Journal of Urban Economics 73 (2013) 30–42

Contents lists available at SciVerse ScienceDirect

Journal of Urban Economics

www.elsevier.com/locate/jue

Does it pay to get an A? School resource allocations in response to accountability ratings

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

[a] Department of Economics, University of Houston, 204 McElhinney Hall, Houston, TX 77204-5019, United States
[b] Department of Economics, Michigan State University and NBER, Marshall Adams Hall, 486 W. Circle Dr. Rm. 110, East Lansing, MI 48824, United States

ARTICLE INFO

Article history:
Received 19 July 2011
Revised 8 June 2012
Available online 20 July 2012

JEL classification:
H72
H75
I22
I28

Keywords:
Education accountability systems
Education expenditures

ABSTRACT

This paper examines public school district budgetary responses to school accountability ratings. We identify school district budgetary changes through a "rating shock" due to a major change in school accountability systems in Texas. Texas implemented a new accountability system and new exam, and allowed schools a "gap" year to adjust to the new test. Using the new Texas exam as an exogenous change, 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 3 years, suggesting temporary budget support to "learn" the new system but no long run institutional change.

1. Introduction

Accountability systems have been fixtures of the US public education system since the late 1990s. 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.[1] 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 a change in accountability regimes in Texas to circumvent the identification problem, and to develop empirical evidence of how school districts employ accountability ratings to finance their schools.

The 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 2 year period from 2002 to 2004. The new accountability system was more rigorous than the old, thus many schools were threatened with a lower accountability grade than they had received on average from the old system. One aspect of the new system was that the exam used to "grade" schools was made more rigorous, and also covered broader material. While the exam was first given during the 2002–2003 school year, school ratings summarizing student performance were not given until 1 year later with the second testing cycle. This 'gap year' 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.

Our empirical examination finds that schools and their districts responded to the new exam by reallocating resources to schools where

* Corresponding authors.
E-mail addresses: scraig@uh.edu (S.G. Craig), imberman@msu.edu (S.A. Imberman).

[1] Dee and Jacob (2009) 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.

0094-1190/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jue.2012.07.002

there was an increased likelihood of a lower accountability grade.[2] Further, almost all of the incremental funds were directed to instruction, at least some of which led to increased teacher hiring. We find that this budgetary response was centered amongst schools that fell to Adequate (the second lowest category) from higher rankings, and amounted to about $75 per student, or 1.2% of total expenditures. As a result, instructional budgets in these schools 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 3 years from the change exhibit no difference from schools that did not experience ratings changes.

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 ratings are related to actual 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 Carnoy and Loeb (2003), Hanushek and Raymond (2004, 2005), Jacob (2005), Figlio and Rouse (2006), and Chakrabarti (2007, 2008). 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; Figlio, 2006) or focusing on marginal students (Chakrabarti, 2007; Reback, 2008; Neal and Schanzenbach, 2010). 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 et al. (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 et al. (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. None 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. In other work we also examine the impact of marginal changes in ratings and find little expenditure impacts (Craig et al., 2012). Our analysis here of the resource allocation response to accountability is considerably broader than both studies. First, we consider the response along different points of the rating distribution, so that we allow schools with different ratings to experience different levels of budgetary response. 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 (2009) and Craig et al. (2012).

It is important to note that our rating shock analysis differs from the Figlio and Kenney case, because the new Texas accountability exam differs from the previous exam in both its rigor, and in the subjects that are covered. In Florida, which they study, the exam and subjects covered remained the same. Thus, in Texas, it was difficult for school district administrators to predict how their schools will perform compared to other schools in the district, as well as compared to the testing boundary standards. Thus our example offers a clean break with the past, strengthening the identification that separates the funding decision from school quality. Our research response to the break in regimes is to exploit the gap year between programs, and thus to project the expected rating from the new process using the new standards. This prediction process is credible because the gap year test scores allowed administrators to forecast how well the students in their school would perform in the new testing subjects and when faced with the new exam rigor. The question we ask here is therefore how administrators use this new predicted information through budgetary allocations to address potential changes in school ratings.

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 virtually all of the increased resources are directed toward instruction, and primarily to reduce student–teacher ratios. Nonetheless, we also find that the resource increase is temporary, as we find it disappears within 3 years.

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).[3] 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.[4] A school's accountability rating under TAAS is based on the share of tested students in various student groups who pass the state-wide exam in each subject. Higher performance by one student does not compensate for lower performance by another; the criteria are Rawlsian based solely on whether each student passes the exam by attaining a score higher than a pre-specified minimum.[5] 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). 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.[6] With the exception of receiving a low rating, there were no direct punishments imposed on schools by the state, and awards provided by the state for high performance were

[2] 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 less formulaic ways.

[3] 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.

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

[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] 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/.

Table 1
Distributions of accountability ratings.

	Accountability rating				Rating changes		
	Low (%)	Acceptable (%)	Recognized (%)	Exemplary (%)	% Lower	% Same	% Higher
TAAS							
1997–1998	0.1	56.4	26.7	16.2	–	–	–
1998–1999	0.9	52.6	29.4	17.1	14.4	67.6	18.0
1999–2000	1.6	47.4	31.9	19.1	14.4	65.9	19.7
2000–2001	0.7	39.3	36.9	23.1	12.4	62.9	24.8
2001–2002	1.4	32.0	37.8	28.9	15.3	59.4	25.3
Transition year (2002–2003)							
Estimate based on 2003–2004 Rules[a]	8.6	70.3	18.2	2.8	62.3	34.2	3.5
TAKS							
2003–2004[b]	0.9	55.5	37.2	6.5	47.8	42.2	10.0
2003–2004 without Req Improvement[c]	1.1	62.2	30.2	6.5	49.7	41.9	8.4
2004–2005	3.0	64.7	28.1	4.2	24.4	68.1	7.4
2005–2006	3.5	47.3	41.6	7.7	8.0	64.8	27.2
2006–2007	3.4	54.2	33.8	8.6	20.0	66.0	14.1

