The Impact of Assessment and Accountability on Teacher Recruitment and Retention: Are There Unintended Consequences?

Donald Boyd, Hamilton Lankford, Susanna Loeb and James Wyckoff

Public Finance Review 2008; 36; 88
DOI: 10.1177/1091142106293446

Public Finance Review
Volume 36 Number 1
January 2008 88-111
© 2008 Sage Publications
10.1177/1091142106293446
http://pfr.sagepub.com
hosted at
http://online.sagepub.com

The Impact of Assessment and Accountability on Teacher Recruitment and Retention

Are There Unintended Consequences?

Donald Boyd
Hamilton Lankford
State University of New York at Albany
Susanna Loeb
Stanford University, California
James Wyckoff
State University of New York at Albany

This article uses data on every teacher in New York State public elementary schools from 1994-1995 through 2001-2002 to examine the response of teachers to the implementation of state-mandated testing. The authors ask whether the introduction of testing in the fourth grade has increased the turnover of fourth-grade teachers, whether testing differentially affected the decisions of teachers with particular attributes, and whether the characteristics of teachers entering the fourth grade changed with the introduction of testing. The authors find that the turnover rate of fourth-grade teachers decreased relative to teachers in other elementary grades since testing began. In addition, entering fourth-grade teachers are less likely to be inexperienced teachers than those moving into other elementary school grades.

Keywords: *educational accountability; education policy*

Authors' Note: We thank Amita Chudgar and Michelle Reininger for excellent research assistance. We appreciate comments from Johnny Yinger and participants at the Syracuse Conference on Teacher Quality and Teach Retention and two anonymous reviewers. We thank the U.S. Department of Education for financial support (No. R305T01008002). The views expressed in the article are solely those of the authors and do not necessarily reflect those of the U.S. Department of Education. Similarly, all errors are attributable to the authors.

1. Introduction

Assessment-based accountability systems, long employed in the private sector, are increasingly utilized to manage large public organizations, such as health care and K-12 education. These applications are noteworthy as they attempt to substitute incentives for a long history of oversight and regulation. Yet we know little about how the actors in such systems respond. Do these systems produce unintended consequences that mitigate program goals? As an example of one such unintended consequence, it is often suggested that student assessment-based accountability systems will induce good teachers to avoid teaching situations where the pressure that accompanies the scrutiny and sanctions is greatest, for example, teaching in tested grades in low-performing schools. If this were to occur, it would add to a long list of job characteristics that make it less likely that the best teachers will teach in the schools where their talents are needed most.[1] In what follows, we explore this question employing data for New York State public schools.

State-mandated testing has become the norm in public elementary and secondary schools across the nation. In many states, test results are used as a measure of the effectiveness of districts, schools, and teachers; and sanctions are tied to low performance. Indeed, there is some evidence that test-based accountability systems may result in higher student achievement (Jacob 2005; Carnoy and Loeb 2002; Hanushek and Raymond 2005). States are strengthening these provisions to meet the No Child Left Behind Act of 2001 (NCLB). NCLB requires that all children be tested in English and math in grades 3 through 8; that the results of these tests be made public in report cards that are widely disseminated; that states set annual performance goals for schools, based on the percentage of students scoring proficient on state reading and math assessments; and that subgroups of students by race, income, and other groupings show consistent progress until, within twelve years, all students demonstrate proficiency in reading and math. Provisions for adequate yearly progress have already begun to place substantial pressure on school administrators and teachers, and these pressures are likely to increase as large numbers of schools fail to meet the targets. In response to these pressures, there is evidence that administrators and teachers manipulate testing situations to demonstrate improvement in student performance (Jacob 2005; Figlio and Getzler 2002; Cullen and Reback 2002).

Clearly there are reasons that teachers may be dissatisfied with the recent changes. Interview and survey research suggests that teachers feel pressure to deliver high student test scores (Barksdale-Ladd and Thomas 2000; Hoffman, Assaf, and Paris 2001). In addition, many teachers indicate that they view the high-stakes tests as an imposition on their professional autonomy, an invasion into their classrooms, a message that the state views them as incompetent, and a hindrance to professional creativity (Luna and Turner 2001). As districts and schools put more emphasis on test performance, teachers may not have as much flexibility in their classrooms. They may face pressures to teach topics that that they are less interested in or believe are less important for students or they may need to teach in ways that increase test scores but not other important skills. Teachers also may have more day-to-day distractions as parents and administrators scrutinize the details of their classrooms. Moreover, teachers may worry about the security of their jobs, particularly if they teach in schools with low-performing students, which are more likely to encounter repercussions from the state. Clotfelter et al. (2004), for example, found that North Carolina's accountability system makes it more difficult for low-performing schools to retain teachers. Provisions in teacher contracts may make teacher transfers more likely. Many contracts provide seniority-based preference to teachers to fill within-district vacancies. As a result, more experienced teachers are able to avoid difficult working conditions, one of which may be teaching in tested grades. Anecdotal accounts suggest that this policy is common in New York school district teacher contracts. Since there are about seven hundred school districts, each of which has its own teacher contract that is typically renegotiated every three years, we have only limited evidence of the prevalence of seniority-based hiring preferences.[2]

However, the increased emphasis on accountability and rewards may benefit teachers as well. Standards-based reforms can provide opportunities for schools to focus on student learning. While this was sure to have been the case in many schools prior to the recent reforms, there were other, poorly functioning schools preoccupied in other ways. Administrators may use accountability policies as leverage with the district to get rid of ineffective or distracting teachers and may simply focus more on trying to create a school that benefits students. Teachers may prefer to teach in these environments rather than in environments that do not recognize success in the classroom. In addition, as a result of the pressures on schools, administrators may encourage their best teachers to move into the grades where testing is mandatory, in hopes of raising their schools' scores. One way to encourage such reassignment of teachers, for example, would be to target additional resources to the testing grades. If this were the case, we may see high-quality teachers moving into grades that most directly affect student scores on the standardized tests.

