Economics of Education Review 49 (2015) 55–68

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Do administrators respond to their accountability ratings? The response of school budgets to accountability grades

Steven G. Craig[a], Scott A. Imberman[b,*], Adam Perdue[a]

[a] Department of Economics, University of Houston, Houston, TX 77204-5019, United States
[b] Department of Economics, Michigan State University, 486 W. Circle Drive, 110 Marshall-Adams Hall, East Lansing, MI 48824, United States

ARTICLE INFO

Article history:
Received 30 July 2014
Revised 12 May 2015
Accepted 20 July 2015
Available online 10 August 2015

JEL classification:
I22
H72
I28

Keywords:
Accountability
School finance

ABSTRACT

This paper examines how school administrators reallocate resources to schools in response to marginal changes in accountability ratings. We study this through an analysis of budgetary changes for schools on the margin of distinct rating boundaries. By determining how close each school is to an accountability grade change we are able to conduct a regression discontinuity analysis on schools that are on either side of the sharp line that separates school ratings. If administrators care about accountability ratings on the margin we would expect to see changes in budgetary allocations that reward higher performing, or punish lower performing, schools. Using data in Texas from 1994 to 2002, we find evidence suggesting that schools with higher ratings received more funds than others, and the differential funds were targeted toward administration/training, counseling and extra-curricular activities.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Accountability systems have been a rapidly growing element of the US public education system since the late 1990's. These systems generally evaluate schools based upon student performance on statewide standardized tests, and assign simple ratings based on the aggregate test score results of all tested students and students in certain sub-groups. The ratings are designed to be informative to parents and state legislators, and one objective of school accountability ratings appears to be to direct pressure from these two groups onto school and district 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.[1] If parents and/or legislators use the ratings for any of their decisions on school choice, 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 from other causes of budgetary choices. Our work here develops a regression-discontinuity framework of schools on the margin between one rating and the next, and analyzes whether the budgetary response of school districts depends on whether a school lies on one side of the rating boundary or the other.

There are two ways to think about how accountability grades might impact resource allocation by school districts when they are making resource allocation decisions between schools. On the one hand, the district might only be worried about "good" versus "bad" schools, and thus base resource allocations based on the long-term impression from

* Corresponding author. Tel.: 517 355 4667.
E-mail address: imberman@msu.edu (S.A. Imberman).

[1] Dee and Jacob (2011) and Neal and Schanzenbach (2010) find evidence that the Federal No Child Left Behind (NCLB) law increases achievement. Chiang (2009), 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. Rouse et al. (2013) also show schools change behavior in response to accountability ratings. To our knowledge, only Craig, Imberman, and Perdue (2013) and Chiang (2009) check budgetary responses to accountability ratings.

http://dx.doi.org/10.1016/j.econedurev.2015.07.005
0272-7757/© 2015 Elsevier Ltd. All rights reserved.

accountability ratings. This reasoning might imply that there are no marginal decisions; e.g., resources are directed toward schools based on their long-term performance. On the other hand, given the substantial resources that state governments and school districts invest in administering as well as assessing school performance on an annual basis, it would make sense that district decisions would still respond to short-term incentives. In this instance, especially if pressure to improve in the short-term is applied from the state government, districts may incentivize schools by providing rewards to highly rated schools and punishments to underperforming schools, or alternatively districts may attempt to bolster schools that under-perform.

In addition to an examination of whether school districts allocate funds based on a school's accountability grade, we examine the within-school allocation of resources. This aspect of budget allocations may be the result of school district behavior, or may result from choices made by the school's principal and other decision makers. That is, schools which barely succeed, or which marginally fail, to obtain the next higher accountability grade may reallocate resources within the school. This reallocation may serve to increase the chances of surpassing the threshold in the next year, or to reward employees and students for performance in this period. Examples would be that schools that find they fall just short of the next grade might reallocate resources toward instruction, or schools which barely are able to achieve the next grade may "reward" students with more funding for extracurricular activities.

To test the impact that annual accountability ratings have on school district as well as individual school financial allocations, we utilize a regression discontinuity (RD) design to compare the budgetary response to annual changes in rating for schools marginally on either side of each rating boundary. The sharp discontinuity occurs because school grades are based on the percentage of students that pass the accountability exam—if the school misses the cutoff by just one student it receives a lower accountability 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). To implement the RD strategy, therefore, we carefully re-create the scoring matrix and identify schools where the rating is marginal based mainly on exam performance. 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 independent of any change in underlying school quality.

Our analysis here offers a different strategy for finding allocation changes compared to Craig, Imberman, and Perdue (2013), who use a "rating shock" strategy based on the change in the rating system in Texas. They find that school districts reallocate funds to schools which were threatened with a drop in their accountability grade, but that the incremental resources were temporary and generally disappeared after 3 years. This paper differs in that, while Craig, Imberman and Perdue (2013) consider responses to a potential long-term change in ratings, we investigate whether administrators respond to the annual changes in school ratings.

Whether and how administrators respond to such marginal changes in school performance is important for a few reasons. First, such an analysis provides insight into the objective function of school administrators which is poorly understood. While Craig, Imberman and Perdue (2013)'s findings suggest that administrators care about long-term school quality, administrators may also worry about the reputational consequences and sanctions—both explicit (e.g., punishments imposed by accountability systems) and implicit (e.g., loss of enrollment)—from published rating changes due to marginal differences in underlying factors. Second, if administrators do provide either inducements or punishments for changes in ratings, such behavior could increase incentives to game accountability systems or "teach to the test." Third, in this study we look at the impacts of accountability ratings on school finances under a routine setting, the impacts of which could differ substantially from cases such as in Craig, Imberman and Perdue (2013) where unique and non-repeatable settings are used. Fourth, while we cannot fully separate the behaviors of district administrators from principals (with the exception of total funding which is entirely under the purview of the district) we can nonetheless gain some insight into how principals respond when their accountability pressure is relaxed (increased) by getting a higher (lower) rating in accountability systems by estimating impacts on changes in categorical expenditures. Our knowledge of such principal behaviors is rather thin as it is difficult to separate principals' efforts from teachers'. Rouse et al. (2013), for example, find evidence that getting a failing rating leads to more teacher resources via a survey, they are not able to assess the impacts on specific spending categories, overall school funding, or the impacts of getting a high rating. Thus, our study complements Rouse et al. (2013) and Craig, Imberman and Perdue (2013) by providing some needed insight into how principals respond to accountability pressures.