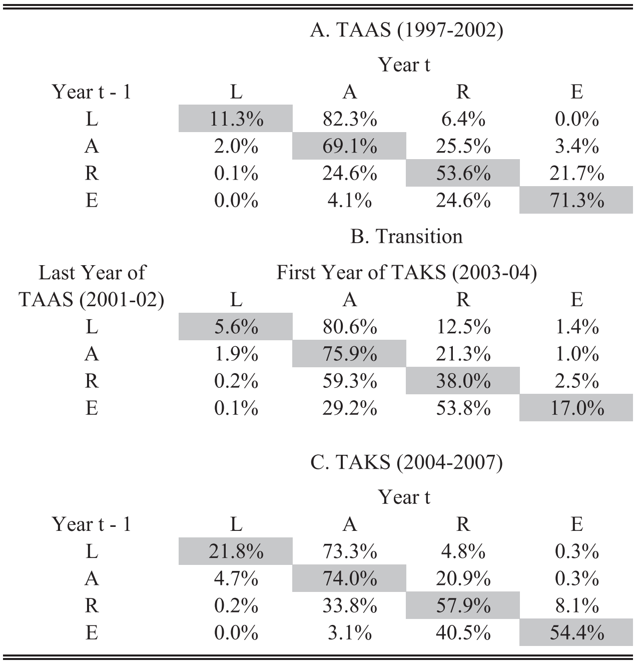
Sample is limited to schools that received an L, A, R, or E rating. For changes the school must have received a rating in both years. Schools that received ratings on appeal, were paired with another school, were identified as having undergone "special analysis," or have fewer than 200 students are excluded.

[a] Only for schools that have a regular rating in 2001–2002. We do not account for required improvement as the implementation of a new exam makes such a calculation impossible.

[b] Ratings changes are relative to 2001–2002.

[c] This is the actual ratings had required improvement not been added to the criteria, which corresponds to our transition year estimate.

Table 2
Rating transitions under different accountabilty 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	First Year of TAKS (2003-04)			
TAAS (2001-02)	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.

extremely small.[7] Thus this initial rating system primarily acted 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 1 shows the distribution of school ratings by year. 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. Further, school accountability grades were steadily increasing over time as the share of schools rated A fell and schools rated R or E rose.

After a 1 year transition we refer to as the "gap year" during 2002–2003, the new accountability system called the Texas

[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 two years in a row from 1998–1999 to 2001–2002. 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 $5490.

Assessment of Knowledge and Skills (TAKS) was implemented for the 2003–2004 year. Under TAKS, the ratings distribution shifted downwards, demonstrated in the lower half of Table 1, as fewer schools were awarded E's and more were rated L or A. Further, the rate at which the ratings were increasing over time seems to have significantly slowed, if not stopped altogether. 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.[8] 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, requirements for science and social studies in more grades, and stricter passing requirements. Details on the requirements are provided in Appendix Table A2. As a result of the 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 2002–2003, but no school accountability grades were released. Thus our rating shock strategy is able to examine how districts and schools financially respond to the 2002–2003 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. Rating shock methodology

Our examination of school district behavior concentrates on the total operating expenses of each school, which reflects the school district's allocation strategy.[9] Further, we expand our examination by analysis of the allocation of funds among categories of expenditure within each school. We view expenditure by funding category as a reduced form measure reflecting the combined decisions of both the school district, and the individual school principals. One reason it is difficult to disentangle district from school decisions is that school teachers are generally paid equally throughout a district, including adjustments for special training and experience. Thus to control for salary differences compared to resource allocation differences we will also examine the number of teachers and their characteristics rather than just expenditures. Our research strategy is therefore to determine how total expenditure, expenditure by category, and physical inputs have changed as a result of the rating shock forecast by the exam results from the gap year between the old TAAS system and the new TAKS accountability assessment.

While administrators were not provided with explicit ratings during the transition year between the two systems, administrators at both the school and school district level were experienced with accountability systems. Thus the changes we examine by both school districts and within schools are likely to result from the administrators' ability to approximate their rating from the testing information. Further, the administrators experience will inform them as to the importance of the ratings, both from the perspective of interaction with the public and politicians, as well as any internal labor market characteristics. Hence, to implement the 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.[10] In particular, we test how schools that faced a rating reduction in 2002–2003 were either granted new resources by their school districts, and/or reallocated their budget during 2003–2004, the first year in which the new TAKS system was fully implemented.[11] The rating shock is therefore a strong test of the response to the ratings, because the 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.

A key to the rating shock strategy as outlined above is that the rating we calculate in the gap year of 2002–2003 is correlated with the information administrators use to respond to the test results from the new TAKS exam. When the gap year ratings are compared to ratings for the first year when accountability grades are given – shown in Table 1 – we see that the first year accountability grades are considerably higher than those we calculate for the gap year. While this could reflect district and school responses, a complicating factor is that we are not able to account for an additional administrative feature of TAKS called "required improvement (RI)," which essentially allows schools to get the next higher rating if they show sufficient improvement. Such a calculation requires knowledge of prior achievement on TAKS, which is not possible in 2002–2003 since this is the first year of the new exam. Hence, we calculate the gap year ratings ignoring required improvement. As a result, in order to make a more relevant comparison to later ratings, the row of Table 1 labeled "2003–2004 without Req Improvement" shows the ratings that would have resulted without this administrative feature.[12] While we do not know all of the causes of differences in accountability results between the gap year and the first year of TAKS, this row in Table 1 shows that required improvement is not the only factor that caused a change in the ratings after the gap year results were known.[13]

The regression using the rating shock examines the change in total expenditures, categorical expenditures, or staffing from the 2002–2003 school year to the 2003–2004 school year based on whether the estimated accountability rating in 2002–2003 fell from its rating in 2001–2002, the last year of the TAAS system. Hence we estimate:

$$\begin{aligned}\Delta R_{i,2003-2004} = \alpha &+ \beta_1 Drop_{i,2002-2003} \times Rating^{L}_{i,2002-2003} \\ &+ \beta_2 Drop_{i,2002-2003} \times Rating^{A}_{i,2002-2003} \\ &+ \beta_3 Drop_{i,2002-2003} \times Rating^{R}_{i,2002-2003} \\ &+ \beta_4 Increase_{i,2002-2003} \\ &+ \sum_{r\in \mathrm{L,A,R}} \sum_{t=1999-2000}^{2001-2002} \gamma_{jt} Rating^{r}_{i,t} + \boldsymbol{X_i\Omega} + \varepsilon_i \qquad (1)\end{aligned}$$

[8] 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.

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

[10] 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 the new (2003–2004) rules to estimate the 2002–2003 ratings.

