The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes

Helen F. Ladd *

Sanford Institute of Public Policy, Duke University, Durham, NC, 27708 USA

Received 2 November 1996; accepted 1 August 1997

Abstract

Consistent with the current emphasis on performance-based accountability in K-12 education, several states and a few local districts have introduced school-based incentive programs. This paper provides one of the few evaluations of the effects of such programs on student outcomes. Using a panel data set for schools in large Texas cities, it measures the gains in student performance in Dallas relative to those in other cities. It finds positive and relatively large effects for Hispanic and white seventh graders, but not for black students. Potentially positive effects also emerge for drop-out rates and principal turnover rates. [JEL I20] © 1998 Published by Elsevier Science Ltd. All rights reserved.

Keywords: Evolution; Accountability; Education policy; School reform

1. Introduction

Consistent with the current emphasis on performance-based accountability in K-12 schools, several states and a few local districts have introduced school-based performance incentive programs.¹ Such programs rank schools based on their success in increasing the achievement of their students, provide financial rewards for the schools at the top of the ranking, and may impose sanctions on the schools at the bottom of the ranking. Among these programs, that introduced by the Dallas Independent School District (DISD) in the fall of 1991 is particularly interesting. The sophistication of its methodology, the magnitude and nature of the rewards associated with the program, and the fact that proponents view the overall accountability program as the driver of a

more comprehensive set of reforms make the Dallas program worthy of detailed investigation. In a separate paper with Charles Clotfelter, I describe and evaluate the design aspects of the program in the broader context of recognition and reward programs for schools (Clotfelter and Ladd, 1996). This paper focuses on the measurable impacts of the Dallas program on student outcomes, such as test scores and drop-out rates, and on one measure of internal change, the turnover of school principals.²

Given the widespread interest in school-based recognition and reward programs, it is surprising how little evaluation has been done of their impacts. In some cases, the reason is clear. Consider, for example, the state of South Carolina which has had a relatively sophisticated school incentive program since 1984. Given that the incentive program is only a small part of a much larger reform initiative, it would be difficult, if not impossible, to isolate the impacts of the incentive program. For purposes of evaluation, a major advantage of the Dallas pro-

* Tel: + 1-919-613-7352; E-mail: hladd@pps.duke.edu

¹ According to the Southern Regional Education Board, *Career Ladder Clearing House*, April 1994, and updated by telephone conversation, the states with school performance incentive programs include South Carolina, Georgia, Indiana, Kentucky, Texas, North Carolina, and Tennessee (Cornett and Gaines, 1994).

² The Clotfelter and Ladd paper reports some summary measures of program impact. This paper extends the analysis of program impacts by adding an additional year of data, presenting more detail, and reporting results about principal turnover.

gram was that the incentive program was a central part of the accountability program, which, in turn, was the centerpiece of the Dallas reform effort. Of course, from the perspective of education reform, Dallas's focus on accountability to the exclusion of other components of reform, such as curriculum and professional development, could be problematic. I return to some of these considerations below.

2. Nature of the program and its incentives

The Dallas program emerged from the work of a special commission appointed by the Dallas School Board in 1990 to design a reform proposal for the district. The Commission, which had strong involvement and support from the local business community, recommended a sophisticated accountability system "to hold all levels of the district accountable for results and thus to stimulate continuous quality improvement over time."³ As part of this accountability system, financial awards were to be given to the personnel in the most effective schools, with half of the \$2.5 million cost of the awards initially to be funded by the local business community. The program began in the 1991–92 school year and continued at the same level of funding through 1995, the final year of this study.⁴

The Dallas program stands out from other school-based incentive programs in the sophistication of its methodology for ranking schools. Starting with data on test scores of individual students in sequential years, the District used multiple regression analysis to rank each school on the magnitude of the contribution the school made to the learning of its students during the year, where the contributions are calculated relative to the state average. For this purpose, Dallas used a two-stage regression procedure. In the first stage, individual test scores for both the current and the prior year were estimated as a function of various student characteristics including race, gender, and eligibility for free and reduced price lunch. This procedure generated a set of residuals that can be viewed as test scores purged of the effects of socio-economic status and race. In the second stage, the residuals from the regressions for the current year were regressed on the residuals for the prior year. The residuals at this stage provide estimates of "gains"

in student learning. These "gains" are then aggregated to the school level by calculating the mean across all subjects for all students enrolled in the school during the year.⁵

By fully adjusting individual test scores for the socioeconomic status of the students and their race, Dallas tried to assure that the estimated value-added measures were not biased toward schools serving more advantaged children and that all schools would have an equal ability to win an award. Concerned about corruptibility of the process in the sense that teachers might teach narrowly to specific tests or might try to influence the composition of students taking the test, Dallas relied on multiple measures of student outcomes, including two tests given annually, a criterion-referenced state test tied to the state curriculum, the Texas Assessment of Academic Skills (TAAS) and a nationally normed test, the Iowa Test of Basic Skills (ITBS). In addition, Dallas supplemented these tests with a variety of other end-of-course tests and school-wide measures such as student attendance and drop-out rates.⁶

In each of the winning schools, which typically represented about 20 percent of the schools during the first three years of the program, financial bonuses were given to all members of the school's staff. Principals and teachers received bonuses of \$1000 and other staff such as janitors and secretaries received bonuses of \$500. In addition, \$2000 was given to the school's activity account.⁷ Thus, regardless of the quality of their individual contributions to the common goal, all teachers in a winning school received awards. Although the program could potentially generate free-rider problems, this group-based incentive scheme had a major potential advantage over a program of merit-based pay for individual teachers in that it would encourage the staff of the school to work together toward the common goal of student achievement.

⁵ The ranking methodology is described in Clotfelter and Ladd (1996). As reported there, the Dallas methodology appears to generate no clear bias among schools with different average test scores or demographic characteristics.

⁶ The test score gains are supplemented by student attendance and promotion rates in elementary schools; those variables plus drop-out rates and enrollment in accelerated courses in middle schools; and those plus SAT and PSAT average scores and percent taking the tests in the high schools and end-of-course exams in high schools. In all cases, the measurement of student and school gains are measured relative to the average rather than to some absolute standard.

⁷ The reward structure was altered for the 1994–95 school year to assure more winners. For that year, approximately the top 20 percent of the schools received the awards mentioned in the text and a second tier of schools were given smaller awards. Under this new structure, about half the schools would receive awards.

³ Commission for Educational Excellence, 1991, *Final Report*, June 18, 1991, p. 3.

⁴ A new school board with a new chair has since continued the program, but cut the budget to \$1.2 million. There is some indication that the Board is interested in restoring funds for the incentive payments if they are available in the future. In any case it appears as if the Board continues to support the ranking of schools and will continue with them even if money is not available for incentive payments.

Proponents of the Dallas incentive program praised its focus on the school as the unit of accountability, emphasized that the measures of school effectiveness were not biased toward schools serving children from advantaged backgrounds, and that the use of multiple outcome measures minimized the incentives for teachers to teach narrowly to a single test. Opponents pointed out that the ranking system was so complicated that principals and teachers did not fully understand how the system worked, that rankings were based on how schools did relative to one another rather than relative to an absolute standard of student achievement, that school officials could manipulate the system by keeping some children from being tested or by outright cheating, and that implementing this incentive program in the absence of more fundamental educational reforms was problematic.