For our analysis we focus on the accountability system in place in Texas from 1994 through 2002 called the Texas Assessment of Academic Skills (TAAS). Under this system, schools were given ratings based on student performance on test scores and, to a lesser extent, attendance, dropout and graduation rates.[2] While the system has since been replaced, first with the Texas Assessment of Knowledge and Skills (TAKS) until 2012 and then the State of Texas Assessments of Academic Readiness (STAAR) afterward, we only analyze the TAAS system here. Our preliminary analyses showed discontinuities in the densities of the forcing variable under the TAKS regime. Under that regime schools were able to acquire temporary "exceptions" that allowed them to negate falling below ratings cutoffs for some subgroups. This led to bunching above the thresholds leaving us unable to determine how much of the bunching was due to the structure of the system and how much was due to manipulation that would negate the validity of the regression discontinuity design.[3] Fortunately, we find little evidence of similar problems during the TAAS regime. For the STAAR system, the implementation is too recent to conduct a full analysis and thus we leave that to future research. Hence, under the TAAS system, we find evidence that school districts provided small budgetary

[2] The attendance requirements were abandoned in 1999–2000.

[3] For the estimates using data from the TAKS period see Craig, Imberman and Perdue (2009).

Table 1
Distributions of accountability ratings by year.

	Accountability rating				Rating changes		
	Low	Acceptable	Recognized	Exemplary	% Lower	% Same	% Higher
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%

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, and identified as having undergone "special analysis," or have fewer than 200 students are excluded (see text).

rewards for schools that were able to achieve a higher accountability grade, most of which went to increasing administrative/training, counseling and extra-curricular expenses.

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 provided by ratings over and above the measured learning output of schools. One caution in their results, however, is that they find the housing market response seems to decline over time, possibly suggesting that the use of accountability grades falls over time as residents learn 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 responses found by those examining within school behavior. Examples of how schools internally respond to accountability systems include Carnoy and Loeb (2003), Chakrabarti (2013b, 2008), Hanushek and Raymond (2004, 2005); Jacob (2005), and Rouse et al. (2013). Researchers have also found, however, that some gains may be due to schools "gaming" the system (Figlio and Getzler, 2006; Figlio and Winicki, 2005; Jacob, 2005; Cullen and Reback, 2006; and Figlio, 2006) or focusing on marginal students (Chakrabarti, 2013a; Reback, 2008; and Neal and Schanzenbach, 2010). Nonetheless, this range of results suggests that, at a minimum, 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. 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. Rouse et al. (2013) 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 (2012) find that schools that receive rewards for higher ratings generally put the money into teacher bonuses. None of these papers, however, looks at resource allocation responses directly as a result of ratings. Other than Craig, Imberman and Perdue (2013) the only paper that considers this question, to our knowledge, 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. Specifically, we examine the full range of schools from elementary through high school, we analyze a span of 6 years, and we also assess the impacts of receiving marginally higher ratings above the lowest rating—those that distinguish between whether schools are adequate or exceptional.

2. The Texas Assessment of Academic Skills

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). Ratings are determined by whether schools exceed 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.[4] Table 1 shows the distribution of ratings by year and how often ratings change. The table shows that very few schools received the lowest ratings. Initially most schools received the second lowest rating "A," but over time the distribution shifted so that by 2001 most schools received one of the two highest ratings. The right-hand-side of the table shows that changes in ratings were common with a little fewer than 40% of schools moving up or down each year.

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

[4] 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.

Table 2
Rating transitions.

Year $t-1$	Year t 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%

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, and identified as having undergone "special analysis," or have fewer than 200 students are excluded (see text).

could also affect the rating.[5] Thus the system is based on test score levels rather than student gains, and the rating for the entire school is determined by a Rawlsian metric, as the school's rating is equal to the rating of the lowest performing subject-group of sufficient size. Table A1 provides a description of the requirements for achieving each rating in the years of our analysis.[6] With the exception of receiving a low rating, there are no direct punishments imposed on schools by the state, and state awards for high performance were extremely small.[7] 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. 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 annual movement of schools between ratings certainly suggests that, if they choose to do so, districts would be able to use the ratings to evaluate, and therefore to reward or punish, schools depending on their performance.

3. Empirical methodology

Our examination of school district behavior is through the total operating expenses in each school, which reflects the school district's financial allocation strategy.[8] 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 research strategy is to 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. This procedure will yield a local average treatment effect of the impact of the rating of schools on each rating margin by comparing schools that otherwise are presumably equal, and will detect whether the accountability grades are valued independently of other indicators of school performance. School districts will be found to reallocate 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 other indicators of school performance, and/or 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. Such beliefs would indicate that annual accountability ratings may not be providing new information about school performance.

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.[9] To implement this strategy, we calculate the accountability rating for each subject-group cell that meets minimum size requirements from the accountability system (this is generally 30 students). 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 by one rating category. 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.[10] The measure of N_b is therefore:

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

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_s is the school's actual accountability

[5] The attendance requirements were abandoned in 1999–2000.

[6] 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/.

[7] 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 2 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.

[8] The district's allocation decision can be a combination of formulas, explicit exceptions, or an administrative decision process.

[9] 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.

[10] 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 Fig. 1 shows that there were virtually no schools that were incorrectly assigned a different rating than they actually received under TAAS.

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 toward 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.[11] 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 falls for any cell that is equal to the school's overall rating, 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 let fail before their rating falls. As such, the second forcing variable, N_a, can be calculated as:

$$N_a = \min_{s,g} \left(1\left(\text{Rating}_{sg} = \text{Rating}_s\right) \times 1\left(\text{Size}_{sg} \geq \text{Min}_{sg}\right) \times M_{sg}\right) \quad (2)$$

where the definitions are the same as in (2), and where M_{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.[12] Hence if a school is precisely at the boundary while receiving the higher rating both N_a and N_b would equal zero.

We conduct our regression-discontinuity analysis of schools where N_a or N_b is close to zero for each boundary. Specifically, we conduct local-linear regressions of the form:

$$R_{i,t+1} = \alpha + \beta_1 N_{b,it} + \beta_2 N_{a,it} + \beta_3 \text{Above}_{it} + \varepsilon_{it} \quad (3)$$

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 are as defined by Eqs. (1) and (2), and *Above* is a dummy variable indicating whether the school is above the rating threshold. We select bandwidths for the RD through leave-one-out cross validation.[13] As suggested in Lee and Lemieux (2009), we use a rectangular kernel that involves limiting the sample to narrow bands around the cutoffs without re-weighting the data.[14] Fig. 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.[15]

It is possible that schools, or districts, do not equally value the three boundaries that exist in the accountability system. Thus, to explore whether the reactions by schools and districts may differ across the separate boundaries, we run (3) 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. For example, special attention is paid even in the legislation to "failing" schools, indicated by the L rating. Conversely, it is possible that ratings at the top (E) also garner special attention. The RD methodology will allow us to assess whether the accountability system provides new information to administrators in charge of allocating resources, who may then use the accountability system information to alter resource allocations indicated by the left hand side variable *R*.