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

[12] In Online Appendix Table A1 we provide the complete two-way distribution of ratings (with and without RI) We also provide the distribution of actual ratings in 2003–2004 relative to predicted ratings based on TAAS rules. We note though that these are poor approximation of what the ratings would have been under TAAS due to the fact that the TAKS exams are harder and hence it is much more difficult to reach the higher TAAS passing cutoffs using the TAKS exams. Further, Texas changed how it calculated and incorporated dropouts into the accountability system from TAAS to TAKS. Hence, we must ignore dropouts in these calculations. At best, therefore, this gives us an extreme upper bound on what the ratings would have been had the rules and the test difficulty stayed the same.

[13] Note that the gap year ratings are not a forecast for 2003–2004, just illustrations of the effects of the new more difficult exam. Since we show required improvement does not account for the entire difference in the 2 years' results, the budget changes below are credibly a result of the new exam results consistent with the rating shock strategy. If administrators predicted in 2002–2003 that required improvement would be a feature of the new system and were able to approximate whether their schools met the requirements, our estimate of the budget response is an under-estimate since we "over-predict" the number of schools that fall to a specific accountability grade.

Table 3
School characteristics by rating.

	TAAS (1997–1998 to 2001–2002)				TAKS (2002–2003 to 2006–2007)			
	Low	Acceptable	Recognized	Exemplary	Low/unacceptable	Acceptable	Recognized	Exemplary
% Asian	1.1	1.8	2.0	3.2	0.7	1.9	3.1	6.4
	(2.3)	(3.4)	(4.1)	(5.3)	(1.5)	(3.5)	(5.5)	(8.5)
% Black	29.5	18.6	11.1	7.1	30.2	15.6	11.0	7.3
	(26.8)	(22.6)	(16.6)	(11.7)	(29.8)	(19.4)	(14.7)	(10.2)
% Hispanic	48.4	43.8	39.4	24.5	52.6	49.3	39.6	22.0
	(30.1)	(31.7)	(32.3)	(28.6)	(32.3)	(31.5)	(30.9)	(24.7)
% White	20.8	35.6	47.1	64.9	16.4	32.8	45.9	63.
	(23.6)	(29.0)	(31.3)	(29.1)	(22.4)	(29.0)	(30.4)	(26.1)
% Economically disadvantaged	70.9	59.9	51.0	31.9	76.9	62.4	51.6	28.1
	(22.5)	(25.0)	(26.4)	(28.0)	(19.0)	(24.9)	(27.2)	(28.0)
% LEP	23.5	16.6	13.3	8.0	19.6	17.7	14.8	9.5
	(23.1)	(19.4)	(17.7)	(14.3)	(22.3)	(19.8)	(18.0)	(13.9)
% Special ed.	18.3	14.1	12.0	13.6	28.8	19.9	6.5	1.8
	(26.4)	(23.7)	(23.8)	(26.6)	(30.4)	(28.7)	(17.7)	(10.7)
% Gifted	7.7	7.6	7.5	9.5	6.7	7.3	7.2	9.0
	(6.3)	(6.7)	(6.6)	(9.5)	(4.7)	(5.2)	(7.0)	(10.1)
Enrollment	812	749	633	601	781	783	606	595
	(503)	(502)	(410)	(383)	(519)	(567)	(336)	(226)
Observations	285	12,111	8614	5550	593	12,537	7985	1537

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.	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)
	B. TAKS (year t)			
Resources in Year $t+1$	Low/unacceptable	Acceptable	Recognized	Exemplary
Total operating exp.	7089	6343	6039	5766
	(1567)	(1214)	(1232)	(982)
Instructional exp.	4982	4581	4518	4412
	(957)	(790)	(827)	(669)
Admin. and training exp.	778	625	577	540
	(280)	(179)	(179)	(149)
Counseling exp.	288	243	217	201
	(128)	(109)	(100)	(91)
Extra-curricular exp.	214	180	80	36
	(315)	(293)	(188)	(89)
Student–teacher ratio	14.0	14.6	14.8	15.2
	(2.4)	(2.3)	(3.1)	(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 per student, deflated by the CPI.

where ΔR is the change in resources per student in a given category from the transition year (2002–2003) to the next, for either expenditures or employees in school *i*. *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, limited English Proficiency (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 2002–2003 with *Drop* – a dummy variable which equals one if the calculated rating in 2002–2003 is lower than the last TAAS rating in 2001–2002. 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 2001–2002 but where one received the same (imputed) rating in 2002–2003 while the other received a lower rating based on the new test. Table 1 shows that

[14] Our estimate of the effect of a rating drop is not sensitive to the inclusion of these past ratings. We also cluster the standard errors by school district.

Table 5
Estimates of impact of ratings changes on resource changes in transition year.

	Δ From 2002–2003 to 2003–2004					
	Total	Instruction	Admin. and training	Counseling	Extra-curricular	Student–teacher ratio
	(1)	(2)	(3)	(4)	(5)	(6)
Estimated rating falls to L	−16.4	12.6	−0.7	−2.4	2.9	−0.13
	(49.9)	(39.2)	(5.7)	(3.1)	(3.7)	(0.08)
Estimated rating falls to A	72.9**	66.2**	8.0	2.3	−4.1	−0.28***
	(34.2)	(26.3)	(6.0)	(2.9)	(2.8)	(0.07)
Estimated rating falls to R	−5.8	13.8	−1.5	1.5	−3.1	−0.11
	(38.0)	(27.2)	(6.2)	(3.5)	(3.1)	(0.07)
Estimated rating increases	−15.3	−10.5	1.1	−2.5	0.7	0.13
	(83.0)	(72.5)	(14.0)	(6.6)	(4.7)	(0.13)
Low rating in 2001–2002	187.1	71.8	77.8***	27.7***	9.6	−0.46*
	(179.3)	(176.3)	(28.8)	(8.8)	(7.8)	(0.26)
Acceptable rating in 2001–2002	168.5***	151.6***	22.4***	8.1*	−3.4	−0.33***
	(48.5)	(38.2)	(8.0)	(4.4)	(4.1)	(0.08)
Recognized rating in 2001–2002	64.9**	49.0**	12.5***	4.8**	0.6	−0.05
	(26.5)	(19.7)	(3.8)	(2.4)	(2.5)	(0.05)
Low rating in 2000–2001	−213.2**	−254.8***	24.4	6.1	−3.0	−0.24
	(108.4)	(94.1)	(40.0)	(8.8)	(8.6)	(0.15)
Acceptable rating in 2000–2001	−31.3	−42.9*	−5.1	−2.7	9.1**	0.04
	(31.6)	(24.9)	(4.7)	(2.7)	(3.7)	(0.06)
Recognized rating in 2000–2001	20.4	−3.9	−0.3	−1.7	9.7***	−0.03
	(25.3)	(19.7)	(3.6)	(2.2)	(3.1)	(0.05)
Low rating in 1999–2000	−127.2	−126.7*	4.4	−4.6	−7.2	0.35*
	(98.3)	(74.7)	(12.6)	(6.7)	(5.3)	(0.20)
Acceptable rating in 1999–2000	−72.8**	−41.3*	−7.7*	−1.4	−8.4**	0.12**
	(29.8)	(23.1)	(4.6)	(2.9)	(3.5)	(0.06)
Recognized rating in 1999–2000	−88.9***	−63.4***	−2.8	−2.1	−5.0	0.12**
	(25.9)	(19.9)	(3.8)	(2.7)	(3.3)	(0.05)
Joint significance test for any drop in rating (*F*-statistic)	3.70**	3.39**	1.74	0.65	1.56	7.62***
Observations	4958	4958	4958	4958	4958	4957