This final criticism deserves further discussion. Undoubtedly, the Dallas Independent School District (DISD) had and continues to have some major problems. Like many large urban school districts, it was highly centralized and bureaucratic, and the academic performance of the students was abysmal. The issue was how to change the system. One strategy could have been to introduce a coherent set of reforms consistent with the goal of achieving high outcomes for all children. Accountability and incentives might well have been part of such a strategy, but they would have followed improvement in curriculum and professional development and the decentralization of governance to the school level.⁸ Alternatively, the DISD might have worked more closely with teachers and principals to develop a bottom-up reform strategy. Instead, the DISD chose to use its accountability and incentive system as a top-down catalyst for other changes. In the short run, it was intended to increase student performance simply by getting school staff to focus their attention on student learning. Within a longer time frame, DISD officials hoped to encourage teachers to demand better development opportunities, to provide information to help administrators identify ineffective schools, and to make it politically more feasible for the school board to give more authority to the schools, a change that many education reformers would support both for its own sake and to increase the benefits from the accountability program.

Has the Dallas strategy been effective? My approach in this paper is to judge it on its own terms, that is, in terms of student outcomes. Specifically, I focus primarily on the impacts of the program on student test scores during the first four years of the program. If the program had been effective, one might expect to see some immediate short run improvement in test scores as teachers focused more attention on academic achievement and test-taking skills. Of greater interest is whether the program induced a sufficient number of other changes to generate more permanent gains in learning over time. Should such gains in learning emerge, it would then be desirable to try to identify the mechanisms through which they occurred. For example, was the leadership of the schools improved, were teachers replaced, did teachers receive better professional development, or did schools get more parents involved in the schools? Of these latter questions about process or inputs, the only one for which I have any partial data relates to school leadership. Looking at the other mechanisms would require a very different study design. Hence, I focus here almost exclusively on student outcomes, an appropriate first step given that accountability and incentive programs are specifically designed to improve such outcomes.

3. Measuring impacts

To measure the impacts of the Dallas program on student outcomes, it would be helpful to have a consistent set of test scores over time. Unfortunately, Dallas's testing arrangements are dictated largely by state decisions, and the state altered its testing program, especially its use of nationally normed tests, in significant ways during the early 1990s.⁹ Consequently, test results on nationally normed tests in Dallas for each year since the introduction of the Dallas reform are not directly comparable to test results for any year prior to its introduction. In an internal study comparing Dallas test scores on the Iowa Test of Basic Skills in 1994 to scores in 1991, the authors had to make some rather arbitrary adjustments to account for the fact that Dallas used the long form of

⁸ To that end, the District would have had to supplement the state's basic curriculum to make it more oriented toward higher-order thinking skills, to introduce assessment systems consistent with the new curriculum, to provide for professional development for the district's teachers consistent with the higher expectations for children, to give school officials significantly more authority over how to use the resources within a school, and only then to set up a system for holding schools accountable for the academic performance of their students. This type of reform is commonly referred to as systemic reform. (See O'Day and Smith, 1993.)

⁹ Through 1990–91, Texas used the full Iowa Test of Basic Skills and the Tests of Academic Proficiency (ITBS/TAP). For 1991–93, it used the Norm-Referenced Assessment Program for Texas (NAPT), which by 1993 was a bit of a hybrid between a nationally normed test and a test referenced to the state curriculum. Distressed by the inappropriate norming of the NAPT, DISD requested a waiver from the NAPT in 1993–94, and replaced it with a short form of the ITBS/TAP. Although the short form was normed to the full ITBS, it includes a higher proportion of more challenging questions. (See Anderson et al., 1994.)

the test in 1991 and the short form in 1994.¹⁰ The results were at best mixed. For reading, the authors concluded that DISD held its own relative to the nation, with gains in some of the higher grades and losses in the lower grades; for math, DISD gained relative to the nation in most grades except grades 7 and 9. Hispanics generally gained relative to national norm groups but blacks lost. However, how the district would have done relative to national norms in the absence of the program is unclear.

I focus instead on student performance on the Texas Assessment of Academic Skills (TAAS), a test that is linked to the state's curriculum and that serves as the basis for statewide accountability. The link with the state curriculum provided teachers a strong incentive to incorporate the objectives of the TAAS, and therefore the objectives of the state curriculum, into their teaching. Given that student performance on the TAAS was one of the outcome variables included in the Dallas accountability system, it is reasonable to expect that a successful Dallas program would lead to better student performance on the TAAS.

My basic outcome measure is the pass rate on the reading and math parts of the test. One disadvantage of this measure is that it provides more information on schools with low-performing students than on those with high-performing students. In the latter, any success by the schools in increasing the test scores of students who would otherwise have passed the test is not reflected in the results. Nonetheless, pass rates are an acceptable outcome measure for the purposes of measuring the impacts of the Dallas reform program especially since the pass rates are relatively low. Because the nature of the tests, and statewide pass rates, vary somewhat from year to year, I express all pass rates relative to the state average for that year.

Fig. 1 shows the average pattern of third grade and seventh grade pass rates in Dallas for the period 1991 to 1995. The 1990–91 school year represents the year before the Dallas program was introduced and 1994–95 the fourth year of the program. These averages are simple unweighted averages across all schools in the district. Note that in all cases the pass rates in Dallas are below the state average, an outcome that is typical for urban school districts. The figures show that while the third grade pass rates may have increased slightly during 1992 and 1993, they fell relative to the average in 1994 and 1995. Over time, Dallas seventh graders apparently improved relative to the state average in reading, but not in math.

However, the relevant analytical question is not simply whether student performance on the TAAS improved in Dallas during the 1990s but rather whether it improved

relative to what would have been expected in the absence of the program. To address this question, I use panel data techniques to compare average student performance by school in specific grades in Dallas on the state administered TAAS with student performance by school in the other five large Texas cities (Austin, El Paso, Fort Worth, San Antonio, and Houston) for the school years 1990–91 to 1994–1995.¹¹

Although the six Texas cities differ in terms of size and the racial makeup of the schools, the schools are comparable to those in Dallas in that they are all subject to the same statewide curriculum, the same statewide testing requirements, and the same statewide reform and accountability efforts. In addition, all six cities are minority Anglo and have large proportions of children from disadvantaged families. City and school-specific differences can be controlled for statistically.¹²

A thorny issue arises about the counterfactual against which the effects of the Dallas accountability system should be measured. One possible counterfactual is that had Dallas not introduced its specific incentive-based reform it would have implemented a reform effort comparable to the average of the local reform efforts under-

¹¹ I would prefer to focus on changes in performance for specific cohorts of students. Unfortunately the data for that type of study design are not available. Although I have information on gains by cohorts after the introduction of the program, we do not have comparable information on gains by similar cohorts prior to the program. Hence, no direct inferences about program impacts can be made from the post program changes in test scores by cohort.

