NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

School Based Accountability and the Distribution of Teacher Quality Among Grades in Elementary School

SARAH C. FULLER
AND HELEN F. LADD

WORKING PAPER 75 • APRIL 2012

School Based Accountability and the Distribution of Teacher Quality Among Grades in Elementary Schools

Sarah C. Fuller
Duke University

Helen F. Ladd
Duke University

Contents

Acknowledgements

This paper was initially prepared for the Association for Education Finance and Policy, Spring Meeting, March, 2012 in Boston. The authors are grateful for financial support from the Smith Richardson Foundation and to the Center for the Analysis of Longitudinal Data in Education Research (CALDER).

CALDER working papers have not gone through final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication.

The views expressed are those of the authors and should not be attributed to the American Institutes for Research, its trustees, or any of the funders or supporting organizations mentioned herein. Any errors are attributable to the authors.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

School Based Accountability and the Distribution of Teacher Quality Among Grades in Elementary Schools

Sarah C. Fuller and Helen F. Ladd

CALDER Working Paper No. 75
April 2012

Abstract

We use North Carolina data to explore the extent to which teachers in the lower grades (K-2) of elementary school are lower quality than in the upper grades (3-5) and to examine the hypothesis that accountability contributes to a shortfall in teacher quality in the lower grades. Our concern with early elementary grades arises from recent studies that have highlighted that children's experiences in the early school years have long lasting effects on their outcomes, including college going and earnings. Using licensure test scores as the primary measure of teacher quality, we find that concern about teacher quality in the lower elementary grades is warranted. Teachers in those grades are of lower quality than teachers in the upper grades. Moreover, we find that accountability, especially the form required by the federal No Child Left Behind legislation, increases the relative shortfalls of teacher quality in the lower grades and increases the tendency of schools to move teachers of higher quality from lower to upper grades and teachers of lower quality from upper to lower grades. These findings support the conclusion that accountability pressure induces schools to pursue actions that work to the disadvantage of the children in the lower grades.

Introduction

Many studies have documented differences across schools in the quality of teachers, where quality is most often measured by teacher credentials, such as years of experience or teacher licensure test scores. Such studies consistently show that schools serving large proportions of disadvantaged students have teachers with weaker credentials than those serving more advantaged students (e.g. Clotfelter, Ladd , and Vigdor, 2007). At least one study has found a similar pattern for teacher quality as measured by value added, although the differences based on value added measures tend to be relatively small (Hannaway et al., 2010). To the extent that teacher credentials are predictive of student achievement, the uneven distribution of teacher credentials across schools is detrimental to the learning of disadvantaged students.

In this paper, we shift the focus away from differences across schools to how teacher quality is distributed among grades within elementary schools. Specifically we explore the extent to which teachers in the lower grades (K-2) are of lower quality than those in the upper grades (3-5). Our concern with the early elementary grades arises in part from recent studies that have highlighted that what happens to children in the early school years has long lasting effects on their subsequent outcomes, including their college going behavior and their earnings (Chetty et al., 2010; Dynarksi, Hyman, and Schanzenbach, 2011).

Such findings for investments in the early years of regular schooling are fully consistent with the findings from random assignment studies of early childhood programs. High quality programs such as the Perry/High Scope Project and the North Carolina Abecedarian program, for example, generate gains well into the children's adult years (Schweinhart, 2005; Currie, 2006; Mervis, 2011). Although the studies of larger programs, including Head Start, have generated somewhat mixed results, the general research consensus is that high quality early childhood programs are crucial for both the cognitive and non-cognitive development of children (Barnett, 2011). Consequently, both the federal government, through its investments in Head Start and Early Head Start, and many states have been investing in early childhood programs.

Regardless of how effective those early childhood programs may be, their effectiveness is likely to be diminished if the program participants attend poor quality elementary schools. Indeed, researchers (e.g. Currie and Thomas, 2000) have implicated poor school quality experienced by black children as an explanation for the “fade out” of the effects of Head Start on black children but not for white children. In the present paper, we examine the possibility that investments in early childhood programs may be weakened by elementary school practices that lead to weaker teachers within a school being assigned to the lower grades.