Thus, it is unclear, a priori, how the recent reforms will affect teachers. They may dissuade potential teachers from entering the classroom, increase transfers, or increase the probability that teachers will quit. These effects may be greatest in low-performing schools that already have difficulty attracting and retaining well-qualified teachers. However, the effects may work in the exact opposite direction if testing and accountability have made teaching more satisfying, especially in schools that had been mismanaged prior to reform.

This article proceeds as follows. Section 2 describes assessment-based accountability in New York, section 3 describes the data and methods in more detail, section 4 provides the results, and section 5 concludes with policy implications. We find, somewhat surprisingly, that the turnover rate of fourth-grade teachers has not increased and has, in fact, decreased relative to teachers in other grades since the implementation of the fourth-grade test. This decrease in the quit rate is consistent across urban, suburban, and rural areas and across schools with differing levels of student performance. In addition, we find little evidence that high-ability teachers are more likely to leave fourth grade; in fact, in some cases these teachers are less likely to leave. We do find some evidence that more experienced teachers decrease their turnover less than more recent hires, especially in suburban districts. We also find substantial differences in the characteristics of teachers moving into the fourth grade. New fourth-grade teachers are markedly less likely to be first-year teachers than those moving into other elementary school grades. In the lowest-performing schools, new fourth-grade teachers are also more likely to have attended highly competitive undergraduate institutions. These results provide evidence that the distribution of teachers is changing as a result of the test, though not necessarily in the predicted directions.

2. Accountability in New York State

New York has a long history of assessment-based accountability. Since 1878, it has had curriculum-based high school exit exams for awarding the Regents high school diploma with a different curriculum and exams that award the less prestigious local high school diploma. In addition, it has employed curriculum-based elementary and middle school exams as diagnostic tools for student remediation since the mid-1970s. From 1975 to 1998, these exams were given in third- and sixth-grade math and English language arts. However, the accountability environment has changed substantially in recent years. The Regents adopted learning standards in 1996 that substantially altered high school graduation requirements and that are supported by a new curriculum-based assessment system. Beginning with the first administration of the fourth- and eighth-grade tests in 1998-1999, all students are held to the same high learning standards, and exit exams in five subject areas that are required for a high school diploma were phased in over the next several years. The learning standards define curriculum in every grade with statewide assessments in the fourth and eighth grade intended to gauge progress.[3]

In 2000, the Regents adopted an accountability system that established school performance standards based upon student performance in elementary, middle, and high school exams. Failure to meet this standard triggers a planning process. Continued failure to meet the standard results in designation as a school in need of improvement and could lead to designation as a School under Registration Review (SURR). Poor performance by SURR schools can result in the dissolution of the school. These provisions augmented the school report card system that had been in place since 1998-1999. New York is ranked highest among the states in the adoption of the standards and accountability.[4]

Prior to 1998, assessment in New York, like many states, was designed primarily to provide information to local administrators and teachers to diagnose and remediate poor performing students. The system did not hold teachers, schools, or districts accountable for the results, except in the most extreme cases where schools could potentially lose their registration. Since 1997, accountability has been more explicit. Beginning with the 1998-1999 school year, report cards that are widely disseminated and receive extensive public discussion have brought pressure on schools and districts. In 2001, consequences for schools were increased. Thus, the 1998-1999 year is the first year with the new fourth- and eighth-grade exams and a public report card of the performance of individual schools. For purposes of our analysis, 1998-1999 is viewed as the beginning of the postreform period.

Unlike some states, New York has no direct consequences for teachers. However, indirectly fourth- and eighth-grade teachers have felt considerable pressure. Many observers associate the performance of the students in the tested subjects (e.g., fourth-grade English language arts and math) to those teachers. Since there are consequences for schools for poor performance, these teachers are likely to be subject to greater pressure than teachers in nontested grades and subjects. This occurs, despite the fact that the test measures performance to an absolute standard, rather than year-to-year changes, which implies that it is cumulative from previous grades. As evidence of this pressure, there is anecdotal evidence that some New York teachers have cheated to improve student scores. According to a spokesperson (a Mr. Burman) for the state education department,[5]

- "Teachers have taken actions such as illicitly reviewing tests in advance and tailoring their instruction to match specific questions; improperly giving students passing grades when they score tests for the state; and telling students to correct answers the teachers knew to be wrong."
- "Most of the incidents involved testing in elementary and middle schools, where the state uses test scores in report cards to chart school progress."

Cheating by teachers on the state's high school Regents exams, which students must pass to graduate, was less common, according to the spokesperson. These examples are consistent with those reported in other states and with empirical evidence that in some instances accountability systems have led to cheating (Jacob and Levitt 2003).

3. Empirical Approach

This article uses data on New York State to study the response of teachers to assessment and accountability. The data set includes information on every teacher that has been a part of the New York State public schooling system from 1994-1995 through 2001-2002. This eight-year record allows us to track individual teachers across grades and schools over their course of employment in any New York State public school, identifying the grade level a teacher taught both before and after the implementation of testing. In addition, we have linked to this data set information about the qualifications of the teachers including experience, gender, race/ethnicity, selectivity of undergraduate institution, and performance on state certification exams.

The article is driven by four questions. First, has the introduction of high-stakes testing in the fourth grade increased the turnover of fourth-grade teachers? We look at both transfers across grades and exits from teaching in New York State. We focus on the fourth-grade test and not the eighth-grade test because fourth-grade teachers typically teach only students in one grade and are likely to have more opportunities for mobility across grades within schools, where eighth-grade is often the highest grade in middle schools.[6] This results from greater variability in the grade configurations for schools with eighth grades than those with fourth grades. We also assess whether the impact of these tests has differed systematically across schools by urbanicity and student achievement.

Second, has the introduction of high-stakes testing differentially affected the transfer and exit decisions of teachers with different characteristics across all elementary grades? We look at differences by teachers' experience, ranking of undergraduate institution, and teacher certification exam failure.