4. Data

Our data covers all of the 5,803 public schools in the state of Texas, and comes from three datasets provided by the Texas Education Agency (TEA). First is the Public Education Information Management System (PEIMS), which provides financial reports on expenditures and resources for each school in Texas by category of expenditure.[16] We examine total expenditures plus; instruction, administration and training, counseling, and extra-curricular expenditures. The PEIMS data also contains the number of full-time equivalent faculty in each school. We use this data to test using the number of teachers separately from their cost, which can vary based on training and experience. We convert all expenditure levels to per student amounts, and further use the CPI to convert all dollar amounts to real 2007 dollars. The student data are contained in the Academic Excellence Indicator System (AEIS), which provides data on enrollment and student demographics for each school. The third data source is the TEA accountability reports that provide the data used to calculate the accountability ratings, including the performance in each of the subject-group cells which allow us to carefully distinguish schools on the margin of a rating boundary.[17]

[11] Since the overall school rating is based on the lowest rated subject-group, no group that counts toward the rating would have a rating lower than the school-wide rating.

[12] 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.

[13] 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.

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

[15] Given these results, we rely on a "strict" RD design. Nonetheless estimates using a "fuzzy" design were nearly identical. There is an appeal process for a school's rating, and we believe this accounts for the few exceptions across the rating boundary.

[16] We use the actual expenditures by schools rather than budgeted (planned) expenditures.

[17] All data sources are publically available on the website for the TEA.

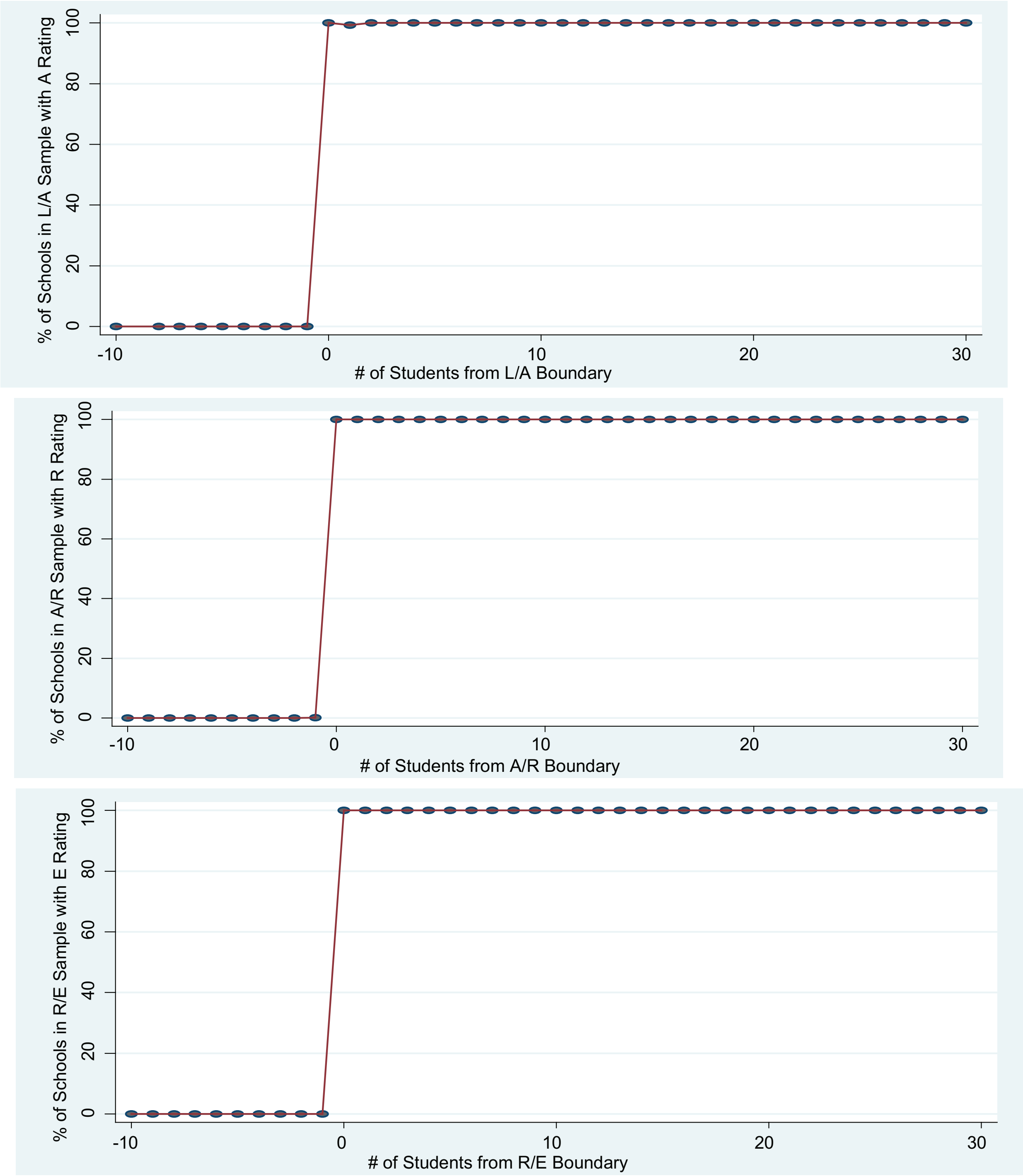


Fig. 1. Changes in accountability rating at rating boundaries. Sample is restricted to schools one rating above boundary or one rating below boundary where they miss the higher rating in only one subject-group. Schools that are paired with another school, received a rating on appeal, were identified as receiving "special analysis", or had fewer than 200 students are also excluded.

The RD regressions are pooled across the years 1997 through 2002, though we separately estimate each accountability boundary. 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.[18] After these restrictions, we have 26,500

[18] Alternative schools targeted to specific groups of students operate under a separate accountability system. These schools along with charter schools

Table 3
School characteristics by rating.

	Low	Acceptable	Recognized	Exemplary
% Asian	1.1	1.8	2.0	3.2
	(2.3)	(3.4)	(4.1)	(5.3)
% Black	29.5	18.6	11.1	7.1
	(26.8)	(22.6)	(16.6)	(11.7)
% Hispanic	48.4	43.8	39.4	24.5
	(30.1)	(31.7)	(32.3)	(28.6)
% White	20.8	35.6	47.1	64.9
	(23.6)	(29.0)	(31.3)	(29.1)
% Economically	70.9	59.9	51.0	31.9
Disadvantaged	(22.5)	(25.0)	(26.4)	(28.0)
% LEP	23.5	16.6	13.3	8.0
	(23.1)	(19.4)	(17.7)	(14.3)
%Special Ed	18.3	14.1	12.0	13.6
	(26.4)	(23.7)	(23.8)	(26.6)
% Gifted	7.7	7.6	7.5	9.5
	(6.3)	(6.7)	(6.6)	(9.5)
Enrollment	812	749	633	601
	(503)	(502)	(410)	(383)
Observations	285	12,111	8,614	5550

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, and identified as having undergone "special analysis," or have fewer than 200 students are excluded (see text).