Sample is limited to schools that received an L, A, R, or E rating in 1999–2000, 2000–2001, 2001–2002 and 2003–2004. 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 by school district in parentheses.

* Significance at the 10% levels.

** Significance at the 5% levels.

*** Significance at the 1% levels.

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 provides 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 re-allocate funds between schools, and how those funds are used within the schools. It is this last set of potential responses that motivates our work here.

The estimates from Eq. (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 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.

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.[15] 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.[16]

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

[15] We use the actual expenditures by schools rather than budgeted expenditures.

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

Table 6
Placebo estimates of “Impact” of ratings changes in transition year on resource changes from 2001–2002 to 2002–2003.

	Δ From 2001–2002 to 2002–2003					
	Total (1)	Instruction (2)	Admin. and training (3)	Counseling (4)	Extra-curricular (5)	Student–teacher ratio (6)
Estimated rating in 2002–2003 falls to L	31.3	1.0	−3.9	2.4	2.5	0.00
	(44.0)	(31.6)	(7.1)	(4.0)	(3.6)	(0.07)
Estimated rating in 2002–2003 falls to A	−13.7	−28.7	−1.3	2.3	2.4	0.12**
	(32.3)	(23.8)	(5.0)	(3.2)	(2.9)	(0.06)
Estimated rating in 2002–2003 falls to R	18.8	7.1	−1.2	−2.2	2.7	−0.04
	(39.1)	(29.4)	(6.2)	(3.9)	(3.4)	(0.08)
Estimated rating in 2002–2003 increases	−2.6	−34.7	21.5	4.4	−1.7	−0.14
	(64.4)	(52.7)	(15.5)	(5.3)	(3.3)	(0.16)
Low rating in 2001–2002	273.4***	229.7***	−33.7	2.5	2	−0.11
	(101.8)	(79.2)	(34.0)	(9.3)	(6.5)	(0.24)
Acceptable rating in 2001–2002	87.6**	42.3	9.1	6.4	5.2	−0.02
	(41.8)	(32.1)	(6.5)	(4.2)	(4.4)	(0.08)
Recognized rating in 2001–2002	59.8**	40.2**	10.1***	3.8	5.9**	−0.07
	(24.7)	(19.0)	(3.6)	(2.3)	(2.8)	(0.05)
Low rating in 2000–2001	357.3***	226.2**	7.4	−4.5	20.9**	−0.44*
	(117.5)	(90.3)	(40.5)	(11.0)	(10.5)	(0.25)
Acceptable rating in 2000–2001	78.9**	43.9*	10.6**	1.7	−2.5	−0.10
	(34.8)	(25.7)	(5.2)	(3.2)	(3.9)	(0.06)
Recognized rating in 2000–2001	16	0.1	3	−1.8	−4.1	−0.02
	(26.8)	(20.4)	(4.1)	(2.5)	(3.1)	(0.05)
Low rating in 1999–2000	−92	−56.5	−15.4	8	−0.7	−0.05
	(85.0)	(68.7)	(23.0)	(7.2)	(5.7)	(0.17)
Acceptable rating in 1999–2000	−38.1	−16.8	−2	1.1	2.2	0.06
	(31.9)	(24.3)	(4.7)	(3.0)	(4.0)	(0.06)
Recognized rating in 1999–2000	11.5	18.7	1.5	1.6	2.4	-0.06
	(28.5)	(21.6)	(4.2)	(2.7)	(3.5)	(0.05)
Observations	4997	4997	4997	4997	4997	4996

Sample is limited to schools that received an L, A, R, or E rating in 1999–2000, 2000–2001, 2001–2002 and 2003–2004. 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 by school district in parentheses.

* Significance at the 10% levels.

** Significance at the 5% levels.

*** Significance at the 1% levels.

expenditures are divided by enrollment to provide per-student measures and inflated to 2007 dollars using the CPI.

Our data starts using all of the schools in Texas. We then drop alternative schools, charter schools, “paired” schools, those with special analysis, and schools under 200 students.[17] After these restrictions, we have data on 4958 schools across Texas in 784 school districts. The largest district (Houston) has 279 schools, and 627 districts have only one elementary school. While we have data from the 1997–1998 school year through the 2006–2007 school year, our primary analysis utilizes the change in expenditures from 2002–2003 to 2003–2004 as this is the time period encompassing the change in accountability regime.

Table 3 presents the means of school characteristics for each of the two accountability regimes. In general, the table shows 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 per student than those with higher ratings, although this gap widens in TAKS. These differences generally hold across all expenditure categories.

5. Empirical results

Table 5 shows our primary ratings-shock results. To keep the identification as clear as possible, this specification is estimated based on resource data only from the gap year (2002–2003), and the year immediately after the gap year (2003–2004).[18] Thus the results pertain to how the initial student test performance from the TAKS exam, which the school could not know before their students experienced the new exam, affected the budget of the subsequent year. The estimated parameters show the change in resources from the 2002–2003 transition year to the first official year of TAKS (2003–2004) 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 3 years.[19]

The evidence from Table 5 suggests school districts allocate $73 per student to schools that experience a drop in their rating to A.[20]

[17] 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 ratings.

[18] Although note that since we include lagged ratings as controls, the schools had to be rated in the previous 2 years as well.