Third, have the characteristics of teachers entering fourth-grade teaching positions changed with the introduction of testing? Again, we ask whether the effect of testing on the characteristics of new teachers has differed in different types of schools.[7]

Finally, has the introduction of mandatory testing led to an increase in turnover across elementary grades in low-performing schools relative to higher-performing schools? Since other factors have been changing at the same time as the introduction of the test, the comparison between schools based on student performance is not as clean a method for assessing the effect of testing as is the comparison across elementary grades; however, it provides a framework for interpreting the other results. As suggested in the introduction, it is entirely possible that assessment and accountability may affect teacher retention and teacher quality through both demand and supply. The data available to us do not permit estimation of a structural model. As described below, we estimate reduced form models to address the four research questions.

3.1. Data

Several data sets collectively provide information for every teacher and administrator employed in a New York public school at any time from 1994-1995 through 2001-2002. The core data come from the Personnel Master File (PMF), part of the Basic Education Data System of the New York State Education Department. The PMF identifies approximately two hundred thousand teachers each year. We have linked annual records for individuals from these data sets, yielding detailed data characterizing the career history of each individual. The data contain a range of information about the qualifications of teachers as well as the environments in which these individuals make career decisions. We obtain a measure of college selectivity by combining the names of the institutions from which individual teachers earned their undergraduate degrees, with the Barron's ranking of college selectivity. We also have access to all certification exam scores for each teachers and whether they passed the exams on their first attempt. A school-level data set adds information on the location, grade span, student composition, and student performance for each school.

This article relies on data on teachers in the first through sixth grades in the years surrounding the implementation of the test, 1994-1995 through 2001-2002. Table 1 shows that, of the teachers in the New York State system during these years, 72 percent had attended a noncompetitive or less competitive undergraduate institution, more than half had more than ten years experience in the New York State system, and three-quarters were white. Over the sample period, an average of 29.3 percent of teachers did not return to the same grade in the following year, and 6.8 percent did not continue teaching in the New York State system. Nearly one-third (30.8 percent) of teachers had not taught in the same grade the year before.[8] The characteristics of teachers do not differ substantially across grades, though the percentage of teachers who have failed a certification exam is lower in the upper grades than in the lower grades.[9] Over time, the percentage of teachers in their first year of teaching in New York State has increased as a result of increased enrollments and retirements. In addition, more teachers receive their undergraduate degree from noncompetitive colleges, and fewer fail their certification exams.[10]

Following the 1994-1995, 1995-1996, and 1996-1997 academic years, an average of 29 percent of fourth-grade teachers left their grade, compared with 27 percent of teachers in other elementary school grades. After implementation in 1998-1999, 31 percent of fourth-grade teachers left, as did 31 percent of teachers in other grades. A simple difference-in-difference suggests that the fourth-grade turnover rate decreased (rather than increased, as one might have expected) relative to other grades. However, this result may be due to changes in the composition of teachers in these grades instead of changes in the propensity of similar teachers to leave. Below, we use a multivariate framework to address this possibility.

3.2. Method

We begin by asking whether or not the introduction of high-stakes testing in the fourth grade has increased the turnover of fourth grade teachers. A logit model estimates the probability of a teacher leaving the grade they are teaching in:

$$\Pr(\text{leaving}) = \frac{e^{x\beta}}{1 + e^{x\beta}} \tag{1}$$

where $x\beta = \beta_0 + T\beta_1 + B\beta_2 + F\beta_3 + E\beta_4 + S\beta_5 + Y\beta_6 + G\beta_7 + G^4Y^{\text{post}}\beta_8$.

Here, the probability that a teacher will leave the grade in the following year, either to teach in another grade or to exit teaching all together, is a function of the teacher's years of teaching experience (T) as measured by whether he or she is a first-year teacher; teaching experience and the square of teaching experience; the competitive level of the undergraduate institution attended (B) as measured by three dummy variables; whether he or she failed a certification exam (F); the teacher's ethnicity (E); characteristics of

Table 1
Sample Descriptives

Variable	Observations	Mean (Standard Deviation)
Year = 1995	359,962	0.135
Year = 1996	359,962	0.136
Year = 1997	359,962	0.139
Year = 1998	359,962	0.141
Year = 1999	359,962	0.145
Year = 2000	359,962	0.152
Year = 2001	359,962	0.152
Grade = first	359,962	0.190
Grade = second	359,962	0.173
Grade = third	359,962	0.168
Grade = fourth	359,962	0.156
Grade = fifth	359,962	0.153
Grade = sixth	359,962	0.157
Failed general knowledge portion of National Teacher Exam or New York State (NYS) Teacher Certification Exam on first attempt	157,182	0.187
Failed general knowledge missing	359,962	0.563
Highly competitive undergraduate	304,757	0.091
Competitive undergraduate	304,757	0.185
Less competitive undergraduate	304,757	0.576
Noncompetitive undergraduate	304,757	0.149
College ranking missing	359,962	0.153
Experience in NYS	359,962	14.7 (10.4)
First year in NYS	359,962	0.070
Two to five years' experience in NYS	359,962	0.198
Six to nineteen years' experience in NYS	359,962	0.378
Twenty or more years' experience in NYS	359,962	0.354
White	199,645	0.778
Black	199,645	0.096
Hispanic	199,645	0.073
Other	199,645	0.053
Race/ethnicity missing	359,962	0.445
Urban school	359,962	0.412
Rural school	359,962	0.172
New York City metro area	359,962	0.571
% black or Hispanic students	359,962	0.372 (0.380)
% students on free lunch	359,962	0.415 (0.331)
% low-scoring students—math	359,962	0.338 (0.224)
% low-scoring students—English language arts	359,962	0.407 (0.212)
Exit grade in following year	359,962	0.292
Exit system in following year	359,962	0.068
New to grade	359,962	0.299

the school where he or she taught (S) including urbanicity, whether the school is in New York City, the percentage of black and Hispanic students, the percentage of students eligible for free lunch, and the percentage of students who scored at the lowest level on the New York State fourth grade mathematics exam and on the English language arts exam; dummy variables for the year (Y) and the grade (G); and interactions of fourth grade with years 1997-1998, 1998-1999, 1999-2000, 2000-2001, and G^4Y^{post}. These final interactions capture the impact of test implementation on teachers' exit decision. The coefficient on the interaction of fourth grade with 1997-1998 assesses whether teachers were more likely to leave just before test implementation, thus avoiding teaching in a testing grade. The coefficient on the interaction of fourth grade with the other years assesses whether teachers were more likely to leave in the year following test implementation.[11]

We estimate these models separately for two different groupings of schools based on the geographical setting of the school—urban, suburban, and rural—and the quartile of student achievement based on the students' test scores. In addition, we run a multinomial specification to distinguish teachers who transferred to a different grade within New York State and those teachers who left the system entirely. For illustrative purposes, we use an interaction of grade four with a binary variable equal to one in the tested years instead of including a separate interaction for each year.