Table 4
Mean school resources by rating.

Resources in Year $t + 1$	Low	Acceptable	Recognized	Exemplary
Total operating exp.	6397	5956	5980	5969
	(2011)	(1238)	(1257)	(1505)
Instructional exp.	4612	4376	4461	4470
	(1699)	(843)	(832)	(985)
Admin and training exp.	679	585	562	550
	(308)	(210)	(240)	(340)
Counseling exp.	249	230	216	210
	(127)	(106)	(98)	(102)
Extra-curricular exp.	115	117	128	146
	(191)	(210)	(246)	(282)
Student–Teacher ratio	15.2	14.9	14.7	14.7
	(2.6)	(2.5)	(2.5)	(2.5)

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, and identified as having undergone "special analysis," or have fewer than 200 students are excluded (see text). All expenditures are in 2007 dollars.

school-year observations for TAAS, consisting of over 5800 schools in 829 school districts in Texas.

Table 3 presents the means of school characteristic data. In general, schools with higher ratings contain fewer students on free and reduced price lunch, have fewer minorities, have fewer total students, and have fewer special needs students. Table 4 provides summary statistics for resources in the year after a school receives a rating. Schools with lower ratings have higher per student expenditures than those with higher ratings. These differences generally hold across expenditure categories.

5. Empirical results

Fig. 2 graphically shows the discontinuity in total expenditures as school test score results get closer to each of the three rating boundaries—the lowest rankings boundary L/A, the middle rankings boundary A/R, and the highest rankings boundary, R/E. The figure illustrates how much schools spend per pupil in academic year $t + 1$ when the accountability rating is released late in the spring of academic year t using a fitted local polynomial. It shows clear jumps in expenditures as schools pass from L into A ratings and from A into R ratings, though we acknowledge that the small sample size of schools around the L/A boundary generates a substantial amount of noise around the cutoff. The results for the R/E boundary are consistent with the other boundaries but smaller. Table 5 presents the regression coefficients for models analogous to those shown in Fig. 2. In column (2), we provide estimates that use the cross-validation optimal bandwidths along with estimates that include smaller bandwidths (column 1) and larger bandwidths (columns 3 and 4) to test sensitivity. While the estimates for the L/A border vary by bandwidth choice given the small sample sizes they nonetheless show a qualitatively consistent and positive effect. The other boundaries show more consistent positive effects, though only A/R is statistically significant at the optimal bandwidth.

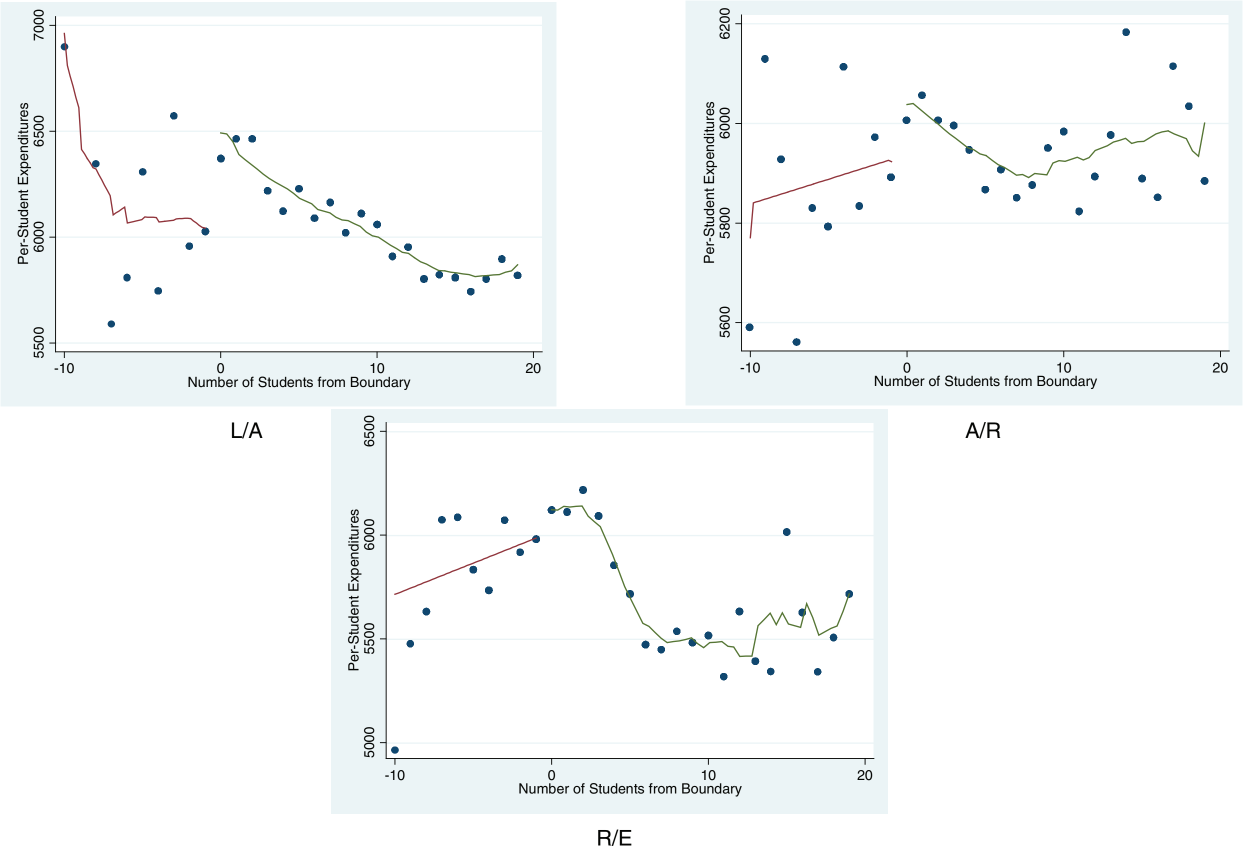
As a result, while the pattern is positive, for the L/A and R/E boundaries we cannot rule out a zero effect, though taking the L/A estimate at the optimal bandwidth at face value this amounts to an increase of 8% of mean expenditure. We see significant impacts for the A/R boundary of $108 per student, approximately 1.5% of mean total expenditure. Thus, despite the relatively small size and marginal significance, all three boundaries show consistent results that indicate some small positive rewards were given to schools that exceeded cutoffs, or conversely some small punishments were imposed on schools that just miss the cutoffs.[19]

Table 6 explores the issue of timing. The accountability results are made 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 in the initial specification.[20] Thus, in Table 6 we show results using alternative timing assumptions. We find that the results are similar to the base case in Table 5. That is, expenditures in year t are found to rise by about the same amount as in year $t + 1$ which we reported in Table 5, albeit with slightly higher standard errors. This implies that either some schools provide the bonuses in year t and some in $t + 1$, or that the schools provide the funds for at least 2 years on average.