¹² Although the Dallas Independent School District (DISD) is virtually coterminous with the city of Dallas, cities and school districts are typically not coterminous in the other big Texas cities. To cover as much of each city as possible, we have combined the major school districts in those cities that are served by more than one district. Austin ISD, Dallas ISD, and Fort Worth ISD were used for Austin, Dallas, and Fort Worth, respectively. For Houston, we combined the Houston ISD, Alief ISD, and Spring Branch ISD. For El Paso, we combined El Paso ISD, Canutillo ISD, and Ysleta ISD. For San Antonio, we combined San Antonio ISD, Edgewood ISD, Harlandale ISD, South San Antonio ISD, and Alamo Heights ISD. The 1992 population of the six cities ranged from 1.7 million in Houston to 454 000 in Fort Worth. Dallas is the second largest city with 1 million people. Dallas has the highest proportion of black students with 44.4 percent, and is followed by Fort Worth with 34.0 percent and Houston with 31.7 percent. Dallas has the fourth highest percentage of Hispanic students. San Antonio has 83.8 percent Hispanic students, El Paso 78.6 percent, Houston 45.0 percent and Dallas 39.7 percent. With respect to economically disadvantaged students, Dallas ranks second with 71.2 percent to San Antonio's 82.8 percent. The source of the population figures is Census Bureau Press Release CB94-15 and Population Listing PPL-2. All other figures are from 1993–94 and originate from the MicroAEIS (Academic Excellence Indicator System) computer data base and the AEIS.

¹⁰ See Anderson et al., 1994, p. 5 for the adjustment and throughout for the results.

Fig. 1. Pass rates in Dallas schools (relative to state average).

taken by the other large Texas cities during the period. An alternative counterfactual is that Dallas would have implemented no local reform effort. The difference between the two counterfactuals in practice depends on the extent to which the other large Texas cities engaged in significant school reform during the study period.

Telephone inquiries indicated that Fort Worth started a reform initiative in 1989 that engaged the local corporations and the community in a program to assure that students were prepared for the work place. In 1992–93, Houston introduced a multi-year school improvement plan under which schools submitted goals and implementation plans and received school report cards from the district indicating their performance relative to their goals. Finally, while El Paso ISD had no special reform initiative, Ysleta ISD, which serves 50 000 students in El Paso, had a school incentive program that was similar in spirit to that of Dallas in that it provided awards to high-performing schools.

In the basic model discussed below, student performance in Dallas is evaluated relative to the performance of students in all the other large Texas cities. To the extent that the other reform efforts generated gains in student performance, students in Dallas would have had to do even better for the Dallas program to generate stat-

istically significant positive impacts. Hence, this model represents a relatively stringent standard for evaluation. An extended model changes the basic comparison to schools in districts not engaged in local reform efforts and also permits comparisons between Dallas schools and schools in each of the other three reforming areas.

For reasons I explain below, the seventh grade data are more amenable to the adjustments needed to make the data comparable over time than are the third grade data. Hence I begin by examining the impact of the program on seventh grade pass rates and look only briefly at third grade pass rates. I then examine trends in school attendance and drop-out rates, and conclude with an analysis of turnover rates of principals. The third grade results are not encouraging. However, those for the seventh grade are more positive, at least for some subgroups of the population. In addition, drop-out rates appear to be falling relative to those in other Texas cities, and principal turnover rates are higher than in the past, a change that I view as positive given the low turnover rates of the pre-reform period.

4. Impacts on seventh grade pass rates

The panel analysis focuses on pass rates in two subjects, reading and math. For each subject, I look at pass rates for all seventh graders by school, and for the following subgroups: children from economically disadvantaged households, blacks, Hispanics, and whites. Unfortunately, the TAAS, which was administered during the fall for the first three years of the sample period, was administered in the spring during the final two years. Hence, for the first three years, the seventh grade tests essentially covered material through the sixth grade. The switch to spring administration in 1994 means that for the most recent two years, the test covered seventh grade material. To make the coverage as comparable as possible, I use the seventh grade test results for the first three years and the sixth grade results for the last two years. However, because of the long summer vacation, a student taking a test on sixth grade material in the spring may do better than one who takes it in the fall, and this differential is probably greatest for students from disadvantaged households. To control for the fact that the tests are not fully equivalent, I include several terms interacting the years 1994 and 1995 with various school characteristics.

In the following basic model, the subscript i denotes the school and t the year. For ease of exposition, subscripts for the subject or subgroup are excluded:

$$\text{PASS}_{it} = \alpha + \beta_1(\text{YR92}_i * \text{Dallas}) + \beta_2(\text{YR93}_i * \text{Dallas}) + \beta_3(\text{YR94}_i * \text{Dallas}) + \beta_4(\text{YR95}_i * \text{Dallas}) + \sum_k X_{kit}[\gamma_k + \delta_k(\text{YR94 or YR95})] + \sum_m \epsilon_m + \sum_r \eta_r + \mu_{it}$$

where

- PASS_{it} = seventh (or sixth) grade pass rate in school i in year t (for a specific subject and subgroup), expressed relative to the statewide average pass rate for that subject in that year.
- YR92_i = an indicator variable that takes on the value 1 if year = 1992 and 0 otherwise in year t , where 1992 represents the 1991–92 school year.
- YR93_i , YR94_i , and YR95_i are similarly defined.
- Dallas = an indicator variable that takes on the value 1 if the school is in Dallas and 0 otherwise.
- X_{kit} = k th characteristic of school i in year t . Characteristics include the percent of the students in the school who are black, the percent Hispanic, the percent with limited English proficiency, and the percent economically disadvantaged and the mobility rate of students.
- ϵ_m = year effects: 1992, 1993, 1994, 1995. 1991 is the base year.

η_r = city effects: Dallas, Austin, El Paso, Fort Worth, and San Antonio. Houston is the left out city.

$\beta_1, \beta_2, \beta_3, \beta_4, \gamma_k$ (for $k = 1, \dots, 5$), δ_m (for $m = 1992, 1993, 1994, 1995$) are parameters to be estimated.

μ_{it} is a random error term that varies by school and by year.

The crucial variables for this analysis are the four Dallas-year interaction terms, that is, the first four variables in the equation. Given that the Dallas program was initiated during the 1991–92 school year, and the TAAS was given during the fall of that year, the 1992 interaction term should be viewed as prior to the introduction of the program. However, a small positive impact might be predicted given the publicity associated with the reform effort during the previous spring and summer after the Commission recommendations were released and accepted. The other three interaction terms measure the cumulative impacts of the program in each year relative to the average change in pass rates in the other large Texas cities compared to 1991. Positive coefficients on these variables would indicate that the Dallas program was successful in raising seventh grade test scores relative to 1991.

The other variables serve as control variables. The school characteristic variables (the X s) control for differences among schools and over time in the composition of the student body and include variables such as the percentage of children who are from disadvantaged families. The city effects control for any differences in the average performance on the TAAS across cities. As a result, the observation that Dallas schools typically may have either lower or higher pass rates than schools in the other cities does not affect the measure of program impact. The year effects control for all the factors that affect average pass rates from year to year, such as the change in the passing standard in 1994. Since the dependent variable is expressed relative to the state average, the year effects are expected to be small.