Next we assess how the leaving behavior of teachers differs for teachers with different characteristics. To do this, we interact teacher characteristics with the fourth grade by year interaction. The coefficients on these new interactions measure the differential effect of the test on teachers with different attributes. We characterize the teachers by whether they are in their first year teaching in New York State, by whether they have more than ten years of experience, by whether they have a degree from a highly competitive undergraduate institution, and by whether they have failed a certification exam. We add three interaction terms to the original logit model: interactions between the characteristic of interest and dummy variables representing the tested years, fourth grade, and fourth grade by the tested years. For attributes with missing data, we also include interactions between a missing dummy and the above five measures so that the comparison is between teachers with available data.

In the third portion of the analysis, we look to see whether the attributes of teachers entering the fourth grade changed with the introduction of testing. Logit models determine the likelihood that a new teacher has certain attributes, including whether she or he is a first-year teacher or has two to five years of experience, six to nineteen years of experience, or twenty or more years of experience; whether she or he attended a highly competitive undergraduate institution; and whether she or he failed a core certification exam. We include the school characteristics from equation (1), dummy variables for the years 1995-1996 through 1999-2000 as well as for first through sixth grade, and an interaction term for fourth grade with the tested years. The coefficient on the interaction term measures to what extent new teachers in the fourth grade are more or less likely to have a certain characteristic given the characteristics of teachers in other grades, than we would predict if the test had not been implemented. Equation (2) summarizes this specification, where Z is the dummy variable representing the characteristic of interest.

$$\begin{gathered}\Pr(Z=1)=\frac{e^{w\gamma}}{1+e^{w\gamma}}\\ w\gamma=\gamma_0+S\gamma_1+Y\gamma_2+G\gamma_3+G^4Y^{99,00}\gamma_4+\varpi.\end{gathered}\tag{2}$$

4. Results

4.1. Has the Introduction of High-Stakes Testing in the Fourth Grade Increased the Turnover of Fourth-Grade Teachers?

The results from estimating equation (1) appear in table 2. In keeping with other studies, we find the probability of leaving higher for first-year teachers, for teachers from more competitive colleges, and for teachers in urban schools and in schools with higher proportions of black, Hispanic, and low-performing students (Grissmer and Kirby 1987; Ingersoll 2001; Murnane et al. 1991; Kirby, Naftel, and Berends 1999). We also find that teachers working in schools with low math achievement, those teaching in the New York City metropolitan area, and Hispanic teachers are more likely to leave. Individuals teaching in rural areas are less likely to leave. In addition, most of the year and grade dummies are significant.

Contrary to the hypothesis that testing has led to increased turnover of fourth-grade teachers, however, we find that the probability of a teacher leaving the fourth grade is lower in the testing years. The coefficient on the interactions between fourth grade and year indicate that teachers were approximately 7 percent less likely to leave the fourth grade in the year before implementation and 18 percent less likely to leave after the year of implementation than in previous years and other grades. Similarly, in the next two years, fourth-grade teachers were 3 and 5 percent less likely to

Table 2
Logit Estimates of a Teacher Leaving the Grade in the Following Year

Variables	Odds Ratio (*z*-Statistic)	Variable	Odds Ratio (*z*-Statistic)
Teacher attributes			
First year in New York State	1.12 (6.98)	% low math	1.65 (10.12)
Experience	0.94 (32.82)	% low English	1.13 (2.49)
Experience squared	1.001 (34.69)	Other attributes	
Highly competitive	1.24 (14.78)	Year = 1995-1996	1.01 (1.05)
Competitive	1.02 (1.98)	Year = 1996-1997	1.11 (7.28)
Noncompetitive	1.01 (0.85)	Year = 1997-1998	1.10 (6.52)
Ranking missing	1.14 (11.28)	Year = 1998-1999	1.17 (10.10)
Failed	1.01 (1.02)	Year = 1999-2000	1.19 (11.44)
Exam missing	0.91 (7.09)	Year = 2000-2001	1.18 (11.26)
Black	1.00 (0.23)	Grade = first	0.88 (7.27)
Hispanic	1.07 (3.71)	Grade = second	0.97 (1.83)
Other	0.96 (1.81)	Grade = third	0.96 (2.49)
Ethnicity missing	0.88 (11.67)	Grade = fourth	1.01 (0.69)
Urban	1.42 (23.95)	Grade = fifth	0.93 (4.52)
Rural	0.97 (–2.11)	Fourth grade in 1997-1998	0.93 (2.11)
New York City metro area	1.03 (3.46)	Fourth grade in 1998-1999	0.82 (6.40)
Student attributes		Fourth grade in 1999-2000	0.97 (1.13)
% black or Hispanic	1.39 (13.83)	Fourth grade in 2000-2001	0.95 (1.94)
% free lunch	0.94 (–1.68)		
Sample size = 359,942			
Psuedo R^2 = .046			

leave the fourth grade than the model predicts they would have been had the tests not been implemented.