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.

[19] One potential concern is that these changes are responses to changing enrollment patterns rather than awards/punishments. To address this we also estimate models, provided in the online appendix, that control for a quartic in enrollment in the year the funds are spent. The results change little in this model and, in fact, smaller standard errors provide consistently significant and positive impacts across bandwidth choices.

[20] 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.



L/A

A/R

R/E

Fig. 2. Total per-student expenditures in TAAS. 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 5
Regression discontinuity: the effect of rating on per-student total operating expenditures in following year ($t + 1$) sensitivity to bandwidth.

Year t rating ↓	CV - optimal bandwidth minus 1 (1)	CV - optimal bandwidth (2)	CV - optimal bandwidth plus 1 (3)	CV - optimal bandwidth plus 3 (4)
Low/acceptable (L/A)	736.9**	496.1	540.1*	355.9
	(348.0)	(308.7)	(279.1)	(224.3)
Obs	511	779	1099	2008
Bandwidth below	3	4	5	7
Bandwidth above	3	4	5	7
Acceptable/recognized (A/R)	84.6	108.2**	112.1**	120.4**
	(57.8)	(55.2)	(52.8)	(50.8)
Obs	7756	8731	9409	10116
Bandwidth below	8	9	10	12
Bandwidth above	4	5	6	8
Recognized/exemplary (R/E)	113.7	113.7	94.7	165.1**
	(77.9)	(77.9)	(74.4)	(68.90)
Obs	2658	3777	4653	5758
Bandwidth below	18	19	20	22
Bandwidth above	1	2	3	5

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 (see text). 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 6
RD estimates of impact of rating on total expenditures by time since rating.

Year *t* rating ↓	Year *t* + 1 (Baseline) (1)	Year *t* (2)	2 Years (*t*, *t* + 1) (3)	Year *t* – 1 (4)
Low/acceptable (L/A)	496.1	520.1*	1016.2*	98.4
	(308.7)	(299.3)	(594.9)	(162.3)
Obs	779	779	779	633
Bandwidth below	4	4	4	4
Bandwidth above	4	4	4	4
Acceptable/recognized (A/R)	108.2**	83.4	191.6*	−6.9
	(55.2)	(53.7)	(104.9)	(29.9)
Obs	8731	8731	8731	7050
Bandwidth below	9	9	9	9
Bandwidth above	5	5	5	5
Recognized/exemplary (R/E)	113.7	115.3	229.1	−3.8
	(77.9)	(71.8)	(145.6)	(33.6)
Obs	3777	3777	3777	3144
Bandwidth below	19	19	19	19
Bandwidth above	2	2	2	2

Estimate is for the intercept term for receiving a higher rating from a linear regression with the provided bandwidths. CV optimal bandwidths for total expenditures from Table 5 are used in all models. 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, and identified as having undergone "special analysis" are excluded, as are schools with fewer than 200 students (see text). All expenditures are in 2007 dollars. Estimates in column one are identical to column two in Table 5. Robust standard errors clustered by school in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 7
Tests of discontinuities in school characteristics.

Outcome in year *t* → Year *t* rating ↓	Enrollment (1)	Δ Enrollment (2)	% White (3)	% Black (4)	% Hispanic (5)	% Disadv (6)	%LEP (7)	% Special education (8)	% Gifted (9)
Low/Acceptable	53.7	−15.0	4.0	1.7	−5.6	−4.0	−10.4*	6.4	−11.4**
	(96.5)	(15.9)	(6.6)	(6.3)	(7.6)	(5.4)	(5.3)	(4.9)	(5.2)
Obs	786	562	786	786	786	786	786	786	786
Acceptable/recognized	19.0	−3.2	−1.9	0.4	1.5	0.6	0.7	0.7	0.8
	(20.0)	(3.9)	(1.2)	(0.7)	(1.2)	(1.0)	(0.6)	(1.0)	(0.6)
Obs	8768	6619	8768	8768	8768	8768	8768	8768	8768
Recognized/exemplary	71.6***	0.4	−1.2	0.7	0.6	−1.2	−0.2	2.7**	−0.2
	(21.0)	(6.3)	(1.5)	(0.7)	(1.4)	(1.3)	(0.7)	(1.3)	(0.6)
Obs	3795	2690	3795	3795	3795	3795	3795	3795	3795

Estimate is for the intercept term for receiving a higher rating from a linear regression with the provided bandwidths. CV optimal bandwidths for total expenditures from Table 5 are used in all models. 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, and identified as having undergone "special analysis," or have fewer than 200 students are excluded (see text). 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.

One potential concern is that perhaps the schools that exceed the cutoff happen to be schools that receive more money in general. As a test of this, in column 3 we show the impact of exceeding the cutoff with expenditures in the year prior to the rating receipt. If the schools that exceed the cutoff are larger spenders we would expect to see significant effects. In this case we do not, and in fact the estimates are close to zero in all three cases. We thus conclude that the finding of larger total expenditures as a result of exceeding the rating boundary is likely to be a valid description of school district behavior.

To further validate the regression-discontinuity strategy, in Table 7 we provide estimates of reduced-form RD impacts on observable characteristics of schools in year *t* that should not be affected by the rating in year *t* since the rating is released after the observables are measured. The significance rates are a bit higher than would be expected to occur randomly—3 estimates are significant at the 5% level or higher, 4 at the 10% level or higher—but do not show a clear pattern and are spread across the boundaries. Further, there are no significant effects for the A/R boundary, which is the case where there were the clearest positive effects. Finally, in Table A2 we provide estimates that control for observables, and find qualitatively similar results to the baseline results for schools on the A/R and R/E boundaries, although we do not find that the L/A expenditure impacts are statistically robust to this test. Further, the lack of significant impacts from lagged expenditures shown in column (4) of Table 6 provides additional evidence for the validity of or RD design.