Concerns of this type provided the immediate motivation for this paper, which is based on North Carolina data. The state of North Carolina has been investing heavily in early childhood programs – in the form of the state’s highly touted Smart Start Initiative for children aged zero to five since the early 1990s and its More at Four pre-kindergarten program since the early 2000s. The concern is that the positive effects of those programs may be being dissipated as the children enter elementary schools not only because the schools themselves may be weak but also because the schools may be assigning their weaker teachers to the youngest children in the schools. Although we are in not in a position in this paper to shed light on the larger issue of program dissipation, we are able to examine the extent to which schools are making teacher assignment decisions of the type hypothesized. Hence, the first purpose of this paper is to simply examine the extent to which North Carolina elementary schools assign their weaker teachers to the lower grades, and, if they do so, to determine how the practice differs across groups of schools defined by the disadvantage of their students.

A second, closely related purpose is to examine the extent to which test based accountability for schools is implicated in any shortfalls in teacher quality in the lower grades relative to the upper elementary school grades. Because school based accountability programs are typically based on student test scores starting in grade three, such programs give school principals intent on maximizing the measured performance of the school powerful incentives to assign their stronger teachers to the upper grades. The availability of teacher data for North Carolina over the extended period, 1995 to 2009, allows us to examine how changes in accountability regimes – starting from no accountability in the early years, to a state based accountability program between 1997 to 2002, to the Federal No Child left Behind program starting in 2003 – have affected the within- school distribution of teachers.

This component of the analysis contributes to two literatures. One is the growing literature on the unintended side effects of accountability programs. Critics have pointed, for example, to how test based accountability programs can narrow the curriculum and, in situations of extremely high stakes, lead teachers to cheat (Jacob and Levitt, 2004). Researchers have shown that schools have responded to high stakes accountability by identifying more students as special needs to get them off the testing rolls (Jacob, 2005; Cullen and Reback, 2006; Figlio and Getzler, 2006), altering their disciplinary decisions to keep some low performing students from being tested (Figlio, 2006), and changing their nutrition policies to enhance test results (Figlio and Winicki, 2005). A North Carolina study has also documented that the state's school based accountability program exacerbated the problems that low performing schools face in retaining teachers (Clotfelter et al., 2004). The present paper provides evidence of another unintended side effect of test-based accountability – the potential that it reduces the quality of teaching provided to children in the early elementary grades relative to what it would be without high stakes accountability.

This paper also contributes to the small and emerging literature, which is discussed further below, on the extent to which school leaders are using data from student test scores to make staffing and other resource decisions within schools. Although school leaders and districts may well use test score data in ways that would enhance the quality of the school, our focus in this paper highlights its potential to be used to the disadvantage of some students, namely those in the non-tested grades.

The concept of teacher quality is central to our analysis. Because we are focusing on all grades in elementary schools including the lower grades where the children do not take state tests, we cannot estimate test-based value added measures, which, for better or for worse, have become the standard approach for measuring the effectiveness of teachers who teach in tested grades and subjects. Instead, we use as our proxy measure of teacher quality the average licensure test scores of each teacher. Only in one section – when we are looking at the probability that a teacher will be moved from the upper grades to the lower grades – are we able to use a value added measure of teacher effectiveness.

Based on the analysis presented below, we conclude that in elementary schools, the licensure test scores of teachers in the lower grades are typically below those in the upper grades, and the pattern is true throughout the 1995-2009 period. Contrary to our initial concern that the differences might be particularly stark in the schools serving disadvantaged students, however, we find that they tend to be larger in the more advantaged schools. At the same time, we find that strategic responses of school leaders to the accountability pressures associated with No Child Left Behind have increased the shortfall of teacher quality between the lower and the upper grades, and that the increase has been greatest in the most disadvantaged schools.

Conceptual Framework and Prior Research

The distribution of teachers among grades within a school reflects decisions by school principals that are based, not only on their educational goals for the school, but also on the preferences of individual teachers. In particular, the outcome depends on a variety of decisions – placement decisions at the time of initial hiring, subsequent decisions about moving teachers between grades once teacher effectiveness is revealed to teachers and principals, and decisions by teachers to leave the school.

No High Stakes School Based Accountability

In the absence of high stakes accountability based on student test scores, teacher preferences may or may not play a role in how teachers of different quality are distributed among grades. One can imagine individual teachers coming into a school with preferences to teach particular grades. However, those preferences will only affect the distribution of teacher quality if those who prefer to teach at one level, for example the upper level, are higher quality than those who prefer the other level. Once teachers have taught for a year or two those who are not successful, and hence arguably of lower quality, in the higher grades might prefer to move to a lower grade, while those who are successful in the low grades might prefer to move up. Although that type of movement would push in the direction of having the higher quality teachers in the upper grades, it is difficult to know how common it is or the extent to which it is more prevalent in the more or less advantaged schools.