The school characteristics are interacted with the 1994 and 1995 year indicator variables (consolidated into a single variable) to control for the fact that the 1994 and 1995 tests were given to sixth graders in the spring rather than to seventh graders in the fall and to allow for different effects by socioeconomic group.¹³

Given the inclusion of the city and year effects, it is reasonable to assume that the error term is uncorrelated with the explanatory variables and that it has a mean of 0. However, the variation in school sizes across the sample makes it unreasonable to assume that the error terms

¹³ We experimented with other interactions of this form in the estimated equations and none of the terms were statistically significant.

are drawn from distributions with the same variance across schools. Consequently, all equations reported below were weighted either by the number of students in the school or by the number of students in the school in the particular subgroup.¹⁴

4.1. Basic results and interpretations

Table 1 reports the basic results in reading and math for all students aggregated by school. The first two columns report the basic model in which the changes in Dallas schools are compared to the changes in schools in all the other big Texas cities. The final two columns represent results from the extended model which tests whether students in the Dallas schools improved their performance relative to students in the large Texas cities that were not undertaking their own local reform initiatives.

The results of primary interest are in the first four rows of Table 1. In the basic model, the coefficients of the four key interaction terms for reading range from 9 to 12 percent of the state average for all four years and for math range from 12 to 17 percent. In other words, relative to the 1991 base year, pass rates rose more in Dallas than they did in the other large Texas cities. However, the interpretation of these coefficients is complicated by the unexpected finding that the 1992 coefficient for reading is approximately the same size as the subsequent coefficients and for math is higher. Because the 1991–92 TAAS was given in the fall of 1991, the pass rates for 1992 represent gains that occurred before the Dallas accountability program could have had much impact on student learning.¹⁵

The explanation for the positive 1992 coefficients is unclear. They could reflect the phenomenon of regression to the mean or, alternatively, the effects of positive publicity about the new reform program with

its emphasis on test scores.¹⁶ To the extent that the first explanation is valid, the 1992 estimates represent the relative growth that would have occurred in Dallas compared to 1991 in the absence of the accountability program. The fact that the 1993, 1994, and 1995 estimates are similar to or lower than the 1992 estimates would then imply that by 1995 the accountability program exerted no positive impact on the average performance of students in Dallas. However, to the extent that the 1992 results reflect positive publicity and the new focus on student outcomes, the interpretation would be more positive. Given that one would predict that any “Hawthorne” effects—that is, effects caused by change itself—would disappear over time, the finding of a relatively constant effect over time for reading, is encouraging. That is, with respect to reading, the results could be interpreted as follows: not only did the Dallas schools tighten up their operation to the benefit of the students in 1992 at the beginning of the program, but they apparently were able to maintain whatever gains in student test performance achieved that first year over the next few years despite the declining novelty of the new program. Note, however, that in a truly effective program, one would expect the gains to increase over time as seventh grade students have more years of schooling under the program.

Consistent with the view that reforms in some of the other Texas cities raised student performance, the results from the extended model provide a somewhat more positive picture of the effects of the Dallas program. Once again, the 1992 coefficients are positive and statistically significant, but, in this case, the evidence suggests a small upward trend over time in both reading and math. For reading, the estimates increase from 0.103 in 1992 to 0.150 in 1995. For math, the estimates increase from 0.135 to 0.168. In neither case, however, does the 1995 estimate differ significantly from the 1992 estimate. Thus, while the upward trend relative to 1992 could reflect chance alone, the evidence is at least consistent with a positive program impact relative to both 1991 and 1992.

Comparing the effects of the Dallas reform with those undertaken by Fort Worth, Houston, and Ysleta (see the next set of entries in columns 3 and 4) shows that the Dallas coefficients exceed the comparable coefficients for Fort Worth and Houston. Only Ysleta (the small dis-

¹⁴ Ideally, the weights should reflect the number of students tested in each grade rather than the number of students in each school. However, we do not have data on the number of students in each grade. In general, school size would be correlated with grade size except for those elementary schools that have fewer grades than the average elementary school. A case can also be made that weighting is undesirable on the ground that within school correlation offsets the effect of school size on the variance of the error term. In any case, unweighted equations yield results that are similar to the weighted results reported here. See Dickens (1990).

¹⁵ If pass rates for the seventh grade rather than the sixth grade tests are used for the final two years of the analysis, the coefficients on the four key interaction terms for reading are almost identical to those reported in Table 1. For math, the 1994 and 1995 coefficients are 0.115 and 0.083 rather than 0.121 as reported in Table 1.

¹⁶ Officials in Dallas emphasize the role of publicity given that the reform initiative was discussed publicly during the spring of 1991 and adopted during the summer of 1991. Hence, the 1992 coefficient may well represent a program effect. However, given the timing, it is more likely to reflect the new focus on testing rather than significant new learning. With this interpretation of the 1992 coefficient, rather than a regression-to-the-mean interpretation, the estimated coefficients for 1993 and 1994 may well measure true program effects.

Table 1

Seventh grade reading and math—basic and extended models^a (absolute value of *t*-statistics in parentheses)

	Basic model		Extended model	
	Reading (1)	Math (2)	Reading (3)	Math (4)
<i>Dallas interaction</i>				
YR92	0.096* (1.83)	0.154** (2.30)	0.103* (1.91)	0.135** (1.96)
YR93	0.117** (2.25)	0.168** (2.54)	0.118** (2.21)	0.153** (2.24)
YR94	0.090** (2.10)	0.121** (2.20)	0.134** (2.89)	0.141** (2.37)
YR95	0.086** (1.99)	0.121** (2.20)	0.150** (3.25)	0.168** (2.82)
<i>Fort Worth interaction</i>				
YR92	—	—	0.053 (0.91)	— 0.018 (0.24)
YR93	—	—	0.007 (0.11)	0.032 (0.43)
YR94	—	—	0.046 (0.79)	— 0.001 (0.01)
YR95	—	—	0.025 (0.43)	0.040 (0.53)
<i>Houston interaction</i>				
YR92	—	—	— 0.039 (1.15)	— 0.060 (1.36)
YR93	—	—	— 0.043 (1.26)	— 0.079* (1.80)
YR94	—	—	0.054* (1.72)	0.017 (0.44)
YR95	—	—	0.097** (3.14)	0.058 (1.47)
<i>Ysleta^b interaction</i>				
YR92	—	—	0.002 (0.03)	— 0.130* (1.77)
YR93	—	—	0.016 (0.28)	— 0.065 (0.88)
YR94	—	—	0.120** (3.53)	0.105** (2.42)
YR95	—	—	0.171** (4.98)	0.168** (3.80)
<i>School characteristics</i>				
Percent Black	— 0.009** (16.47)	— 0.009** (13.54)	— 0.008** (12.99)	— 0.008** (10.47)
Percent Hispanic	— 0.009** (13.53)	— 0.009** (10.81)	— 0.009** (11.52)	— 0.009** (8.84)
Percent LEP	— 0.002* (1.71)	— 0.001 (0.97)	— 0.001 (0.60)	— 0.001 (0.39)
Percent disadvantaged	— 0.002** (3.41)	— 0.002* (1.94)	— 0.003** (4.41)	— 0.003** (2.61)
Mobility rate	— 0.014** (11.37)	— 0.012** (8.00)	— 0.015** (11.57)	— 0.014** (8.62)
<i>Year indicator variables</i>				
YR92	0.003 (0.19)	— 0.032 (1.36)	0.009 (0.39)	— 0.003 (0.09)
YR93	— 0.024 (1.29)	— 0.036 (1.54)	— 0.012 (1.52)	— 0.011 (0.37)