The reduced exit behavior is evident for teachers in urban, suburban, and rural school settings as illustrated in table 3. In urban schools, fourth-grade teachers were 11 percent less likely to leave; in suburban and rural schools, these estimates are 8 and 9 percent, respectively. Table 3 also shows the results when split by the quartile of student performance. Quartile one includes the 25 percent of schools with the lowest proportion of students scoring at the lowest level. Thus, quartile one schools are the highest achieving, and quartile four,

Table 3
Logit Estimates of Teacher Leaving the Fourth Grade Relative to Other Grades by Urbanicity and Quartile of Student Test Performance: Odds Ratios (*z*-Statistics)

Variable	All	Urban	Suburban	Rural
Fourth grade post-1998	0.91	0.89	0.92	0.91
	(4.25)	(3.64)	(2.47)	(1.69)
n	359,962	148,390	149,769	61,803
	Highest Quartile	Quartile 2	Quartile 3	Lowest Quartile
Fourth grade post-1998	0.91	0.93	0.94	0.88
	(2.21)	(1.74)	(1.36)	(3.32)
n	89,938	89,026	90,061	90,937

Note: Student test scores are the percentage of students scoring at the lowest levels, level 1 or level 2, on the fourth-grade math exam. The first quartile is the schools with the lowest proportion of these scores. Models include all controls from table 2.

the lowest achieving. Again, across all four quartiles teachers are less likely to leave the fourth grade and the effect is the strongest in the lowest-achieving schools.

The reduced exit behavior of fourth-grade teachers could be driven by a reduction in quits or a reduction in transfers from fourth grade to other grades. To distinguish between these types of exit behavior, we run a multinomial logit with a three-level outcome variable: remaining in the same grade, switching grades but remaining in teaching, and leaving teaching. Table 4 gives these results. The positive effect of testing on teacher retention in the fourth grade is driven primarily by reduced transfers from the fourth grade. However, we see some reduction in the decision to quit teaching, as well.

4.2. Has the Introduction of High-Stakes Testing Differentially Affected the Transfer and Exit Decisions of Teachers with Different Characteristics?

Table 5 shows some difference in the impact of test implementation on the probability of leaving across groups of teachers. Teachers just completing their first year of teaching show greater increased retention in the fourth grade following reform than do more experienced teachers. First-year teachers were 25 percent less likely to leave the fourth grade following

Table 4
Multinomial Logit Estimates Distinguishing Leaving the Grade from Leaving the System: Relative Risk Ratios (*z*-Statistics)

Variable	All	Urban	Suburban	Rural
Leaving grade				
Fourth grade post-1998	0.91	0.88	0.91	0.94
	(4.23)	(3.78)	(2.55)	(0.98)
Leaving system				
Fourth grade post-1998	0.94	0.94	0.96	0.84
	(1.54)	(0.97)	(0.63)	(1.77)
n	359,962	148,390	149,769	61,803

Note: Models include all controls shown in table 2.

reform (*z*-statistic = 3.80); teachers with two to five years of experience were 14 percent less likely to leave the fourth grade following reform (0.75 × 1.147), although with a *z*-statistic of 1.61 they are not statistically significantly so at typical significance levels of .10 or greater; teachers with six to nineteen years of experience were 5 percent less likely to leave (*z*-statistic = 2.92); and those with twenty or more years of experience were estimated to be 4 percent more likely to leave (*z*-statistic = 3.46). Thus, none of the teacher experience groups is more likely to leave following implementation.

Teachers that received undergraduate degrees from highly competitive institutions also show greater increased retention in the fourth grade following reform than do their colleagues from less competitive colleges. The odds ratio for the interaction suggest that these teachers are 13 percent less likely to leave as a result of the tests than other teachers.

When urban, suburban, and rural schools are assessed separately in table 5, two trends emerge. First, in suburban schools, fourth-grade teachers from more competitive colleges are less likely to leave the fourth grade in the year following test implementation than are teachers from less competitive institutions relative to other grades; this trend is less strong in urban and rural schools. Second, the general differences across experience groups in retention noted above result from the behavior of suburban teachers. For example, in urban schools, teachers in other experience groups are no more likely to leave than first-year teachers. However, in suburban schools, first-year teachers are 40 percent less likely to leave, teachers with two to five years of experience are 16 percent less likely to leave (0.604 × 1.395), and those with more experience are unaffected by the reform. When the schools are divided into quartiles of student achievement,[12] we see a similar trend. In high-achieving schools more

Table 5
Logit Estimates of Leaving the 4th Grade for Teachers with Different Characteristics by Urbanicity and Student Performance Quartile: Odds Ratios (*z*-Statistics)

Variable	Post-1998 × Grade 4	Experience = 2 to 5 Years	Experience = 6 to 19 Years	Experience = 20+ Years	Most Competitive	Failed Exam
Characteristic × Post-1998 × Grade 4	0.750	1.147	1.273	1.382	0.877	1.032
	(3.80)	(1.61)	(2.92)	(3.46)	(1.83)	(0.41)
Urban schools						
Characteristic × Post-1998 × Grade 4	0.849	1.050	1.033	1.120	0.907	1.065
n = 148,390	(1.62)	(0.28)	(0.29)	(0.87)	(0.72)	(0.71)
Suburban schools						
Characteristic × Post-1998 × Grade 4	0.604	1.395	1.710	1.880	0.835	.939
n = 149,769	(3.87)	(2.27)	(3.71)	(3.90)	(1.79)	(0.33)
Rural schools						
Characteristic × Post-1998 × Grade 4	0.731	1.014	1.443	1.390	0.969	.942
n = 61,803	(1.36)	(0.05)	(1.46)	(1.19)	(0.19)	(0.17)
Highest test quartile						
Characteristic × Post-1998 × Grade 4	0.504	1.692	2.147	2.148	0.760	0.835
n = 89,938	(4.01)	(2.75)	(4.05)	(3.65)	(2.04)	(0.68)
Quartile 2						
Characteristic × Post-1998 × Grade 4	0.729	1.133	1.473	1.752	0.979	1.073
n = 89,026	(1.88)	(0.66)	(2.09)	(2.69)	(0.16)	(0.32)
Quartile 3						
Characteristic × Post-1998 × Grade 4	0.824	0.962	1.089	1.209	0.830	1.255
n = 90,061	(1.22)	(0.22)	(0.49)	(0.98)	(1.22)	(1.32)
Lowest test quartile						
Characteristic × Post-1998 × Grade 4	.858	1.075	1.022	1.028	1.007	1.013
n = 90,937	(1.24)	(0.2)	(0.16)	(0.17)	(0.04)	(0.12)

Note: Models include all controls shown in table 2 plus interactions between each characteristic and grade 4 and post-1998.

experienced teachers are relatively more likely to leave, while in low-achieving schools this is not the case.