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

[20] In Online Appendix Table A2 we estimate models that use 2001–2002 as the base year instead of 2002–2003. Doing this makes the estimate less susceptible to anticipatory effects but also increases measurement error in the dependent variable. Given that the budgets were determined well before the 2002–2003 exams were taken, we think that it is unlikely that the districts made anticipatory responses. Nonetheless, the estimates for total and instructional spending in Online Appendix Table A2 fall to insignificance but remain positive, and not statistically different from those in Table 5. The estimate for teacher–student ratios remains statistically significant, and also not different from the estimate in Table 5.

Table 7
Estimates of impact of ratings changes in transition year on prior changes in student characteristics.

	Δ From 2001–2002 to 2002–2003					
	% African–American (1)	% Hispanic (2)	% Econ disadv. (3)	% LEP (4)	% Gifted (5)	% Special education (6)
Estimated rating drops to L	0.029	0.007	−0.229	−0.287	−0.193	0.436
	(0.133)	(0.168)	(0.230)	(0.209)	(0.213)	(0.434)
Estimated rating drops to A	0.237*	0.059	0.273	0.214	0.234	0.002
	(0.123)	(0.143)	(0.215)	(0.150)	(0.144)	(0.282)
Estimated rating drops to R	−0.063	−0.058	0.074	0.077	0.114	0.360
	(0.138)	(0.179)	(0.250)	(0.169)	(0.164)	(0.301)
Estimated rating increases	0.149	−0.994***	−1.134*	−0.147	−0.371	−0.326
	(0.206)	(0.378)	(0.583)	(0.395)	(0.476)	(0.458)
Low rating in 2001–2002	−0.131	0.702	0.539	−0.326	0.535	−0.003
	(0.363)	(0.465)	(0.762)	(0.672)	(0.857)	(1.084)
Acceptable rating in 2001–2002	0.149	0.334*	0.211	0.329	0.384*	0.442
	(0.147)	(0.192)	(0.297)	(0.208)	(0.197)	(0.406)
Recognized rating in 2001–2002	0.048	0.039	0.094	0.088	0.087	0.740***
	(0.080)	(0.107)	(0.166)	(0.120)	(0.113)	(0.261)
Low rating in 2000–2001	−0.558	0.933**	1.285	−0.530	−0.112	2.173
	(0.437)	(0.469)	(0.894)	(0.545)	(0.601)	(1.642)
Acceptable Rating in 2000–2001	−0.039	0.071	0.248	0.022	0.079	−0.568
	(0.102)	(0.154)	(0.216)	(0.155)	(0.147)	(0.360)
Recognized rating in 2000–2001	0.021	0.103	−0.033	0.177	0.103	−0.419
	(0.081)	(0.122)	(0.169)	(0.120)	(0.114)	(0.273)
Low rating in 1999–2000	−0.292	−0.009	−0.383	−0.695	−0.039	−0.142
	(0.269)	(0.370)	(0.597)	(0.564)	(0.755)	(0.701)
Acceptable rating in 1999–2000	−0.073	0.160	0.271	0.064	−0.035	0.213
	(0.096)	(0.148)	(0.216)	(0.150)	(0.142)	(0.348)
Recognized rating in 1999–2000	0.002	0.137	0.332*	0.091	0.074	0.150
	(0.088)	(0.127)	(0.179)	(0.126)	(0.122)	(0.277)
Observations	4997	4997	4997	4997	4997	4997

Sample is limited to schools that received an L, A, R, or E rating in 1999–2000, 2000–2001, 2001–2002 and 2003–2004. 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.

* Significance at the 10% levels.

** Significance at the 5% levels.

*** Significance at the 1% levels.

This is about 1.2% of the total school budget for the average A school. The second column of the table shows that virtually all of the funds are allocated to instructional expenses (the coefficient of 66.2 is not significantly different from 72.9). This 1.5% increase in instructional resources is found to generate a statistically significant reduction in the student/teacher ratio in column 6 of almost 0.3%.[21] For schools where the predicted rating falls to L or R we find no statistically significant changes in any resource.[22]

Table 6 provides estimates from a "placebo test," where we estimate expenditure in the prior year as a function of the future rating shock. We expect this test to show no effect, significant results would suggest potential endogeneity. All of the estimates are insignificant except student teacher ratios, which are found to have a positive effect while the primary effect shown in Table 5 is negative.[23] Table 7 provides an additional validity test, where we test whether the rating shock drops are associated with prior student characteristics, including the percent of students who are African–American, Hispanic, economically disadvantaged, LEP, gifted and need special education from 2001–2002 to 2002–2003. Of these, only the percent African–American for schools that fall to an A is significant and only at the 10% level.[24] Table 8 explores some of the potential mechanisms that could generate the findings in Table 5. First, we divide the student–teacher ratio into each of its components, and find no significant differences in either student enrollment as seen in column (1), or in FTE teachers as in column (2). The lack of a clear differentiation suggests both process could be at work, but the relatively large standard errors prevent a definite conclusion. Further, the other factors in Table 8 point to, if anything, a reduction in expenses per student, as average teacher experience falls slightly, and the share of new teachers rises slightly (though neither is statistically significant), both factors that should reduce dollars per student through lower teacher compensation. Thus we believe the weight of the evidence is that a substantial portion of the $72.90 per student in expenditure increase that we find in Table 5 is the result of explicit policy change by school districts to bolster the quantity of teaching resources in schools that are threatened with a decrease in accountability ratings.

It could be reasonably expected that school districts which desire to re-allocate resources to schools will be more likely to do so depending on the opportunity cost. That is, to the extent district resources are fixed, a re-allocation is more likely to occur if the institutions which lose money in the reallocation do not suffer

[21] We get qualitatively similar estimates with school district fixed effects, presented in online Appendix Table A1.

[22] If we apply Bonferroni corrections to the estimates their statistical significance does not change.

[23] We also provide estimates for placebo tests three years prior to the transition in Online Appendix Table 3. The total and instructional expenditures are significant for schools that fall to an L but only at the 10% level. Nonetheless we do find statistically significant increases in extra-curricular spending for schools that fall to an A and for those that fall to an R, but the amounts are economically small at $6 and $9 per student for A and R schools, respectively.

[24] In Online Appendix Table 4 we also provide similar tests for changes from 2002–2003 to 2003–2004. Since this is during the transition period, the estimates could reflect changes in student body characteristics in response to policy changes and the gap year test results. Nonetheless, in most cases the estimates are statistically insignificant. Percent African–American for schools that drop to an L and percent LEP for schools that drop to an A are significant but only at the 10% level. Percent Hispanic is significant at the 1% level for schools that fall to an R. Even so, we note that for each type of rating shock only one estimate is significant.