In Fig. 3 we show distributions of the density of schools around the cutoffs. As can be seen from the overlaid kernel density smoother, the histograms appear to track the

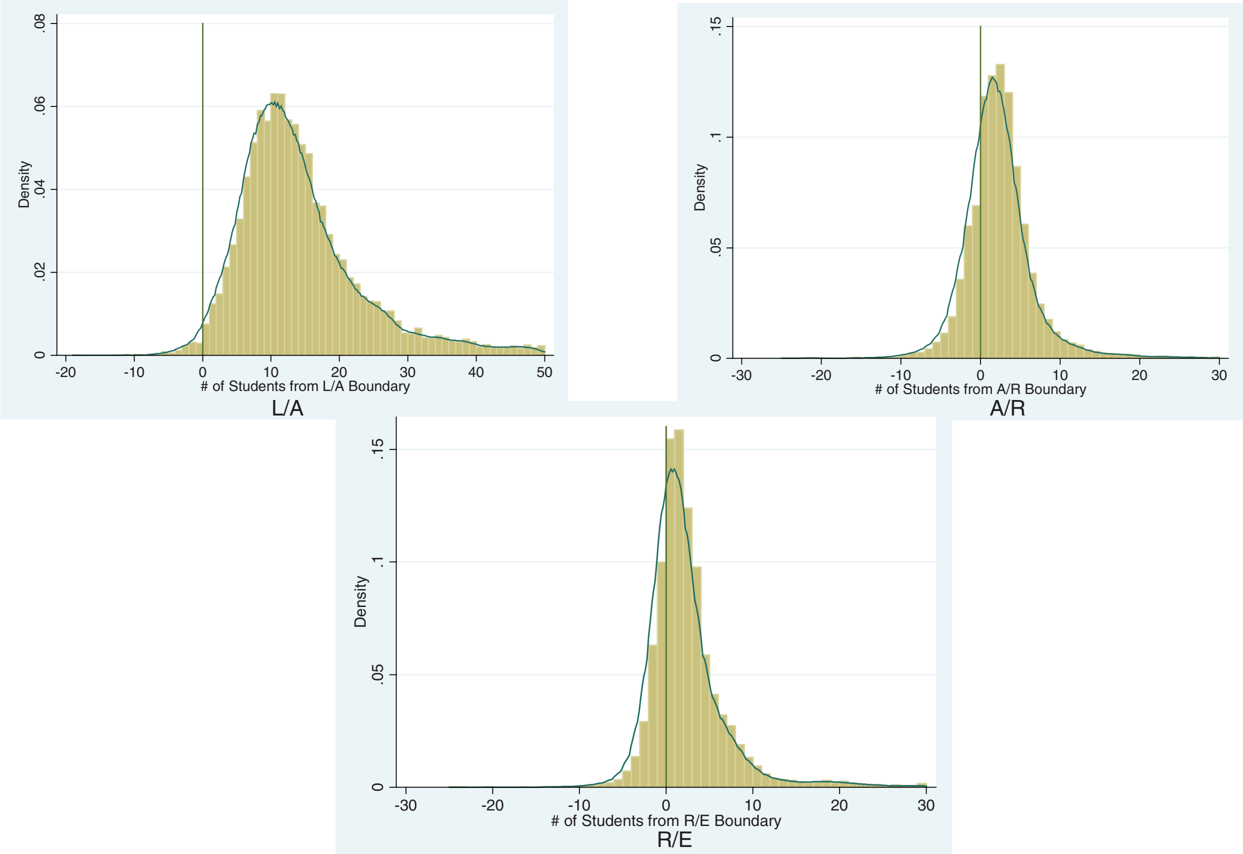


Fig. 3. Distributions of distances from ratings boundaries sample is restricted to schools one rating above boundary or one rating below boundary where they miss the higher rating in only one subject-group. Schools that are paired with another school, received a rating on appeal, and identified as receiving "special analysis", or had fewer than 200 students are also excluded.

smoothed density pretty closely. However, there is some indication of a potential for very localized—within one student of the cutoff—manipulation. To test whether this is a major concern, we estimate "donut-hole" models that drop observations at the cutoff and one student below the cutoff. These are provided in Table A2 and show little change in the quantitative estimates for the L/A and A/R boundaries, though standard errors increase enough to render them insignificant. The R/E estimate, on the other hand, drops essentially to zero. In general, the results in Tables 6 and 7 and Fig. 3 indicate that, while there remains some possibility of manipulation, the likelihood that this affects the estimates in a substantial way is small. Hence, we conclude from the sensitivity analysis that at least some school districts in Texas are likely to have used the TAAS accountability system to reward success at reaching certain ratings or punish districts that fail to meet the rating thresholds.

The Table 8 results show how various categories of expenditure respond to the budgetary changes from being just over the accountability ratings border.[21] We test for these categorical expenditure changes to examine how the impacts on total expenditures are split amongst various uses. The expenditure results are less clearly the behavior of the school districts, as presumably school officials, and especially 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 higher total expenditures (83%) for schools just over the LA border is spent on instruction, with most of the remainder on extra-curricular activities. In contrast, schools on the AR and RE border are found to spend their much smaller extra resources on counseling, plus administration and training.[22] Intriguingly, while the L/A results are consistent with Chiang (2009) finding higher instructional spending, the higher cutoffs show expenditure differences that occur in ancillary categories. Given that he looks specifically at the lowest cutoff in Florida, our results suggest that on higher thresholds, where sanctions are smaller, administrators focus on more ancillary costs. Even so, the substantial effect on extra-curricular activities on the L/A boundary does indicate some improvements

[21] Results are similar when controlling for a quartic in current enrollment and are provided in the online appendix.

[22] It is interesting to speculate on the objective function indicated by the differences in these results between the LA and AR borders, but we do not have empirical evidence to suggest anything further than our findings here.

Table 8
Regression discontinuity results for effect of rating on expenditures and staffing.

Outcome in year $t+1$ →	Instruction	Admin and training	Counseling	Extra-curricular	Student–Teacher ratio
Year t rating ↓	(1)	(2)	(3)	(4)	(5)
Low/acceptable (L/A)	413.3*	34.8	−9.9	114.6***	−0.16
	(238.5)	(34.7)	(16.3)	(24.7)	(0.46)
Obs	336	1503	802	1114	791
Bandwidth below	3	12	14	11	7
Bandwidth above	2	6	4	5	4
Acceptable/recognized (A/R)	47.8	26.7***	9.6*	−1.4	0.02
	(35.9)	(7.5)	(5.2)	(10.3)	(0.11)
Obs	7797	6497	7598	8776	6481
Bandwidth below	10	12	5	20	10
Bandwidth above	4	3	4	5	3
Recognized/exemplary (R/E)	29.2	39.5***	17.8***	2.5	0.17
	(49.6)	(13.4)	(5.2)	(17.6)	(0.14)
Obs	3777	6403	5727	4586	3747
Bandwidth below	20	5	7	5	7
Bandwidth above	2	8	5	3	2

Estimate is for the intercept term for receiving a higher rating from a linear regression with the provided bandwidths. CV optimal bandwidths for each outcome are used in all models. 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, and identified as having undergone "special analysis," or have fewer than 200 students are excluded (see text). 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.

in ancillary funding for schools achieving A ratings as well, but this is dwarfed by the impact on instruction.[23]

In addition to those discussed, we provide a series of diagnostic sensitivity analysis of the above RD results in Table A2. We check whether the estimates change if we use the total expenditure bandwidth for expenditure categories, if we instrument for actual rating using whether the school exceeds the boundary (a "fuzzy" RD), expand the sample such that schools that fall below the boundary can have failed multiple subject-groups as long as the school only fails in one subject, or use a 3-term or 5-term polynomial instead of local-linear regression. While the estimates for the L/A boundary are not completely robust across specifications, the estimates for A/R and R/E remain consistent across all of these specifications.