4.3. Have the Characteristics of Teachers Entering the Fourth Grade Changed with the Introduction of Testing?

Even though the exit rate of fourth-grade teachers decreased in the years of test implementation, some teachers did exit and new teachers took their place. In addition, a growing student population resulted in more new fourth-grade teachers. We next look at the characteristics of entering teachers and whether those characteristics are different for fourth-grade teachers in the years following the test relative to other years and other grades. Table 6 shows that teachers who are new to the fourth grade are less likely to be first-year and very senior teachers and are more likely to have one to four years of experience relative to other grades and pretesting years.[13] This is true across urban, suburban, and rural schools and across student achievement groups. The lowest-performing schools are the least likely to put first-year teachers into the fourth grade when a vacancy occurs. New teachers to the fourth grade are also more likely to have graduated from highly competitive undergraduate institutions. Again, this is particularly true for new fourth-grade teachers in the lowest-performing schools. Given the consistent findings that the first few years of teacher experience substantially increase the student achievement value added and that very senior teachers suffer somewhat of a reduction, this finding is consistent with the goal of raising fourth-grade achievement.[14]

4.4. Has the Probability of Quitting or Transferring Increased across All Elementary School Grades as a Result of the Tests?

So far, the analysis has compared fourth grade to other elementary grades. However, testing may have influenced quit and transfer behavior across grades. While this is an important question, it is also difficult to answer because the effect of the test cannot be separated from general time trends caused by other factors. Table 2 shows an increased probability of leaving a grade, either by quit or transfer, in the year following the test; but there had also been a general trend toward increased leaving since 1994-1995. If teachers are leaving because they worry that their students may not perform well, test-induced leaving may be greater in low-performing schools. Table 7 presents the results of logits estimating the probability that a teacher leaves

Table 6
Logit Estimates of Teacher Characteristics for New Fourth Grade for Teachers by Urbanicity and Student Performance Quartile: Odds Ratios (*z*-Statistics)

Variable	Experience = 0	Experience = 1-4 Years	Experience = 5-18 Years	Experience = 19+ Years	Most Competitive	Failed Exam
Post-1999 × Grade 4	0.92	1.19	0.99	0.86	1.09	0.98
	(2.06)	(4.73)	(0.26)	(3.46)	(1.57)	(0.36)
n	110,296	110,296	110.296	110,296	84,713	63,249
Urban schools						
Post-1999 × Grade 4	0.86	1.17	0.98	0.96	1.14	0.97
	(2.75)	(2.12)	(0.26)	(0.56)	(1.37)	(0.45)
n	62,031	62,031	62,031	62,031	43,562	36,756
Suburban schools						
Post-1999 × Grade 4	1.03	1.20	0.96	0.77	1.07	0.99
	(0.51)	(2.87)	(0.64)	(3.61)	(0.80)	(0.05)
n	34,982	34,982	34,982	34,982	29,770	19,586
Rural schools						
Post-1999 × Grade 4	1.00	1.28	0.96	0.77	1.11	1.20
	(0.03)	(2.33)	(0.43)	(2.32)	(0.68)	(0.89)
n	13,283	13,283	13,283	13,283	11,381	6,907
Highest test quartile						
Post-1999 × Grade 4	0.89	1.27	1.01	0.80	1.06	0.98
	(1.31)	(2.94)	(0.19)	(2.45)	(0.59)	(0.06)
n	20,839	20,839	20,839	20,839	17,349	11,626

Quartile 2						
Post-1999 × Grade 4	1.08	1.24	0.91	0.74	0.97	0.99
	(0.94)	(2.70)	(1.14)	(3.25)	(0.18)	(0.03)
n	21,444	21,444	21,444	21,444	17,470	11.779
Quartile 3						
Post-1999 × Grade 4	1.09	1.19	0.91	0.81	1.00	1.16
	(1.10)	(2.33)	(1.28)	(2.51)	(0.07)	(1.21)
n	26,074	26,074	26,074	26,074	19,770	14,345
Lowest test quartile						
Post-1999 × Grade 4	0.79	1.13	1.05	1.05	1.30	0.93
	(3.66)	(2.05)	(0.88)	(0.7)	(2.50)	(0.99)
n	41,939	41,939	41,939	41,939	30,124	25,499

Note: Models include urbanicity, whether the school is in New York City, the percentage of black and Hispanic students, the percentage of students eligible for free lunch, the percentage of students who scored at the lowest level on the New York State fourth-grade mathematics exam and on the English language arts exam, and dummy variables for the year (Y) and the grade (G).

Table 7
Logit Estimates of the Probability of Leaving Low-Performing Schools in the Testing Years: Odds Ratios (*z*-Statistics)

	All		Urban		Suburban		Rural	
Low-performing	1.11	1.15	1.11	1.07	1.04	1.06	1.12	1.05
	(4.67)	(2.98)	(1.59)	(0.44)	(1.38)	(0.85)	(2.06)	(0.37)
Low-Performing	0.93	0.86	1.02	1.16	0.91	0.87	0.91	1.04
Post-1998	(2.64)	(2.60)	(0.26)	(0.78)	(2.74)	(1.66)	(1.33)	(0.24)
× 1 Year		1.03		0.42		1.22		1.27
Experience		(0.24)		(2.44)		(1.34)		(0.79)
× 6-19 Years		1.13		0.76		0.997		0.82
Experience		(1.62)		(1.14)		(0.03)		(0.93)
× 20+ Years		1.58		1.17		1.17		0.70
Experience		(5.43)		(0.58)		(1.39)		(1.50)
× Most		0.96		1.13		1.09		0.61
Competitive		(0.76)		(0.66)		(1.52)		(3.00)
× Failed Exam		1.29		1.44		1.39		1.04
		(3.30)		(1.83)		(3.23)		(0.13)