Nor, in the absence of accountability pressures, do we have a clear prediction about where school principals would like to place their stronger teachers. On the one hand, principals may view the lower grades as the foundation for the upper grades and may place their stronger teachers in those grades. On the other, principals who view the upper grades as more challenging and more important for children's success might place their stronger teachers in those grades. Moreover, the success of principals in implementing their chosen strategy will be affected by their power to attract and retain quality teachers. Because schools with large proportions of disadvantaged students have difficulty filling slots at any grade, principals of such schools may have less power to place teachers in specific grades than principals of more advantaged schools.

The bottom line is that in the absence of accountability, it is difficult to predict whether the lower or upper elementary school grades are more likely to have the stronger teachers, especially in schools serving disadvantaged students. In more advantaged schools, the combination of teacher preferences and the flexibility that principals have in hiring teachers could well lead to the stronger teachers being placed in the upper grades. Ultimately, how teachers are distributed among the grades is an empirical question.

High Stakes Accountability Based on Student Test Scores

The introduction of a high stakes school accountability system based on student test scores is likely to change the outcome in relatively predictable ways, not so much because of teacher preferences but rather because of the strategic behavior of school principals. Because students are not typically tested until third grade, teachers in the untested grades of K-2 face fewer direct pressures to raise student test scores than those in the tested grades of 3-5. Moreover, the success of the school as a whole depends primarily on the effectiveness of the teachers in the upper grades.

In this case, one might expect some teachers to prefer the lower to the upper grades. One possibility is that those who prefer the lower grades are the weaker teachers who are uncomfortable with the pressures associated with the accountability system or who would prefer not, or are not able, to change their mode of teaching to respond to the accountability system. In this case, such preferences would push in the direction of the weaker teachers being in the lower grades and the stronger teachers in the upper grades. Working in the other direction is the possibility that it is the stronger teachers who would prefer to teach in the lower grades in order to avoid the pressures facing teachers in the upper grades. While those pressures may be particularly salient in schools serving large proportions of disadvantaged students which often have a lower capacity to succeed in raising test scores, the ability of the stronger teachers to affect how they are deployed within a school could well be greater in the more advantaged schools. In any case, to the extent that stronger teachers have more alternatives either within or outside the teaching profession or that openings created by departures are filled by novice or otherwise weak teachers, the outcome could well be that the stronger teachers end up in the lower grades.

In contrast, the introduction of a high stakes accountability system changes the incentives facing school principals in a clear and predictable direction. With annual pressure on the school to raise the test scores of its students, principals have strong incentives to make sure their best teachers are in the high stakes grades, even if that means weakening the quality of the teachers in the lower grades. Principals could achieve this goal though some combination of placing the strongest new hires in the upper grades, moving weak teachers from the upper to the lower grades, or moving stronger teachers from the lower grades to the upper grades.

The strength of the incentives and the ability of principals to respond to those incentives are likely to differ by type of school. One might expect the incentives to be stronger, for example, in schools that historically have not met the required achievement levels, that is, those serving disadvantaged children. But principals are more likely to be able to respond to incentives in schools with adequate capacity to meet the standards, which are typically not the most disadvantaged schools. Finally, some principals may be more constrained than others in their ability to place their more effective teachers in the upper grades. In particular, principals in schools serving large proportions of disadvantaged students may have insufficient market power in the teacher labor market to keep higher quality teachers in the upper grades if the teachers do not want to teach in those grades. Sometimes assuring that there is a warm body in the classroom trumps consideration of quality.

These considerations lead us to the hypothesis that the introduction of a test based accountability system is likely to reduce the quality of teachers in the lower grades relative to the higher grades of elementary school. Less clear, however, is the predicted differential effect of the accountability system on elementary schools serving different types of students.

Previous Research on Accountability and Within-School Staffing Patterns

One early study examined test based accountability in New York State at a time when the only elementary school grade with high stakes testing was fourth grade (Boyd et al, 2008). The authors hypothesized that teachers would seek to avoid the high stakes grade for multiple reasons – fear of unwanted scrutiny, loss of flexibility in the classroom, pressure to teach to the test, and concern about their jobs – and that the stronger teachers would be more successful in avoiding such classrooms than other teachers. The authors recognized, however, that some teachers might prefer the high stakes environment to ones in which there was little or no attention to whether students were learning.