Continued

Table 1
Continued

	Basic model		Extended model	
YR94	– 0.529** (11.38)	– 0.285** (4.82)	– 0.536** (10.99)	– 0.288** (4.60)
YR95	– 0.494** (10.61)	– 0.299** (5.04)	– 0.524** (10.70)	– 0.329** (5.25)
<i>City Indicator Variables</i>				
Dallas	– 0.082** (2.20)	– 0.026** (0.55)	– 0.088** (2.32)	– 0.037** (0.76)
Austin	– 0.029 (1.41)	– 0.096** (3.71)	– 0.003 (0.11)	– 0.089** (2.92)
El Paso	0.069** (3.89)	0.068** (3.02)	0.055** (2.15)	0.040 (1.23)
Fort Worth	– 0.038* (1.88)	– 0.077** (3.01)	– 0.042 (1.00)	– 0.085 (1.57)
San Antonio	0.030 (1.20)	– 0.012 (0.36)	0.066** (2.18)	– 0.004 (0.10)
<i>1994 and 1995 school year and school characteristics interactions</i>				
YR945*Percent Black	0.004** (6.22)	0.002** (2.20)	0.003** (3.28)	0.000 (0.46)
YR945*Percent Hispanic	0.004** (4.33)	0.001 (1.12)	0.002** (2.49)	0.000 (0.17)
YR945*Percent LEP	0.003** (2.72)	0.003** (2.15)	0.002 (1.45)	0.002 (1.38)
YR945*Percent disadvantage	0.002** (2.29)	0.002 (1.50)	0.003** (3.45)	0.003** (2.19)
YR945*Mobility rate	0.008** (5.44)	0.004** (2.17)	0.009** (6.20)	0.006** (3.25)
Constant	1.954** (52.77)	1.924** (41.17)	1.942** (51.06)	1.918** (39.66)
Number of observations	1122	1118	1122	1118
Adjusted R^2	0.67	0.57	0.68	0.58

Note: ^a In both models, the dependent variable is the relative pass rate on the TAAS and observations are weighted by pupils in the school. * Indicates significance at the 10 percent level, and ** indicates significance at the five percent level. ^bthe Ysleta School District serves as part of El Paso.

trict that is part of El Paso) exhibits effects that are potentially as strong or stronger than those in Dallas. Of interest is that Ysleta's reform is similar to that in Dallas in that it focuses on holding schools accountable for student performance. Thus, this evidence suggests that the Dallas reform effort has raised student performance relative to that in the nonreforming districts taken as a group and also relative to Fort Worth and Houston.

Most of the control variables exhibit the expected effects. The school characteristic variables enter with negative and, in most cases, statistically significant coefficients. In other words, the relative pass rate for all students in a grade is lower the higher is the percent of students who are black, the percent Hispanic, the percent of students with limited English proficiency, and the percent who are economically disadvantaged. In addition, the mobility rate matters in the sense that schools with more students moving in and out during the year have

lower pass rates than schools with more stable student populations.¹⁷

The 1992 and 1993 year coefficients provide straightforward estimates of the average difference between relative pass rates in each year relative to the base year of 1991. These coefficients are typically small and insignificant.

¹⁷ In contrast to the other school characteristics, data for the mobility rates are available for only two years, 1993–94 and 1994–95. Hence the coefficient reflects only the effects of variation among schools and not the effects of changes over time. Note that in contrast to the methodology used in the Dallas program, the test results used in these regressions include all children tested, not just those who were enrolled in the school throughout the year. This observation implies that the negative coefficient measures both the adverse effects of mobility on the children who are not continuously enrolled in the school and also the effects on the other students.

nificant. Care must be taken not to overinterpret the negative 1994 and 1995 year coefficients given that those years are also interacted with the socio-economic variables. The city effects, which represent the differences across cities indicate that the base year pass rate in Dallas in reading was lower than that in the base city, Houston.

4.2. Program impacts for subgroups of the population

Tables 2 and 3 report results for subgroups of the population for reading and math, respectively.¹⁸ In all cases, the reported coefficients come from complete models of the form reported in Table 1. The first panel in each table provides the results for the basic model (with the comparison being schools in all the other large Texas districts) and the second for the extended model (with the comparison being schools in districts without major local reforms).

The tables indicate that the results differ quite dramatically across subgroups, with the most positive effects emerging for Hispanics and whites. Specifically, in reading, the pass rates (relative to the state average) of Hispanics rise relative to those for Hispanics in the other Texas cities from 9 to 16 percentage points in 1994 and then down to 13 percentage points in 1995 (panel 1) and relative to cities without reform efforts by 12 to 20 percentage points (panel 2). For white students, the patterns of gains over time were similar to those for Hispanics, rising from 13 to 25 percentage points over time in the basic model and from 12 to 27 percentage points in the extended model.¹⁹ In contrast, there is no evidence that the Dallas accountability had any effect at all on the performance of black students in Dallas. No clear pattern emerges for the category of students from disadvantaged households, which overlaps in complicated ways with the other categories.

Similar patterns emerge in math (see Table 3), with the Hispanics and the white students both making gains relative to the prereform, 1991 school year and relative

to the 1992 school year. Once again no positive effects emerge for black children, and the results are mixed for students from disadvantaged households.

4.3. Extensions—Resources and state policy

The basic model includes no information on the amount of resources available to the school, such as the number of teachers per pupil and the percent of teachers with more than five years experience. To the extent that these resources were changed at a differential rate in the Dallas schools relative to schools in the other cities, the absence of the resource variables from the equation could conceivably bias the measured impacts of the Dallas program. The argument for excluding them from the basic model is twofold. First, the city indicator variables would capture any systematic differences in resources available per student across cities and the year variables would capture systematic differences in all urban districts over time. Second, changes in the amount of resources available to schools in Dallas could be viewed as endogenous. To the extent that the presence of the accountability system generated pressure for a different set of resource decisions than would have occurred in the absence of the accountability system, the resource changes are affected by the accountability system and should not be treated as exogenous. Two examples of this endogeneity may suffice. First, in addition to seeking new leadership and requiring a better school improvement plan, the central office in some cases provided additional resources to schools that, according to the school rankings, were performing poorly. Second, the pressure for accountability throughout the system may have made it possible to move some resources away from central administration down to the schools.²⁰ Hence, even without any overall increase in resources in Dallas, the schools would have received more resources.

Nonetheless, to rule out the possibility that this endogeneity argument is overstated and that our failure to treat resources as an exogenous determinant of pass rates biases the estimates of program impacts, I reran the basic regressions with two additional variables, the ratio of pupils per teacher in school i in year t , and the percentage of teachers in the school with 5 or more years experience (equations not reported). Except for the subgroup of Hispanics in math, none of the teacher variables is statistically significant. For that subgroup, greater teacher experience generates positive impacts on test results. None of the coefficients for pupils per teacher are statisti-

¹⁸ The subsample regressions are based only on the schools for which data on the subgroup are available. For example, schools with no or few blacks are excluded from the sample of schools used to analyze the impact of the program on black students. Thus, the maximum number of schools for each of the subgroups is the full set of 1122 schools for reading and 1118 for math.

¹⁹ One reader of this paper familiar with Dallas wondered whether recent white flight from Dallas could account for the apparent gains for white students. Such a case is difficult to make since the white parents fleeing the Dallas public school system are likely to be those with the higher performing students. Hence, white flight would likely lead to lower average scores for white children over time, not the relative gains observed here.