Note: *Low-performing* is defined as being in the lowest quartile of student performance on the English language arts exam. Controls for all models include percentage of black and Hispanic students in school, percentage of free-lunch students in school, teacher experience, teacher experience squared, first-year teacher, whether the teacher has six to nineteen years of experience, whether the teacher has twenty or more years of experience, Barron's rating of undergraduate schools, whether the teacher failed a certification exam, teacher race/ethnicity, year dummies, and grade dummies. The first two columns also include controls for whether the school is urban, whether the school is suburban, and whether the school is in New York City. Controls for all other regressions include all of the above plus interactions between the characteristics presented in the table and low-performing school and between the characteristics and post-1998.

the schools either to go to another school or to leave the New York State public school system. The interactions between whether the teacher's school is in the lowest quartile of student achievement and the post-1998 dummy measure whether this leaving behavior has increased more in low-performing schools. We find that it has not. Overall and across urban, suburban, and rural areas, teachers are no more likely to leave their school following the implementation of testing than they were before. In suburban schools, the probability of leaving decreased by 9 percent relative to the pretesting years.

Table 7 also gives the interaction of teacher characteristics with the probability of leaving low-performing schools once the tests have been implemented. There are no evident trends based on teacher experience. In urban schools, first-year teachers are even less likely than teachers with moderate experience to leave; while in suburban schools, the trend goes in the other direction. The clearest trend is for teachers who failed their certification exam. Across the board, these teachers are more likely to leave following reform relative to their higher-scoring colleagues. While these estimates describe the change in leaving behavior in the years surrounding test implementation, they do not adequately establish a causal link. There are a number of other contemporaneous changes that could also alter the overall probability of leaving and the relative probability of leaving low-performing and high-performing schools. For example, from 1998 to 2002, expenditures per pupil generally and the average salaries of all teachers increased substantially.[15] These changes could reduce the likelihood that teachers quit.

5. Conclusions and Policy Recommendations

The results provide evidence that the distribution of teachers is changing as a result of the state-mandated testing. New teachers to the fourth grade in the years surrounding the implementation of exams were, on average, less likely to be first-year teachers and more likely to have attended highly competitive undergraduate institutions than new teachers in other grades and in other years. This was especially true in urban schools and in schools that had either very high- or very low-achieving students. This behavior could result from increased resources targeted to the fourth grade to assist with the new pressures of teaching for the exam. Additional analysis of the reallocation of resources other than teachers would help assess whether this is the mechanism for the change in the distribution of teachers.[16]

In addition, the results of this analysis do not accord with the popularly held belief that teachers are leaving tested grades as result of the implementation of high-stakes testing. In fact, we find exit rates decreased in the fourth grade after state-mandated fourth-grade exams were introduced in the 1998-1999 academic year. Fourth-grade turnover decreased across urban, suburban, and rural schools and across schools with different levels of student achievement. The overall decrease in leaving resulted primarily from a decrease in transfers out of the fourth grade for teachers who remained in the New York system but also from differential quits.

There are differences in the probability of transferring as a result of the tests across teachers with different characteristics. Teachers in their first years of teaching reduced their attrition more following testing than did more experienced teachers. We speculate that this differential leaving by more experienced teachers could result from these teachers being less willing to change their teaching styles or curricula to fit testing requirements. The differential by experience is primarily a suburban phenomenon. There is less difference in the impact of testing by experience levels for teachers in urban or rural schools. Similarly, the exit of more experienced teachers is concentrated in high-achieving schools. The trend is not evident in low-achieving schools. In addition to the differential by experience, we find a differential by the competitiveness of the undergraduate institution that teachers attended. Teachers with degrees from highly competitive colleges reduced their attrition more following reform than did their colleagues with degrees from less competitive colleges.

It is important to emphasize that this article does not provide evidence that the testing has reduced the overall quit rate of teachers. In fact, turnover appears to increase over the eight years of the study. It is difficult to attribute the cause of this change. Many factors, including the labor market in alternative occupations, could affect turnover in schools. Turnover did not increase more in the tested grades, and it did not increase more in low-performing schools. The analysis does assess only relatively short-term responses to testing. In the long run, teachers may become more or less invested in the new testing environment, and their responses may change. Pressures associated with testing also could change. What the results do suggest is that schools and teachers are responding to testing. Schools are keeping their teachers in the testing grades at a higher rate and replacing the ones that do leave with more experienced and, in some cases, higher-ability teachers. This shows flexibility not always evidenced in public institutions. However, if resources are reallocated to maximize test performance, then other areas may suffer. A thorough analysis of the impact of testing would need to account for schools' responses to testing and the potential impact of these policies on teaching quality and multidimensional student outcomes in both tested and nontested grades.

Notes

1. See Lankford, Loeb, and Wyckoff (2002) for a description of the sorting of teachers.

2. We have examined contracts in several of the larger urban school districts in New York and some form of seniority-based transfer policy exists in most of these districts.

3. Fourth-grade mathematics exams are given in May and English language arts exams in February. Additional tests were introduced in the following years: fourth-grade science (spring 2000), fifth-grade social studies (November 2000), eighth-grade science (spring 2001) and social studies (June 2001), and an intermediate-level technology education exam (June 2001) that will be given in the year the student has completed one year of study in this subject area (http://www.emsc.nysed.gov/ciai/assess.html).

4. See "State of the States" (2005).

5. See Hoff (2003).

6. For example, less than 15 percent of fourth-grade teachers who teach at least half-time spend less than full-time in the fourth grade. In eighth grade, half the teachers with at least 0.5 full-time equivalent teach in some other grade at least part of the time.