Because their hypotheses focused on the preferences of teachers alone and not the strategies of principals, the authors were surprised to find that the turnover rate among teachers in fourth grade decreased relative to that of teachers in other elementary schools grades after the introduction of the high stakes test. In addition, the evidence suggested that in some cases high ability teachers were less likely than others to leave fourth grade. Further, they found that newly hired fourth grade teachers were less likely to be novice teachers and more likely to have attended a highly competitive undergraduate institution than teachers entering other grades. These findings are fully consistent with the predictions of principal responses to high stakes testing mentioned above.

The one finding consistent with their initial hypothesis about teacher preferences was that more experienced teachers behaved somewhat differently than the less experienced teachers with respect to attrition. The authors interpreted this finding as evidence that compared to the newer teachers, some of the more experienced teachers were "less willing to change their teaching styles or curricula to fit testing requirements" (Boyd et al. 2008, pp. 107-108). Of interest is that the differential by experience was concentrated in the high achievement schools.

The strategic behavior of elementary school principals in the context of accountability plays a far more central role in the hypotheses examined in two more recent papers. The first is a qualitative study in which the author examines the extent to which school leaders are using student test scores to allocate resources within schools (Cohen-Vogel, 2011). The study is based on close analysis of staffing practice in 10 elementary schools, one high performing and one low performing school in five school districts in Florida. Based on the reports of school principals, the author finds that principals are in fact making “evidence-based” staffing decisions. In particular, they use student test scores to identify grades and subjects in which students are not doing well and make hiring and staffing decisions to shore up those areas. When reassigning teachers among grades, the principals reported paying attention to teacher effectiveness. The principals, in some cases, explicitly talked about moving ineffective teachers from tested grades to lower grades. For example, one principal reported:

> “If I know a teacher is really good, and since third, fourth, and fifth grades are the grades you have the FCAT [Florida’s high stakes test] tests, and I really need a stronger teacher there, I will switch people around. “ Cited in Cohen-Vogel, 2011, p. 494.

And teachers in the same school reported:

> “Last year they did a lot of reassigning. They took a couple of teachers that were in the higher [grade] levels and moved them to the lower levels. The rationale? You know, those that had good skills could move up to the higher grades and the students would benefit from that, and those that might have been lacking went down to the lower grades.” Cited in Cohen-Vogel, 2011, p. 494.

In contrast to, but complementary to that study, the second study is a quantitative analysis of the career paths of 25,000 Florida teachers initially in grades four through eight. The study is designed to determine how schools make promotion and reassignment decisions in response to teacher effectiveness as measured by success in raising student test scores (Chingos and West, 2011). Instead of relying on principal self-reports, this study looks at their actual behavior. Of interest is the trajectories teachers follow during their careers, including being on track to become school principals, becoming reading or math coaches, remaining in high stakes classroom positions or moving to low stakes teaching positions. For those remaining in elementary schools, low stakes teaching positions include those in grades K-2. Most relevant for the present study is the authors' finding that those teachers who were "demoted" to low stakes classrooms were consistently less effective classroom teachers than those who were retained in the high stakes classrooms or promoted to administrative positions.

Although the conclusions of both Florida papers are consistent with our predictions about how accountability would affect school staffing outcomes, neither is able to attribute the patterns they find explicitly to accountability because they have no pre-accountability data.

Context, Data, and Approach

The analysis in this paper is based on North Carolina data for the years 1995-2009. North Carolina is a particularly good state for this research because it has been administering statewide tests to students in grades 3-8 since the 1992-93 school year, and it implemented its own sophisticated school-based accountability program (the ABCs) before the introduction of the federal No Child Left Behind (NCLB) program. With data made available to us through the North Carolina Education Research Data Center at Duke University, we are able to examine the within-school distribution of teachers from the 1994-95 school year (henceforth 1995) to the 2008-2009 school year (henceforth 2009). This period covers two pre-accountability years (1995 and 1996), six years of the ABCs program (1997 to 2002), and seven years of NCLB (2003 to 2009). We note that the ABCs program has co-existed with the NCLB program throughout the latter's existence.