6. Summary and conclusion

This paper has examined whether school accountability testing has influenced financial allocations of school districts. Specifically, if legislators believe that accountability ratings provide new information on school quality, we would expect that school districts would respond by using the ratings to assist in resource allocation. Our test of this idea uses a regression discontinuity design by carefully examining the first Rawlsian-style accountability system used in Texas, where we identify the exact scoring cell by subject and demographic characteristic. Our empirical work finds that this early accountability system (TAAS) was used to some degree to make financial decisions—schools just over the rating boundary received more resources compared to schools just under the boundary, although only on the boundary between being rated "Acceptable" or "Recognized" do we see results that are consistently statistically significant at conventional levels across our wide variety of sensitivity analyses. Further, while estimates on the lowest boundary between "Low" and "Acceptable" schools appear large, we caution that they are noisy due to small sample sizes and are not robust to adding control variables. We interpret these results as suggesting that educators and policy administrators believed that accountability rating could be used as new information about school quality, though we cannot determine whether the funding was provided as rewards for good performance or removed as sanctions for poor performance. Further, whether this result persists into the future has not yet been ascertained.

The other interesting distinction in our work is that we find that schools and districts tended to spend the higher funding from exceeding the cutoff in different ways. Our results suggest that schools on the lowest rating boundary that just barely achieve the Acceptable rating (A) use the incremental resources on what might be viewed as basic educational activities, specifically on instruction and extra-curricular activities. On the other hand, schools that are barely able to achieve the higher levels, either Recommended (R) or Excellent (E), are found to use the funding for what might be longer term or enrichment activities such as administration/development or counseling.

This work presents an interesting complement to Craig, Imberman and Perdue (2013) who find that a long-term *decrease* in the rating leads to higher spending in schools. This study shows that administrative responses to short-term marginal changes in ratings were quite different leading districts to reward schools that have an *increase* in ratings, or conversely punish schools with a *decrease* in ratings.

[23] It is possible that some of the differences are due to the different optimal bandwidths for each outcome. To address this in panel II of Table A2 we provide estimates that use the same bandwidths (the optimal bandwidths for total expenditures) for each outcome. While the instruction estimate for L/A shrinks by about 40% it remains positive and the rest of the results are similar to those provided in Table 8.

Combined, our studies suggest that school districts are willing to invest real resources to learn to adapt to a new accountability rating system, but that once a system is up and running at least the bulk of districts are interested in using the accountability system to provide marginal short term incentives to perform.

Acknowledgments

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 Steven Craig at scraig@uh.edu or Scott Imberman at imberman@msu.edu. All errors and opinions are our own. ©2015 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.

Appendix

Table A1
Requirements for TAAS accountability 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.
Notes: To count, all of the subject/student group combinations must be at least either: 30 students and 10% of the student body, or 200 students (prior to 2001)/50 students (2001 and later).

Table A2
R-D sensitivity tests.

Outcome in Year $t + 1$ →	Total (1)	Instruction (2)	Admin and training (3)	Counseling (4)	Extra-Curricular (5)	Student–Teacher ratio (6)
I. Drop Observations That Are One Below or One Above the Boundary						
Low/acceptable	587.8	1084.7	50.5	−4.0	123.9**	0.09
	(874.2)	(1078.2)	(80.7)	(29.0)	(58.0)	(0.76)
Obs	654	386	1378	677	989	666
Acceptable/recognized	97.1	68.3	42.9***	4.6	−14.9	0.14
	(88.0)	(62.2)	(16.6)	(9.5)	(14.5)	(0.18)
Obs	6676	5742	4442	5543	6721	4426
Recognized/exemplary	−1.1	−53.7	42.3**	7.8	1.1	0.03
	(139.8)	(94.1)	(19.7)	(10.7)	(38.9)	(0.32)
Obs	2858	2858	4608	3932	2791	2828
II. All Regressions Use Total Per-Student Expenditure Optimal Bandwidths						
Low/acceptable	496.1	266.9	16.4	−21.7	108.5***	−0.3
	(308.7)	(215.9)	(53.3)	(25.6)	(31.4)	(0.6)
Obs	779	779	779	779	779	775

(continued on next page)

Table A2 (*continued*)