Table 8
Estimates of impact of ratings changes on teacher/staff characteristics in transition year.

	Δ From 2002–2003 to 2003–2004						
	Enrollment	# FTE teachers	Avg. teacher experience	Percent beginning teachers	Percent of teachers 1–5 years exp.	Percent of teachers >5 years exp.	Student-educational aide ratio
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estimated rating falls to L	−5.4	0.2	−0.0	−0.2	0.1	0.1	3.7
	(3.5)	(0.3)	(0.1)	(0.4)	(0.4)	(0.4)	(4.5)
Estimated rating falls to A	−5.2	0.2	−0.2	0.3	0.3	−0.5	−3.2
	(3.4)	(0.3)	(0.2)	(0.4)	(0.4)	(0.4)	(6.2)
Estimated rating falls to R	−3.1	−0.1	0.1	−0.5	0.6	−0.0	6.1
	(4.8)	(0.3)	(0.2)	(0.4)	(0.5)	(0.4)	(9.1)
Estimated rating increases	2.0	−0.1	−0.6**	0.4	−1.7**	1.3	−7.3
	(7.3)	(0.5)	(0.3)	(0.7)	(0.7)	(0.8)	(4.9)
Low rating in 2001–2002	−27.6***	−0.7	0.1	−1.3	2.3*	−1.0	11.6
	(10.6)	(0.8)	(0.5)	(1.2)	(1.3)	(1.1)	(11.4)
Acceptable rating in 2001–2002	−13.9***	−0.1	−0.3	0.3	0.4	−0.7	−1.8
	(4.9)	(0.3)	(0.2)	(0.5)	(0.5)	(0.5)	(6.3)
Recognized rating in 2001–2002	−9.0***	−0.5**	−0.1	−0.3	0.3	0.0	4.5
	(3.1)	(0.2)	(0.1)	(0.3)	(0.3)	(0.3)	(5.3)
Low rating in 2000–2001	−11.6	0.2	−1.1***	−2.1	0.6	1.5	−8.6
	(9.3)	(0.7)	(0.3)	(1.6)	(1.7)	(1.3)	(14.4)
Acceptable rating in 2000–2001	4.7	−0.1	−0.3**	−0.8**	0.7*	0.1	−10.4
	(3.1)	(0.2)	(0.1)	(0.4)	(0.4)	(0.4)	(7.8)
Recognized rating in 2000–2001	−0.8	−0.1	−0.1	−0.7**	0.9**	−0.2	−3.9
	(2.5)	(0.2)	(0.1)	(0.3)	(0.4)	(0.3)	(6.6)
Low rating in 1999–2000	−4.9	−0.8	−0.8***	−0.9	−0.7	1.7*	9.8
	(10.9)	(0.6)	(0.3)	(0.9)	(1.0)	(0.9)	(8.9)
Acceptable rating in 1999–2000	0.2	−0.2	−0.1	0.1	−0.4	0.4	11.7
	(3.8)	(0.2)	(0.2)	(0.4)	(0.4)	(0.4)	(7.8)
Recognized rating in 1999–2000	5.3*	0.0	0.0	0.5	−0.2	−0.2	9.9
	(3.1)	(0.2)	(0.1)	(0.3)	(0.4)	(0.4)	(6.9)
Observations	4958	4957	4956	4957	4957	4957	4821

Sample is limited to schools that received an L, A, R, or E rating in 1999–2000, 2000–2001, 2001–2002 and 2003–2004. 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 by school district in parentheses.

* Significance at the 10% levels.

** Significance at the 5% levels.

*** Significance at the 1% levels.

Table 9
Estimates of impact of ratings changes on resource changes in transition year with school district concentration characterisitcs.

	Δ From 2002–2003 to 2003–2004					
	Total (1)	Instruction (2)	Admin. and training (3)	Counseling (4)	Extra-curricular (5)	Student–teacher ratio (6)
I. Baseline estimates from Table 5						
Estimated rating falls to A	72.9**	66.2**	8.0	2.3	−4.1	−0.28***
	(34.2)	(26.3)	(6.0)	(2.9)	(2.8)	(0.07)
II. Interact with Herfindahl–Hirschman index (HHI) of distribution of ratings in district						
Estimated rating falls to A	212.6***	185.5***	21.2	8.9	−0.3	−0.43***
	(80.7)	(55.9)	(13.8)	(7.2)	(5.7)	(0.15)
Estimated rating falls to A	−218.8*	−185.5**	−19.1	−10.1	−1.9	0.14
∗HHI	(119.2)	(90.7)	(19.4)	(11.9)	(8.9)	(0.25)
III. Interact with share of schools in district that have predicted ratings higher than A in 2002–2003						
Estimated rating falls to A	54.2	62.1	4.3	−2.7	−3.6	−0.32***
	(47.3)	(37.7)	(8.1)	(3.9)	(4.1)	(0.10)
Estimated rating falls to A	89.6	27.7	13.6	21.6**	−4.7	0.11
∗Share of schools rated R or higher	(110.9)	(84.4)	(15.5)	(10.2)	(10.3)	(0.21)
IV. Interact with share of schools in a district where rating falls						
Estimated rating falls to A	113.8	132.2**	−9.5	2.6	−2.8	−0.31**
	(71.2)	(52.5)	(12.4)	(7.5)	(7.7)	(0.13)
Estimated rating falls to A	−81.9	−117.4	29.6	0.7	−3.9	0.07
∗Share of schools where rating falls	(103.9)	(79.6)	(19.0)	(11.5)	(11.5)	(0.18)
V. Interact with whether district has >20 schools						
Estimated rating falls to A	74.2**	68.2**	9.6	−0.0	−3.3	−0.25***
	(36.8)	(29.6)	(6.1)	(3.4)	(3.7)	(0.07)
Estimated rating falls to A	−3.8	−1.5	−2.5	4.2	−1.1	−0.05
∗>20 Schools	(36.8)	(27.6)	(7.2)	(3.6)	(3.2)	(0.08)
VI. Sample restricted to districts with a single elementary school						
Estimated rating falls to A	19.5	43.7	−18.3	−12.9	−7.0	0.15
	(110.1)	(95.0)	(14.5)	(9.8)	(12.6)	(0.17)

Sample is limited to schools that received an L, A, R, or E rating in 1999–2000, 2000–2001, 2001–2002 and 2003–2004. Sample in panel II is limited to districts with 5 or more schools. 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 by school district in parentheses. Each of the segments reported here is reported in full in Online Appendix Tables 5–10.