²⁰ Based on interview with Chad Woolery, superintendent of the Dallas Independent School District, January, 1995.

Table 2

Program impacts for seventh grade by subgroups—reading^a (absolute value of *t*-statistics in parentheses)

	All	Economically disadvantaged	Black	Hispanic	White
<i>Basic model^b</i>					
YR92*Dallas	0.096* (1.83)	0.027 (0.45)	0.102* (1.64)	0.094 (1.37)	0.130 (1.37)
YR93*Dallas	0.117** (2.25)	0.111* (1.84)	0.093 (1.51)	0.110* (1.64)	0.141 (1.48)
YR94*Dallas	0.090** (2.10)	0.035 (0.68)	– 0.069 (1.35)	0.161** (2.85)	0.262** (3.20)
YR95*Dallas	0.086** (1.99)	0.030 (0.59)	– 0.051 (1.00)	0.130** (2.30)	0.246** (2.97)
Number of observations	1122	1091	791	970	644
Adjusted <i>R</i> ²	0.67	0.45	0.44	0.54	0.45
<i>Extended model^c</i>					
YR92*Dallas	0.103* (1.91)	0.057 (0.91)	0.075 (1.03)	0.119* (1.73)	0.123 (1.28)
YR93*Dallas	0.118** (2.21)	0.133** (2.15)	0.045 (0.63)	0.131* (1.94)	0.118 (1.21)
YR94*Dallas	0.134** (2.89)	0.082 (1.50)	0.000 (0.01)	0.203** (3.46)	0.268** (3.17)
YR95*Dallas	0.150** (3.25)	0.100* (1.83)	0.003 (0.05)	0.197** (3.36)	0.273** (3.18)
Number of observations	1122	1091	791	970	644
Adjusted <i>R</i> ²	0.68	0.47	0.44	0.56	0.45

Notes: ^a In both models, the dependent variable is the relative pass rate on the TAAS and observations are weighted by pupils in the school. * Indicates significance at the 10 percent level, and ** indicates significance at the five percent level. ^b Additional variables in the basic model are indicator variables for school years 1992, 1993, 1994 and 1995 (1991 = base); school characteristics (percent black, percent Hispanic, percent LEP, percent disadvantaged, and mobility rate); interactions between 1994 and 1995 school years and the school characteristics; and indicator variables for the cities of Dallas, Austin, El Paso, Fort Worth and San Antonio (Houston = base). ^c The independent variables in the extended model include all of the variables in the basic model plus interactions between the school years and district indicator variables to represent the districts (Houston, Fort Worth and Ysleta) that also have reform programs currently in progress.

cally significant.²¹ Although the inclusion of the resource variables dampens somewhat the measured impacts of the Dallas program, it does not change the basic picture. The patterns by subject and subgroup remain very similar to those reported in Tables 2 and 3.

Another possible complication is that during the study period the state of Texas became more aggressive in identifying low-performing schools and trying to remedy

their problems. To the extent that the state's program was successful, it would be difficult for schools in this low-performing category in Dallas to outperform comparable schools in other Texas cities.

One way to rule out this possibility is to eliminate from the sample the schools that the state might have judged as low-performing based on its criterion of low pass rates on the TAAS. For statistical reasons, it is undesirable to truncate the sample based on the value of the dependent variable. Instead, schools were eliminated on the basis of one of the school-specific explanatory variables, the percent of students from economically disadvantaged households. In particular, I eliminated the 25 percent of the schools with the highest proportion of disadvantaged students on the grounds that high disadvantage is likely to be associated with low pass rates on the TAAS. Once again, the results change a bit, but not

²¹ For a number of reasons, this result should not be interpreted as implying that variation in class size has no effect on student performance. For example, if a smaller ratio of pupils per teacher reflects the provision of compensatory services to low-performing students, it would be hard to detect positive effects of smaller class sizes. For a more careful study of the effects of class-size on student performance, see Ferguson and Ladd, 1996.

Table 3

Program impacts for seventh grade by subgroups—math^a (absolute value of *t*-statistics in parentheses)

	All	Economically disadvantaged	Black	Hispanic	White
<i>Basic model^b</i>					
YR92*Dallas	0.154** (2.30)	0.111 (1.38)	0.212** (2.69)	0.142 (1.62)	0.133 (1.37)
YR93*Dallas	0.168** (2.54)	0.175** (2.22)	0.204** (2.61)	0.184** (2.16)	0.067 (0.69)
YR94*Dallas	0.121** (2.20)	0.075 (1.12)	0.026 (0.39)	0.196** (2.74)	0.198** (2.37)
YR95*Dallas	0.121** (2.20)	0.076 (1.14)	– 0.017 (0.26)	0.212** (2.97)	0.188** (2.22)
Number of observations	1118	1083	778	968	638
Adjusted <i>R</i> ²	0.57	0.26	0.30	0.36	0.43
<i>Extended model^c</i>					
YR92*Dallas	0.135** (1.96)	0.106 (1.28)	0.181* (1.94)	0.145 (1.63)	0.128 (1.30)
YR93*Dallas	0.153** (2.24)	0.187** (2.29)	0.174* (1.89)	0.198** (2.29)	0.042 (0.42)
YR94*Dallas	0.141** (2.37)	0.086 (1.19)	0.152* (1.82)	0.222** (2.93)	0.206** (2.39)
YR95*Dallas	0.168** (2.82)	0.128* (1.77)	0.078 (0.94)	0.259** (3.43)	0.221** (2.54)
Number of observations	1118	1083	778	968	638
Adjusted <i>R</i> ²	0.58	0.27	0.31	0.37	0.43

Notes: ^a In both models, the dependent variable is the relative pass rate on the TAAS and observations are weighted by pupils in the school. * Indicates significance at the 10 percent level, and ** indicates significance at the five percent level. ^b Additional variables in the basic model are indicator variables for school years 1992, 1993, 1994 and 1995 (1991 = base); school characteristics (percent black, percent Hispanic, percent LEP, percent disadvantaged, and mobility rate); interactions between 1994 and 1995 school years and the school characteristics; and indicator variables for the cities of Dallas, Austin, El Paso, Fort Worth and San Antonio (Houston = base). ^c The independent variables in the extended model include all of the variables in basic model plus interactions between the school years and district indicator variables to represent the districts (Houston, Fort Worth and Yselta) that also have reform programs currently in progress.

enough to alter the basic story given that the patterns across years and across subgroups are similar to those for the full sample.

4.4. Summary of seventh grade results

In sum, the results for seventh grade pass rates (based on sixth grade rates in 1994 and 1995) are encouraging in that they indicate positive impacts of the Dallas program on some subgroups both relative to all the other big Texas cities and relative to gains in the other cities with local reform initiatives. Hispanic and white children in Dallas exhibit gains relative to the other cities, but, for reasons that are unclear and worth pursuing in future work, black students do not. The magnitude of the impacts depends on the appropriate interpretation of the 1992 results. If these results represent true program

impacts then the relevant measures of program impacts are given by the reported coefficients for the 1995 interaction terms and are on the order of 10 to 20 percentage points relative to the state average. If, however, the 1992 results represent regression-to-the-mean or simply the gains from narrow teaching to the test, then the estimated impacts of the program would be smaller and would be measured by the difference between the 1995 and the 1992 estimated coefficients. In either case, it should be noted that the 1995 results represent four years of experience with the program. That is, the students who were being tested in 1995 had an opportunity to increase their learning in each of the three prior years. Hence, gains expressed in annual terms would represent approximately one quarter of the coefficient estimates.

