (13.5)	11.2* (6.8)	0.18 (0.29)
Obs	1856	1747	1678	1970	1725
Bandwidth Below	3	2	3	7	2
Bandwidth Above	9	2	5	7	9

Estimate is for the intercept term for receiving a higher rating from a linear regression with the provided bandwidths. Slopes are permitted to vary on either side of the cutoff. Sample is limited to schools that received an L, A, R, or E rating. Marginality below the rating cutoff is defined as affecting only one subject-group. Schools that receive ratings on appeal, were paired with another school, were identified as having undergone "special analysis," or have fewer than 200 students are excluded. All expenditures are in 2007 dollars. Robust standard errors clustered by school in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 1: Requirements for TAAS Accountabilty Ratings

Subject	Math, Reading	Writing	Social Studies	Drop-Outs	Attendance
Grades	3 - 8, 10	4, 8, 10	8	7 - 12	All
Groups	White, Black, Hisp, Econ Dis, All	White, Black, Hisp, Econ Dis, All	All only	White, Black, Hisp, Econ Dis, All	White, Black, Hisp, Econ Dis, All
A. Acceptable					
1998-99	40% or RI	40% or RI	n/a	6% or RI	94%
1999-00	45%	45%	n/a	6%	94%
2000-01	50%	50%	n/a	5.50%	94%
2001-02	50%	50%	n/a	5%	-
2002-03	55%	55%	50%	5%	-
B. Recognized					
1998-99	80%	80%	n/a	3.50%	94%
1999-00	80%	80%; 3-8, 10	n/a	3.50%	94%
2000-01	80%	80%; 3-8, 10	n/a	3.00%	94%
2001-02	80%	80%; 3-8, 10	n/a	2.50%	-
2002-03	80%	80%; 3-8, 10	80%	2.50%	-
C. Exemplary					
1998-99	90%	90%	n/a	1%	94%
1999-00	90%	90%; 3-8, 10	n/a	1%	94%
2000-01	90%	90%; 3-8, 10	n/a	1%	94%
2001-02	90%	90%; 3-8, 10	n/a	1%	-
2002-03	90%	90%; 3-8, 10	90%	1%	-

RI - Required improvement. Schools that do not meet the requirement could get the higher rating by showing sufficient increase in the performance measure.

To count, a subject/student group combination must be at least either: 30 students & 10% of the student body, or 200 students (prior to 2001) / 50 students (2001 and later).

Appendix Table 2: Requirements for TAKS Accountability Ratings

Subject	Math	Reading/ELA	Writing	Social Studies	Science	SDAA	Drop-Outs	Completions
Grades	3 - 11	3 - 11	4, 8, 10	8, 10, 11	5, 10, 11	3 - 11	7 - 8	12
Groups	White, Black, Hisp, Econ Dis, All	White, Black, Hisp, Econ Dis, All	White, Black, Hisp, Econ Dis, All	White, Black, Hisp, Econ Dis, All	White, Black, Hisp, Econ Dis, All	All	White, Black, Hisp, Econ Dis, All	White, Black, Hisp, Econ Dis, All
A. Acceptable								
2003-04	35% or RI	50% or RI	50% or RI	50% or RI	25% or RI	50% or RI	1% or RI	75% or RI
2004-05 [†]	35% or RI	50% or RI	50% or RI	50% or RI	25% or RI	50% (no RI - new exam)	1% or RI	75% or RI
2005-06	40% or RI	60% or RI	60% or RI	60% or RI	35% or RI	50% or RI	1% (no RI - new calc method)	75% or RI
2006-07	45% or RI	65% or RI	65% or RI	65% or RI	40% or RI	50% or RI	2% or RI	75% or RI
B. Recognized								
2003-04	70% or RI	70% or RI	70% or RI	70% or RI	70% or RI	70% or RI	0.7% or RI	85%
2004-05	70% or RI	70% or RI	70% or RI	70% or RI	70% or RI	70% (no RI - new exam)	0.7% or RI	85% or RI
2005-06	70% or RI	70% or RI	70% or RI	70% or RI	70% or RI	70% or RI	0.7% (no RI - new calc method)	85% or RI
2006-07	75% or RI	75% or RI	75% or RI	75% or RI	75% or RI	70% or RI	0.7% or RI	85% or RI
C. Exemplary								
2003-04	90%	90%	90%	90%	90%	90%	0.2%	95%
2004-05	90%	90%	90%	90%	90%	90%	0.2%	95%
2005-06	90%	90%	90%	90%	90%	90%	0.2%	95%
2006-07	90%	90%	90%	90%	90%	90%	0.2%	95%

[†] Although the passing rates did not increase in 2004-05 the standard for whether individual students passed did increase.

RI - Required improvement. Schools that do not meet the requirement could get the higher rating by showing sufficient increase in the performance measure.

SDAA - State Developed Alternative Assessment - Test for certain special education students. Passing rates based on percent of tests taken.

ELA - English Language Arts

To count, a subject/student group combination must be at least either: 30 students, or 10% of the student body, or 50 students. Schools are also granted exceptions for for a certain number of low-scoring subject/group combinations based on the total number of subject/groups that count towards the rating. Exceptions can only increase a rating from L to A.