* Significance at the 10% levels.

** Significance at the 5% levels.

*** Significance at the 1% levels.

accountability rating losses. For example, districts with few schools may have a more difficult time finding resources for schools that perform poorly in the gap year because there are fewer sources of funds. It should be noted that other schools are not the only source of funds, for example money could be taken out of reserve funds, local resources could be increased, or central administration expenses could be reduced.[25] Nonetheless, we explore these issues in Table 9 by providing a series of interaction tests. Note that each panel of Table 9 provides results from a separate regression, and to save space we only show the results for "falling to an A" and the interaction terms. Full results for each specification are provided in Online Appendix Tables 5–10.

The first row of Table 9 repeats the estimates from Table 5. Subsequent panels show results for interactions with features of school districts which may reflect the opportunity cost of the reallocation of resources between schools. In Panel II we create a Herfindahl–Hirschman Index (HHI) of the distribution of ratings within a district.[26] A value of one implies that all schools in the district have the same rating, while a value of .25 implies an even distribution of ratings. The results suggest that districts with more diverse sets of schools in terms of ratings provide more reallocation than districts that have schools concentrated in ratings.[27] Another indicator for more flexibility would be if the district has a large percentage of highly rated schools from which to transfer resources to schools that fall to an A. We thus use the share of schools rated R or higher as an interaction term in Panel III, and find that while insignificant, the point estimates indicate more reallocation with a higher share of R and E schools. Similarly in panel IV we see a negative but insignificant interaction effect of "falling to an A" with the share of schools in a district suffering any drop in rating. Arguably, a district with more schools that are "shocked" has less flexibility to re-allocate funds. Panel V shows that districts with more schools are not any more or less likely to re-allocate funds than small districts. Finally, Panel VI shows that districts with a single elementary school perform little reallocations. We interpret these results as providing suggestive evidence that more reallocation occurs when the districts have more flexibility.

To test whether the results in Table 5 are permanent or temporary expenditure changes, we re-estimate the rating shock regression, but on the budget change 3 years following the change in accountability regimes. Thus estimation results in Table 10 examine how expenditures in the 2005–2006 school-year respond to the rating shock in 2002–2003.[28] These results show that changes in school budgets 3 years out for schools with lower ratings in the

[25] Local resources would have to be from parents, as Texas has a "robin hood" state aid law that adjusts to local tax revenue.

[26] We restrict the sample for this analysis to districts with five or more schools. The Herfindahl index is constructed in the usual way using the share of schools with each of the four accountability grades.

[27] In fact, our estimates show no reallocations if all the schools in the same district have identical ratings ($213–$219).

[28] The 2 year estimates are about halfway between those in Tables 5 and 10.

Table 10
Estimates of impact of ratings changes in transition year on long-term (3 years) resources.

	Δ From 2002–2003 to 2005–2006					
	Total (1)	Instruction (2)	Admin. and training (3)	Counseling (4)	Extra-curricular (5)	Student–teacher ratio (6)
Estimated rating drops to L	49.0	47.8	18.4*	5.0	−0.7	−0.14
	(52.1)	(37.2)	(9.5)	(5.2)	(4.7)	(0.10)
Estimated rating drops to A	17.8	29.7	7.5	3.0	−5.5	−0.10
	(41.4)	(30.6)	(7.2)	(4.3)	(3.9)	(0.08)
Estimated rating drops to R	−73.0	−28.5	−9.6	−2.2	−4.1	0.04
	(50.1)	(37.0)	(8.5)	(5.3)	(4.4)	(0.10)
Estimated rating increases	51.7	9.7	9.3	1.0	2.7	0.14
	(99.0)	(79.8)	(13.5)	(9.1)	(4.9)	(0.17)
Low rating in 2001–2002	−158.0	−129.4	46.4	11.5	−5.4	−0.07
	(201.7)	(163.3)	(28.6)	(13.0)	(8.8)	(0.29)
Acceptable rating in 2001–2002	130.0**	144.4***	34.1***	8.6	−10.8*	−0.08
	(58.4)	(42.2)	(10.0)	(6.0)	(5.9)	(0.11)
Recognized rating in 2001–2002	72.1**	73.5***	12.8**	4.4	−5.6	0.04
	(35.3)	(25.1)	(5.7)	(3.4)	(3.6)	(0.06)
Low rating in 2000–2001	−490.2**	−421.8**	−32.9	8.4	−8.8	0.14
	(203.3)	(170.4)	(41.8)	(13.5)	(12.1)	(0.28)
Acceptable rating in 2000–2001	−9.3	−25.0	−6.2	−2.2	3.8	0.02
	(44.3)	(32.9)	(7.7)	(4.5)	(4.7)	(0.08)
Recognized rating in 2000–2001	16.7	−3.9	0.1	0.6	2.7	−0.06
	(35.9)	(27.1)	(6.0)	(3.7)	(3.7)	(0.07)
Low rating in 1999–2000	−164.0	−110.2	−15.9	−18.1**	4.6	0.32
	(141.3)	(119.4)	(19.7)	(8.8)	(8.0)	(0.21)
Acceptable rating in 1999–2000	−119.7***	−71.9**	−17.8**	−7.8*	−5.5	0.08
	(41.8)	(30.9)	(7.0)	(4.4)	(4.8)	(0.08)
Recognized rating in 1999–2000	−99.2***	−73.7***	−7.5	−5.8	−3.5	0.12*
	(37.4)	(27.6)	(6.2)	(4.0)	(4.5)	(0.07)
Observations	4882	4882	4882	4882	4882	4881

Sample is limited to schools that received an L, A, R, or E rating in 1999–2000, 2000–2001, 2001–2002 and 2003–2004. 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 by school district in parentheses.

* Significance at the 10% levels.

** Significance at the 5% levels.

*** Significance at the 1% levels.

gap year are not significantly different from schools that maintained their rating in the gap year. These results suggest, therefore, that the budget changes we observe in the year following the new rating information, as seen in Table 5, were temporary resources apparently 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.

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 school accountability ratings. If districts respond to accountability grades 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 research is based on a change in accountability systems in Texas. This is an advantage because the responses we examine are based on the actions of experienced administrators, and thus seem likely to be "permanent" behavioral changes based on experience. In particular, the new Texas accountability system offered a much more rigorous test than the old, but also offered a gap year where the new test was administered without grades being publicly released. This natural experiment offers an opportunity to examine how school districts respond to the "rating shock" of a new system, where the response by administrators is presumably based on reasonable expectations of actions by both politicians and parents.