Accountability Regimes

The North Carolina ABCs accountability program was part of a broader state effort to improve the academic performance of the state's children throughout the 1990s. If a school raised student achievement by more than was predicted for that school, all the school's teachers received financial bonuses – $1500 for achieving high growth and $750 for meeting expected achievement growth.[1] Schools that did not meet their expected growth target are identified as such and in some cases subject to intervention from the state. Although the teacher bonuses are based solely on the growth in student achievement, the ABCs program does not completely ignore levels of achievement. In addition to their rankings based on achievement growth, schools also receive various designations based on the percentages of students meeting grade level standards, such as schools of excellence, schools of distinction, and priority schools. However, these designations carried no financial benefit. In addition, schools are designated as "low performing" if they meet neither their school-specific growth expectation nor the state's performance standard of a 50 percent passing rate.

The federal government started holding schools accountability for student achievement with the 2001 reauthorization of the federal Elementary and Secondary Education Act, called No Child Left Behind. This law, which became effective in the 2002-2003 school year, effectively means that each school faces an annual target defined in terms of achievement *levels* rather than in terms of achievement *gains* as under the state accountability system. To make sure schools are on track toward the ultimate goal of 100 percent proficiency by 2014, NCLB assesses schools on the basis of whether their students are making adequately yearly progress (AYP). Failure to meet AYP brings with it a variety of consequences, including allowing children to move to another school and requiring districts to use their federal Title 1 grants to pay for supplemental services. After five years of failure, a school is subject to state takeover. A school that performs well under the state's growth based accountability system may do poorly under the federal system and vice versa.

Accountability systems are designed to change the behavior of school leaders in all schools, not just those in schools that fail to meet the requirements in any one year. Even those schools that

[1] For the first several years of the program, schools were divided into four categories. Exemplary, meets expectation, no recognition, and low-performing. Subsequently, the name of the "exemplary" category, which refers to schools exceeding their growth targets by more than 10 percent, was changed to "high growth."

successfully meet the standards one year must remain vigilant lest they fail to meet them the following year. As a consequence, in this study, we are far less interested in how accountability regimes affect the staffing patterns in individual schools based on their accountability status in the previous year than we are in the effects of each accountability regime averaged across all elementary schools or across groups of schools defined by the characteristics of their students. Only in the final section do we report any results for individual school accountability status.

Measures of Teacher Quality

We use detailed data on teachers, for whom we have information on various qualifications and credentials, including their licensure test scores. Although we also have data on student test scores for elementary school teachers of math and reading in the upper grades, we can only use those data for a small portion of our analysis because students in the lower grades are not tested. We include in our analysis teachers in all non-charter public schools that serve grades K-5. There were 1285 such schools in 2009, up from 1016 in 1995.

Table 1 provides some initial descriptive analysis of average teacher credentials in the lower and the upper grades, for 2009 in Panel A and for 1995 in Panel B. The table includes a number of credentials that have been widely used in the literature on teachers, with some of them more appropriate proxies for teacher quality than for others. They are all defined so that, to the extent the measures are reasonable proxies for quality, larger numbers represent higher quality. The differences – either across all schools or, of more relevance for our purposes, within schools – are defined as the value in the lower grades minus the value in the upper grades. Hence, a negative difference indicates a shortfall in teacher quality in the early grades.

The consensus in the literature is that teachers with three or more years of experience are on average more effective in raising test scores than those who have limited experience (see summary in

Goldhaber, 2008). By this relatively aggregated experience measure, it appears that the lower grades had a small advantage in 2009, a pattern that does not hold for any of the other credentials in either of the years.[2] The more typical pattern is for the lower grades to have a quality disadvantage relative to the upper grades. The pattern is true for the percentage of teachers with master's degrees, their average licensure test scores and, the proportions of teachers with elementary education licenses or with National Board Certification. For the bulk of our analysis, we use licensure test scores of the teachers as our measure of teacher quality.