7. We also look briefly at whether turnover has increased across all elementary grades since the implementation of the tests, especially in low-performing schools. However, this identification comes from changes over time, and there are a number of factors including concurrent demographic changes that may affect overall turnover rates and the turnover rates of teachers in low-performing schools. Thus, we can not be certain that the changes we see are due to the tests.

8. The sample size for these last variables is smaller because we need to follow the teacher to the next year to know whether he or she left the system or the grade; thus, we do not have data for 1999-2000 on these variables.

9. *Failed* refers to having failed either the National Teacher Exam (NTE) General Knowledge Exam or the New York State Teacher Certification Exam (NYSTCE) Liberal Arts and Science Test on the first attempt.

10. Descriptive statistics by grade and year are available upon request.

11. For right-hand-side variables with missing data, we recode the missing observations to zero and include dummy variables for whether the variables are missing. This specification choice does not provide qualitatively different estimates from the one that eliminates observations with missing data.

12. We divided schools into quartiles based on the school average fourth-grade English language arts score in 2000.

13. We employ the term *post-1999* to designate postreform for this question. Since the reform was implemented in 1998-1999, 1999-2000 was the first year following reform that teachers were hired. To be consistent with observing leaving behavior post-1998, we observe hiring in the 1999-2000 year.

14. For evidence of the effect of teacher experience on student achievement value added, see Rockoff (2004); Rivkin, Hanushek, and Kain (2005); and Boyd et al. (2006).

15. This is based on calculations by the authors from data provided by the New York State Education Department.

16. We examined class size changes using an ordinary least squares (OLS) regression specification with class size as the dependent variable; school and teacher characteristics as controls; and year dummies, grade dummies, and fourth grade by year interactions as in equation (1). We found that the class size in the fourth grade was reduced relative to other elementary school grades in the first year of the test. They were also reduced relative to fifth and sixth grades in 1999-2000. However, a class size reduction policy target to kindergarten through third grade in 1999-2000 made the comparison between fourth grade and the earlier elementary grades uninformative for 1999-2000.

References

Barksdale-Ladd, M. A., and K. F. Thomas. 2000. What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education* 51 (5): 384-97.

Boyd, D., P. Grossman, H. Lankford, S. Loeb, and J. Wyckoff. 2006. How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy* 1 (2): 176-216.

Carnoy, M., and S. Loeb. 2002. Does external accountability affect student outcomes? A cross-state analysis. *Education Evaluation and Policy Analysis* 24 (4): 305-32.

Clotfelter, C., H. Ladd, J. Vigdor, and R. Aliaga. 2004. Do school accountability systems make it more difficult for low performing schools to attract and retain high quality teachers? *Journal of Policy Analysis and Management* 23 (2): 251-71.

Cullen, J., and R. Reback. 2002. Tinkering towards accolades: School gaming under a performance accountability system. Working Paper, Economics Department, University of Michigan, Ann Arbor.

Figlio, D., and L. Getzler. 2002. Accountability, ability and disability: Gaming the system. NBER Working Paper 9307, National Bureau of Economic Research, Cambridge, MA.

Grissmer, D. W., and S. J. Kirby. 1987. *Teacher attrition: The uphill battle to staff the nation's schools*. Santa Monica, CA: RAND.

Hanushek, E., and M. Raymond. 2005. Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management* 24 (2): 297-328.

Hoff, D. 2003. New York teachers caught cheating on state tests. *Education Week* 23 (10): 27.

Hoffman, J. V., L. C. Assaf, and S. G. Paris. 2001. High-stakes testing in reading: Today in Texas, tomorrow? *The Reading Teacher* 54 (5): 482-92.

Ingersoll, R. M. 2001. *Teacher turnover, teacher shortages, and the organization of schools*. Seattle, WA: Center for the Study of Teaching and Policy.

Jacob, B. 2005. Accountability, incentives and behavior: Evidence from school reform in Chicago. *Journal of Public Economics* 89 (5-6): 482-92.

Jacob, B., and S. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118 (3): 843-77.

Kirby, S. J., S. Naftel, and M. Berends. 1999. *Staffing at-risk school districts in Texas: Problems and prospects*. Santa Monica, CA: RAND.

Lankford, H., S. Loeb, and J. Wyckoff. 2002. Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis* 24 (1): 37-62.

Luna, C., and C. L. Turner. 2001. The impact of the MCAS: Teachers talk about high-stakes testing. *English Journal* 91 (1): 79-87.

Murnane, R. J., J. D. Singer, J. B. Willett, J. J. Kemple, and R. J. Olsen. 1991. *Who will teach?* Cambridge, MA: Harvard University Press.

Murnane, R. J., J. B. Willett, Y. Duhaldeborde, and J. H. Tyler. 2000. How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management* 19 (4): 547-68.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73 (2): 417–58.

Rockoff, Jonah. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94 (2): 247-52.

State of the states. No small change, quality counts 2005. 2005. *Education Week* 24 (17): 77-137.

Donald J. Boyd is deputy director of the Center for Policy Research, State University of New York at Albany, where his work includes research on teacher labor markets and school district finances. He is part of a research team examining the relationships between teacher preparation, student academic performance, and teacher labor market choices.

Hamilton Lankford is a professor of economics and public policy at the State University of New York at Albany, where he is involved in a variety of activities that link research to

education policy in New York. His academic publications in both economic and education policy journals include research on the teaching workforce, the allocation of education resources, the determinants of school choice, and the effects of enhanced school choice. In ongoing research, he is a principal investigator on the Teacher Pathways Project, focusing on the linkages between teacher preparation, teacher labor markets, and student outcomes.

Susanna Loeb is an associate professor of education at Stanford University. She studies resource allocation, looking specifically at how teachers' preferences and teacher preparation policies affect the distribution of teaching quality across schools and how the structure of state finance systems affects the level and distribution of funds to districts.

James Wyckoff is a professor of public administration and public policy in the Rockefeller College of Public Affairs and Policy, State University of New York at Albany. He currently is working with colleagues to examine attributes of teaching preparation and induction programs that are effective in increasing the retention of teachers and the performance of students.