Our results suggest that the implementation of a new accountability rating system provides a temporary shock to the education system, but in a very particular way. Specifically, the implementation of the new Texas rating system spurred school districts to allocate extra resources to schools which were formerly "above average," but which find their good ratings threatened by the new more rigorous system. In particular, schools whose ratings fall to the second lowest grade, "adequate," during the transition between the old and new systems saw an increase of their budget of 1.2%, most of which went to instruction. We also find that these budgetary changes seem to be temporary, and that within about 3 years the financial allocation system reverts to its former allocations.

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 1 year. Thus it may be that districts believe schools have sufficient incentive to improve, and do not need temporary resources. It is also possible that the "required improvement" rule of the accountability criteria made districts believe that their newly "L" schools would have an easy path to a higher rating and thus negated the need for large budget changes, even temporarily. Finally, we note from the means in Table 4 that expenditures for L schools are already significantly higher than for other schools, and districts may believe there are no further marginal returns to resources.

These results provide important new perspective to understanding the potential costs, and benefits, of school accountability ratings systems. First, our estimates show that school districts are willing to invest scarce resources into performance, which suggests accountability matters to central education administrations in some manner. We also find, however, that these investments are targeted to schools originally rated relatively highly. For those concerned that public schools are only oriented towards the bottom end of the quality distribution, this is interesting evidence that suggests a broader view.

On the other hand, we do not find that school districts necessarily re-orient their financial allocation system because of accountability. Specifically, we find that the financial investments that respond to a threat of a lower rating are temporary. While our work cannot go further into why, coupled with our other work (Craig et al., 2012) that finds no effect of changes in accountability ratings on marginal financial allocations, it indicates a possibility that annual changes in ratings have no value on the margin. This suggests that annual accountability ratings may be an over-investment, and that investments in testing might be undertaken on a less frequent basis.

Acknowledgments

The authors are grateful for helpful advice and comments from Aimee Chin, Meltem Daysal, Janet Kohlhase, Elaine Liu, Chinhui Juhn, Lori Taylor, Mary Beth Walker, as well as participants at the University of Houston/Rice University applied microeconomics seminar, the University of Florida, Georgia State University, and the meetings of the American Economic Association, Southern Economic Association, the American Education Finance Association, and two anonymous referees.

Appendix A

Table A1
Requirements for TAAS accountabiltiy 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					
1997–1998	40% or RI	40% or RI	n/a	6% or RI	94%
1998–1999	45%	45%	n/a	6%	94%
1999–2000	50%	50%	n/a	5.50%	94%
2000–2001	50%	50%	n/a	5%	–
2001–2002	55%	55%	50%	5%	–
B. Recognized					
1997–1998	80%	80%	n/a	3.50%	94%
1998–1999	80%	80%; 3–8, 10	n/a	3.50%	94%
1999–2000	80%	80%; 3–8, 10	n/a	3.00%	94%
2000–2001	80%	80%; 3–8, 10	n/a	2.50%	–
2001–2002	80%	80%; 3–8, 10	80%	2.50%	–
C. Exemplary					
1997–1998	90%	90%	n/a	1%	94%
1998–1999	90%	90%; 3–8, 10	n/a	1%	94%
1999–2000	90%	90%; 3–8, 10	n/a	1%	94%
2000–2001	90%	90%; 3–8, 10	n/a	1%	–
2001–2002	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 and 10% of the student body, or 200 students (prior to 2001)/50 students (2001 and later).

Table A2
Requirements for TAKS accountabiltiy 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								
2003–2004	35% or RI	50% or RI	50% or RI	50% or RI	25% or RI	50% or RI	1% or RI	75% or RI
2004–2005[a]	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
2005–2006	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
2006–2007	45% or RI	65% or RI	65% or RI	65% or RI	40% or RI	50% or RI	2% or RI	75% or RI

(continued on next page)

Table A2 (*continued*)

B. Recognized								
2003–2004	70% or RI	70% or RI	70% or RI	70% or RI	70% or RI	70% or RI	0.7% or RI	85%
2004–2005	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
2005–2006	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
2006–2007	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								
2003–2004	90%	90%	90%	90%	90%	90%	0.2%	95%
2004–2005	90%	90%	90%	90%	90%	90%	0.2%	95%
2005–2006	90%	90%	90%	90%	90%	90%	0.2%	95%
2006–2007	90%	90%	90%	90%	90%	90%	0.2%	95%

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 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.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jue.2012.07.002.

References

Bacolod, Marigee, DiNardo, John, Jacobson, Mireille, 2009. Beyond Incentives: Do Schools Use Accountability Rewards Productively? NBER Working Paper 14775.

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

Chakrabarti, Rajashri, 2007. Vouchers, Public School Response and the Role of Incentives: Evidence from Florida. Federal Reserve Bank of New York Staff Report no. 306 (October).

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).

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

Craig, Steven G., Imberman, Scott A., Perdue, Adam, 2012. Do School Administrators Believe Their Accountability Ratings? The Response of School Budgets to Accountability Grades. University of Houston Working Paper.

Cullen, Julie Berry, Reback, Randall, 2006. Tinkering Toward Accolades: School Gaming under a Performance Accountability System. NBER Working Paper #12286 (June).

Dee, Thomas, Jacob, Brian, 2009. The Impact of No Child Left Behind on Student Achievement. NBER Working Paper #15531.

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: Gronberg, Timothy (Ed.), Advances in Microeconomics. Elsevier.

Figlio, David, Kenny, Lawrenc., 2009. Public sector performance measurement and stakeholder support. Journal of Public Economics 93, 1069–1077.

Figlio, David, Lucas, Maurice E., 2004. What's in a Grade? School Report Cards and the Housing Market. American Economic Review, 94, June, 2004, pp. 591–604.

Figlio, David, Rouse, Cecelia Elana, 2006. Do accountability and voucher threats improve low-performing schools?”. Journal of Public Economics 90, 239–255.

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

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.

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

Jacob, Brian A., 2003. Getting Inside Accountability: Lessons from Chicago. Brookings–Wharton Papers on Urban Affairs, pp. 42–70.

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.

Neal, Derek, Schanzenbach, Diane Whitmore, 2010. Left behind by design: proficiency counts and test-based accountability. Review of Economics and Statistics 92, 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, Turner, Lesley J., 2010. Short-run impacts of accountability on school quality. American Economic Journal: Economic Policy 2, 119–147.

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