5. Impact on third grade pass rates

In many ways it would have been more logical to start this analysis with third grade test results. While improvements in the performance of elementary school students do not guarantee that the performance of students in the later grades will improve over time, the absence of improvements at the elementary level clearly makes it harder to obtain meaningful improvements in later years in the higher grades.

Unfortunately, two factors make it difficult to analyze the performance of elementary school students on the TAAS, both of which are related to the fact that the TAAS is first administered in the third grade (and that is the only grade in elementary school for which test data are available for all the years). First, I am not able to substitute second grade test scores for third grade scores to account for the shift to the spring administration of the test in 1993–94. Consequently, the tests cover more difficult material in the latter two years of the sample than in the first three years, a fact that matters greatly for disadvantaged students in large urban areas. Additional analysis of student cohorts (not reported here) shows that Dallas students lagged further behind other urban school districts as more advanced material was introduced, a fact that ideally should be taken into consideration when measuring impacts of the Dallas incentive program.²² The other problem is that during the early years of the sample period, schools in Dallas, and maybe elsewhere as well, exercised a lot of discretion about whether third grade students with limited English proficiency would be tested.²³ As a result, the average third grade pass rate is a flawed measure of student performance.

Given these limitations, the results for third graders (not reported) should be interpreted cautiously. The patterns over time are pretty clear for the category of “all” students and for most of the subcategories. A large positive coefficient in 1993, the first full year of the program,

is typically followed by a negative coefficient in the following year and a small and insignificant coefficient in 1995. In no case is there evidence that as of 1995 the pass rates of third grade elementary students in Dallas had increased more in Dallas than in other large Texas cities relative to the 1991 base year.²⁴ This bleak picture is only slightly modified when the model is extended to include interaction terms for the reforms in other cities or districts. In that case, the most hopeful result emerges for black students who appear to have gained about 7 percentage points relative to the state average more than black students in other nonreforming cities by 1995. However, this positive outcome is overshadowed by even larger positive impacts for blacks in the other reforming cities. Hence, this analysis provides virtually no evidence that the Dallas accountability program improved third grade outcomes after the first full year of the program. Whether the results would look better if a comparable test had been available over time or if the testing of students with limited English proficiency had been more consistent is hard to say.

6. Drop-out rates and attendance

Because of relatively high drop-out rates among high school students which would contaminate the concept of a pass rate, similar analyses of pass rates on the high school TAAS exit exam are not reported. Instead, I use a panel data analysis similar to the pass-rate analysis to look at changes in high school drop-out rates in Dallas relative to the other cities. Because of a lag in the reporting of drop-out rates, data are available only through the 1993–94 school year. Hence, the equations include interaction terms only for 1992 through 1994.

The results for those variables are reported in Table 4.²⁵ Starting from an overall drop-out rate of 11 percent in Dallas in 1991 which was substantially higher than that in the other five cities during that year, drop-out rates declined more in Dallas relative to other cities in each of the three years, 1992, 1993, and 1994. The extent to which these reductions should be attributed to the accountability program is unclear. Nonetheless the pattern for most of the subgroups is clear. Compared to

²² In the cohort model, fourth grade pass rates were estimated as a function of third grade pass rates in the same schools in the previous year for the same cohort of students, a vector of school characteristics, and an indicator variable for Dallas. The coefficients of the Dallas indicator variable were negative and statistically significant for reading and insignificantly different from zero for math. Based on the reasonable assumption that the Dallas incentive program did not slow the relative rate at which children in Dallas improved their reading skills, the results for reading indicate that in the absence of the program students in Dallas would have lagged behind their counterparts in other cities as the material became more difficult. While this finding complicates the interpretation of the results in the text, I suspect that many people would view the Dallas program as unsuccessful if it was not able even to offset the expected relative decline in test performance.

²³ Based on interview with Dallas administrators.

²⁴ The coefficients (and *t* statistics) of the Dallas-year interaction terms in the equation for all students for reading starting with 1992 are 0.021 (1.01), 0.058 (2.74), – 0.035 (1.62) and 0.011 (0.49) and for math are 0.058 (2.18), 0.082 (3.09), – 0.033 (1.20) and – 0.001 (0.03). The regressions are based on about 3000 observations. The adjusted *R*² in the reading equation is 0.41 and in the math equation is 0.36.

²⁵ We have left out the column for economically disadvantaged students because high school students exhibit erratic patterns of signing up for the free and reduced-price lunch program which serves as the basis for that classification.

Table 4

High school drop-outs—selected results^a (absolute value of *t*-statistics in parentheses)

	All	Black	Hispanic	White
YR92*Dallas	– 1.757 (1.56)	– 1.568 (1.40)	– 3.443** (2.24)	– 2.484** (2.12)
YR93*Dallas	– 4.419** (3.94)	– 3.629** (3.33)	– 7.151** (4.71)	– 4.316** (3.62)
YR94*Dallas	– 6.677** (5.95)	– 5.835** (5.36)	– 8.993** (5.96)	– 5.877** (4.87)
Number of observations	435	403	431	379
Adjusted <i>R</i> ²	0.55	0.46	0.54	0.45

Notes: ^a Dependent variable is the relative pass rate on the TAAS and observations are weighted by pupils in the school. Additional variables in this equation include indicator variables for school years 1992, 1993, and 1994 (1991 = base); school characteristics (percent black, percent Hispanic, percent LEP, percent disadvantaged, and mobility rate); and indicator variables for the cities of Dallas, Austin, El Paso, Fort Worth and San Antonio (Houston = base).

other cities, the drop-out rate decreased more in Dallas relative to the 1991 base year in 1992 and apparently continued its relative fall in 1993 and 1994.

Finally, I use a similar approach to examine attendance at both the elementary and middle schools. Because attendance figures are not broken down by student subgroup and data are available only for 1991, 1993, and 1994, the comparison is between the latter two years and 1991. For elementary schools, the attendance rate increased by about one quarter of a percentage point more in Dallas in 1993 relative to 1991 than it did in the other Texas cities and even more in 1994, outcomes that may well reflect a separate incentive program in Dallas to promote school attendance. For junior high schools, no effect emerges from the data.²⁶

7. Turnover rates of principals

The research on effective schools highlights the key role of the principal as a leader in high-performing schools (see, for example, Purkey and Smith, 1983, and citations therein and Lee et al., 1992). Given this important role for principals, the replacement of ineffective principals could potentially generate significant gains in student learning. Hence, the question for this section is the extent to which the Dallas accountability and incentive program has changed the behavior of the superintendent and the Board of Education with respect to their treatment of school principals.