Outcome in Year $t+1$ →	Total (1)	Instruction (2)	Admin and training (3)	Counseling (4)	Extra-Curricular (5)	Student–Teacher ratio (6)
Acceptable/recognized	108.2**	59.9*	27.2***	7.1*	0.3	0.0
	(55.2)	(36.0)	(7.4)	(4.2)	(11.3)	(0.1)
Obs	8731	8731	8731	8731	8731	8729
Recognized/exemplary	113.7	29.2	41.3**	21.3***	12.5	0.2**
	(77.9)	(49.6)	(20.9)	(5.1)	(15.1)	(0.1)
Obs	3777	3777	3777	3777	3777	3777
III. Use Actual Rating in Regressions and Instrument with Whether Above or Below Boundary ("Fuzzy" R-D)						
Low/acceptable	497.9	413.3*	34.9	−9.9	115.0***	−0.2
	(309.9)	(238.5)	(34.8)	(16.3)	(24.8)	(0.5)
Obs	779	336	1503	802	1114	791
Acceptable/ recognized	108.4**	47.9	26.7***	9.6*	−1.4	0.0
	(55.2)	(36.0)	(7.5)	(5.2)	(10.3)	(0.1)
Obs	8731	7797	6497	7598	8776	6481
Recognized/exemplary	113.7	29.2	39.5***	17.8***	2.5	0.2
	(77.9)	(49.6)	(13.4)	(5.2)	(17.6)	(0.1)
Obs	3777	3777	6403	5727	4586	3747
IV. For schools below boundaries limit to schools that fail one subject rather than those that fail one subject-group						
Low/acceptable	129.7	79.1	9.5	−8.4	82.6***	−0.7
	(282.0)	(217.6)	(31.3)	(14.1)	(29.4)	(0.4)
Obs	819	369	1583	886	1191	853
Acceptable/recognized	107.7**	60.5*	27.2***	10.9**	−8.0	0.0
	(48.6)	(32.3)	(6.9)	(4.7)	(8.4)	(0.1)
Obs	9704	8856	7740	8056	10394	7540
Recognized/exemplary	89.1	41.4	43.5***	14.6***	−11.9	0.1
	(62.9)	(39.2)	(12.6)	(4.6)	(16.0)	(0.1)
Obs	5227	5251	6984	6564	5167	4584
V. Use 3-term polynomial instead of local-linear regression (Range −10 to 10)						
Low/acceptable	338.5	188.2	14.8	−10.3	77.9***	−0.8**
	(238.5)	(190.6)	(34.0)	(14.7)	(26.6)	(0.3)
Obs	4430	4430	4430	4430	4430	4423
Acceptable/recognized	97.6**	65.7**	16.2***	3.2	−5.2	−0.1
	(42.1)	(27.6)	(6.1)	(3.3)	(8.4)	(0.1)
Obs	16101	16101	16101	16101	16101	16099
Recognized/exemplary	196.3***	104.6***	34.0***	11.6***	15.9	0.0
	(51.3)	(34.2)	(11.7)	(3.5)	(9.8)	(0.1)
Obs	9635	9635	9635	9635	9635	9635
VI. Use 5-term polynomial instead of local-linear regression (Range −10 to 10)						
Low/acceptable	35.9	90.8	−15.6	−42.7*	−3.0	0.0
	(261.8)	(178.0)	(49.2)	(23.6)	(48.0)	(0.5)
Obs	4430	4430	4430	4430	4430	4423
Acceptable/recognized	184.5***	99.6**	31.7***	13.2**	8.0	−0.1
	(69.2)	(46.4)	(9.7)	(5.3)	(14.0)	(0.1)
Obs	16101	16101	16101	16101	16101	16099
Recognized/exemplary	73.3	50.0	24.6*	12.8**	−14.4	0.2*
	(78.7)	(52.8)	(12.9)	(5.4)	(15.6)	(0.1)
Obs	9635	9635	9635	9635	9635	9635
VII. Control for once lagged observables						
Low/acceptable	75.4	188.3	29.7	−36.8**	−17.5	0.35
	(282.5)	(262.8)	(33.3)	(17.1)	(24.6)	(0.40)
Obs	616	267	1195	638	892	627
Acceptable/recognized	83.7	44.3	23.1***	11.2*	−5.5	−0.13
	(53.8)	(35.9)	(7.9)	(5.8)	(7.1)	(0.09)
Obs	6981	6223	5156	6076	7018	5145
Recognized/exemplary	129.5	39.8	36.8**	15.7***	1.6	0.15
	(78.8)	(51.0)	(15.0)	(5.7)	(12.6)	(0.11)
Obs	3110	3110	5266	4723	3798	3088

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 (see text). Career/tech and athletics results are already provided for high schools in the main regressions. 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. Except where noted schools below a boundary we further restrict to those who miss the cutoff in only one subject-group. Except where noted, we use the optimal bandwidths from Tables 5, 6, and 7. 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. Covariates for regressions controlling for lagged observables including enrollment, share of school white, black, Hispanic, economically disadvantaged, LEP, special education, gifted, and year dummies.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econedurev.2015.07.005.

References

Bacolod, Marigee, DiNardo, John, & Jacobson, Mireille. (2012). Beyond incentives: do schools use accountability rewards productively? *Journal of Business & Economic Statistics,, 30.1*, 149–163.

Carnoy, Martin, & Loeb, Susanna. (2003). Does external accountability affect student outcomes? a cross-state analysis. *Education Evaluation and Policy Analysis, 24*, 305–331.

Cullen, Julie Berry, & Reback, Randall. (2006). Tinkering toward accolades: school gaming under a performance accountability system. In Timothy J. Gronberg, & Dennis W. Jansen (Eds.), *Advances in Micoreconomics: vol. 14* (pp. 1–34). Amsterdam, NL: Elsevier.

Chakrabarti, Rajashri. (2008). 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.

Chakrabarti, Rajashri (2013a). Accountability with voucher threats, responses, and the test-taking population: regression discontinuity evidence from Florida. *Education, 8*(2), 121–167.

Chakrabarti, Rajashri (2013b). Vouchers, public school response, and the role of incentives: evidence from Florida. *Economic Inquiry, 51*, 500–526.

Chiang, Hanley. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*, 1045–1057.

Craig, Steven G., Scott, Imberman, & Adam, Perdue (2013). Does it pay to get an A? School district resource allocations in response to accountability ratings. *Journal of Urban Economics, 73*, 30–42 January.

Craig, Steven G., Imberman, Scott, & Perdue, Adam (2009). Does it pay to get an A? School district resource allocations in response to accountability ratings. Unpublished manuscript, University of Houston, Available at http://www.uh.edu/econpapers/RePEc/hou/wpaper/2009-04.pdf

Dee, Thomas S., & Jacob, Brian (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418–446.

Figlio, David. (2006). Testing, crime, and punishment. *Journal of Public Economics, 90*, 837–851.

Figlio, David, & Getzler, Lawrence (2006). Accountability, ability, and disability: gaming the system. In Timothy J. Gronberg, & Dennis W. Jansen (Eds.), *Advances in microeconomics: vol. 14* (pp. 35–50). Amsterdam, NL: Elsevier.

Figlio, David, & Winicki, Joshua (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics, 89*, 381–394.

Figlio, David, & Lucas, Maurice E. (2004). What's in a grade? school report cards and the housing market. *American economic review, 94*, 591–604 June.

Hanushek, Eric A., & Raymond, Margaret E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*, 297–327.

Hanushek, Eric A., & Raymond, Margaret E. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association, 2*, 406–415 April-May.

Jacob, Brian A. (2005). Accountability, incentives, and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*, 761–796.

Jacob, Brian A. (2003). Getting Inside accountability: lessons from Chicago. *Brookings-Wharton Papers on Urban Affairs*, 42–70.

Lee, David S. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics, 142*, 675–697.

Lee, David S., & Lemieux, Thomas (2009). Regression discontinuity designs in economics. NBER Working Paper #14723, February

Neal, Derek, & Schanzenbach, Diane Whitmore (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics, 92*(2), 263–283.

Reback, Randall (2008). Teaching to the rating: school accountability and the distribution of student achievement. *Journal of Public Economics, 92*, 1394–1415.

Rockoff, Jonah E., & Turner, Lesley J. (2010). Short run impacts of accountability on school quality. *American Economic Journal: Economic Policy, 2*(4), 119–147.

Rouse, Cecelia Elena, Hannaway, Jane, Goldhaber, Dan, & Figlio, David (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy, 5*(2), 251–281.