The trends in principal changes by type of school are

shown in Table 5. As a percent of the total positions each year, a clear upward trend emerges. In 1991 (which represents changes between the 1990–91 school year and the 1991–92 school year), only 4.7 percent of the principals were replaced. By 1994, the percent had increased to 28.5 percent. The low turnover rate in 1991 in all schools and also for each type of school is striking. In the absence of any additional information on earlier rates of principal turnover, these 1991 turnover rates represent the best estimates of turnover rates before the introduction of the Dallas incentive program. In 1992, which represents the end of the first year of the program, turnover rates increased dramatically in each type of school, thereby providing evidence of a dramatic change, and they continued to increase in 1993 and, for all schools and for elementary schools, in 1994. Thus, it appears that the new emphasis on accountability made the District much more willing than in the past to change principals. Discussion with the superintendent indicated that the turnover does not simply represent the movement of poor principals from one school to another. In fact, many principals were either demoted or fired.

The number of elementary schools is large enough to examine how the changes in principals were distributed among schools with differing school-effectiveness rankings as of the pre-program year. For this purpose the elementary schools were divided into three even categories, labeled as least effective, average, and most effective based on the 1992 school effectiveness rankings, and principal change rates calculated for two periods: the pre-program period of one year, and the post program period of three years. The patterns are reported in Table 6. Prior to the program, the average turnover rate was extremely low in the least-effective schools. During the program years, the turnover rate rose to about 25 percent in all three categories. Thus, there is no evidence that in the post-program period that principals are being

²⁶ The other variables in the equation are 1993 and 1994 year indicator variables; school characteristics (percent black, percent Hispanic, percent LEP, percent disadvantaged, and the mobility rate); five city indicator variables, and a constant term.

Table 5
Turnover rates for principals by year and by type of school

	1991	1992	1993	1994
<i>All schools^a</i>				
Changes (1) ^b	8	28	44	49
Positions (2)	172	172	172	172
(1)/(2) (%)	4.7	16.3	25.6	28.5
<i>Elementary schools^c</i>				
Changes (1) ^b	7	19	32	40
Positions (2)	124	124	124	124
(1)/(2) (%)	5.7	15.3	25.8	32.3
<i>Middle schools</i>				
Changes (1) ^b	1	6	5	4
Positions (2)	21	21	21	21
(1)/(2) (%)	4.8	28.6	23.8	19.1
<i>High schools</i>				
Changes (1) ^b	0	3	7	5
Positions (2)	27	27	27	27
(1)/(2) (%)	0.0	11.1	25.9	18.5

Notes: ^a Excludes magnet schools and academies. ^b Changes occur in the period indicated by the year. For example, for 1991, the change would occur between the 1990–91 school year and the 1991–92 school year. ^c Includes K-3, 4-6 and K-6 schools.

Table 6
Principal turnover rates by school effectiveness (percent)

	Pre-program 1991	Post-program 1992–1994
Least effective	2.4	24.6
Average	6.7	25.0
Most effective	6.3	24.4

replaced more frequently in the least-effective schools. Nonetheless, the fact that the rate of principal turnover increased the most (from 2.4 to 24.6 percent) in the least-effective schools suggests that the program may be generating differentially intense scrutiny in the low-performing schools and could lead to better school leadership in the future. An alternative interpretation of these results is that a high turnover rate of principals could lead to undesirable outcomes as principals become insecure about their jobs. While the optimal rate of principal turnover is unclear, my interviews suggested that some of this turnover of principals was clearly desirable and that it might not have occurred in the absence of the spotlight of the accountability program.

8. Conclusion

During the early 1990s, the Dallas Independent School District engaged in an interesting and sophisticated (at least along some dimensions) experiment in the use of a performance-based accountability system buttressed by financial incentives to school staff to increase student

learning. Because other school districts and state boards are interested in such school-based incentive programs, it is important that the Dallas program be evaluated. Evaluations can take many forms and can focus on different aspects of the program. The evaluation in this paper focuses on student outcomes as measured by pass rates on the state criterion-referenced test, high school drop-out rates, and one measure of internal change, namely the turnover of principals, during the first four years of the program. Additional evaluations that look more closely at what changes are occurring within the schools and that cover a longer time period would be a desirable next step.

The third-grade results notwithstanding, the analysis presented here allows one to be cautiously optimistic about the potential for such a program to have a positive impact on student outcomes. The seventh grade results are consistently positive and relatively large, but primarily for Hispanics and whites. In contrast, less positive results for black students provide cause for concern. Other potentially positive changes include the fall in the Dallas drop-out rate relative to that for other cities and the fact that principals are being turned over more readily

than in the past, both of which could signify real change in the system and could bode well for the future.

However, I remind the reader that this paper focuses on measuring the effects of Dallas's accountability program on student outcomes, with little attention to processes within schools. Given that I have found some positive impacts, the next step would be to look inside the black box to determine first whether the gains in student performance on the TASS translate into gains in student learning, and second to relate the gains in outcomes to information about how principals and teachers changed their behavior in response to the accountability system and how those changes affected outcomes. An additional component would require a closer look at the potentially undesirable unintended effects of such programs with an eye to better understanding the benefits relative to the costs and to designing improved school-centered accountability programs.

Acknowledgements

This paper is part of a larger project on performance-based accountability that the author initiated as a Visiting Fellow at the Brookings Institution. She thanks the Brookings Institution and its Brown Center for Education Policy for financial and collegial support. In addition, she is indebted to the Spencer Foundation and an anonymous donor for financial support of the larger project. She is also grateful to her coauthor on related papers, Charles Clotfelter; her research assistant at Brookings, Tony Shen, and at Duke, Laura Hodges; and Thomas Kane, the Brookings study group on education, and members of the Duke University labor seminar for their insightful comments.

References

- Anderson, M., Babus, S., Bearden, D.K., Mendro, R.L., and Webster, W. J. (1994) Longitudinal Achievement Trends: 1989–94. Approved report of the Division of Research, Planning and Evaluation to the Dallas Public Schools (August).
- Cornett, L.M. and Gaines, G.F. (1994) Reflecting on ten years of incentive programs: The 1993 SREB Career Ladder Clearinghouse Survey. *Southern Regional Education Board Career Ladder Clearinghouse*. Atlanta, Georgia: Southern Regional Education Board.
- Clotfelter, C.T., and Ladd, H.F. (1996) Recognizing and rewarding success in public schools. In *Holding Schools Accountable: Performance Based Reform in Education*. ed. H. F. Ladd, pp. 23–63, Washington, DC: The Brookings Institution.
- Commission for Educational Excellence (1991). *Final Report*. Board of Trustees, Dallas Independent School District (June).
- Dickens, W.T., 1990. Error components in grouped data: is it ever worth weighting? *The Review of Economics and Statistics* 72 (2), 328–333.
- Ferguson, R. and Ladd, H.F. (1996) How and why money matters: An analysis of Alabama schools. In *Holding Schools Accountable: Performance-Based Reform in Education* (Edited by Ladd, H.F.), pp. 265–298. The Brookings Institution: Washington, DC.
- Lee, V.E., Bryk, A.S., Smith, J.B., 1992. The organization of effective secondary schools. *Review of Research in Education* 19, 171–267.
- O'Day, J.A., and Smith, M.S. (1993) Systemic reform and educational opportunity. In *Designing Coherent Education Policy: Improving the System* (Edited by Fuhrman, S.H.), pp. 250–312. Jossey-Bass: San Francisco.
- Purkey, S.C., Smith, M.S., 1983. Effective schools: a review. *The Elementary School Journal* 83 (4), 427–452.