We focus on this single measure largely because the research shows that teachers' test scores are the credential that most consistently emerges as predictive of student achievement across studies of various types (see summary in Goldhaber 2008). Moreover, this measure has the advantage of being a continuous variable which makes it far less lumpy than measures such as experience or characteristics that are measured as a percentage of teachers. For example if a school has only six teachers in the lower grades, three of whom leave in a single year, the proportion of experienced teachers could potentially fall from 100 percent in one year to 50 percent the next year if all the openings were filled with inexperienced teachers. We rejected master's degree as a measure because recent studies show that master's degrees are not predictive of student achievement at the elementary level (Clotfelter, Ladd and Vigdor 2006,2007).[3] While most careful studies, including several based on North Carolina data, show that National Board Certified teachers are more effective at raising student achievement than are those who are not certified (Goldhaber and Anthony 2007; Clotfelter, Ladd, and Vigdor 2006, 2007, 2010), the fact that there were no certified teachers in the 1990s, plus its lumpy nature, rule it out as a useable measure of teacher quality for this study. Finally, the percent of elementary teachers who are licensed

[2] The fraction of novice teachers generates the same picture in that the percentage of teachers who are novices is slightly higher in the upper than in the lower grades in 2009. The 2009 percentages are below 2 percent in both grade categories, however, which suggests that the observed differences may not be informative.

[3] Another measure used in many studies is the quality of a teacher's undergraduate institution, as typically measured by Barron's College ratings. Research using North Carolina data confirms its predictive power at the high school level but not at the elementary level (Clotfelter, Ladd, and Vigdor 2006, 2007, 2010).

as elementary education teachers, though very low in the 1990s, is now sufficiently close to 100 percent to provide little information on differences across the grade levels.

<Insert Table 1>

In sum, with the one exception of teacher experience in 2009, all of the measures show that, for the set of all elementary schools in the state, teachers in the lower grades have weaker credentials than those in the upper grades. In the following sections, we explore the patterns for our preferred proxy for teacher quality by type of school and over time and examine the extent to which accountability pressures have affected the patterns.

Results

School Characteristics and Accountability Regimes

We begin our analysis by using ordinary least squares (OLS) models to look at differences across the whole period between teacher test scores in lower and upper elementary school grades within schools. For Tables 2 through 4, the dependent variable is the within-school difference in average standardized teacher licensure test scores between the two sets of grades. As in Table 1, negative coefficients indicate that teacher quality in the lower grades falls short of that in the upper grades and positive coefficients indicate that the quality of the teachers in the lower grades exceeds that in the upper grades.

Table 2 reports basic descriptive results for the full period, with no specific attention to the role of accountability. We remind the reader, however, that schools were subject to accountability for 13 of the 15 years studied. The negative coefficient of -0.083 in the first column indicates that the average test scores in the upper elementary grades exceed those in the lower grades by 8.3% of a standard deviation.

The following columns show how the patterns differ by schools divided into quintiles based on three different categories of student disadvantage: percent minority students, percent free or reduced lunch students, and performance composite. The performance composite for each school represents the percent of all test scores at the school that met proficiency standards in the previous year and has been reported for each school annually since 1997. Because the performance composite is not available for the early years of our period, the sample size of schools in the fourth column is smaller than for the other columns. Not shown in Table 2 is the fact that the schools within the more disadvantaged quintiles have teachers of lower average quality and greater variance among teachers than those in the more advantaged quintiles (see appendix A for detailed descriptive statistics). This greater variance among teachers in the disadvantaged schools suggests there may be opportunities for larger differences between grade sets in those schools.

Across all three sets of quintiles, the estimated coefficients are all negative and statistically different from 0. Interestingly, however, the quintiles for the more advantaged schools (the lowest quintiles in each category) exhibit the largest differences across grades. For schools in the lowest minority quintile, for example, the teacher test scores in the lower grades fall short of those in the upper grades by 10.0% of a standard deviation, while in the highest minority quintile, the shortfall is only 2.6% of a standard deviation. Similar patterns emerge for the other two sets of quintiles. This pattern is consistent with the hypothesis that the principals of the more advantaged schools are more strategic or have more market power to place their teachers in selected grades than do the less advantaged schools. Without further analysis by accountability regime, however, one should not attribute the patterns specifically to accountability pressures.

<Insert Table 2>

Figure 1 displays the information in Table 2 graphically. In this visual representation, it is clear that the shortfall in teacher test scores in the lower grades occurs across all types of schools but is much larger in the more advantaged schools than in the more disadvantaged schools.

To shed light on the role of accountability pressures, we next examine how the patterns of teacher test scores differ across time and across the three accountability regimes. Table 3 reports the within-school differences in test scores for each year in our time period, from 1995 to 2009, and for the three accountability regimes: no accountability, ABCs and NCLB.

The negative and significant coefficients in the first column indicate that teacher quality in the lower grades falls short of that in the upper grades in every year. Beginning in 2003, however, the shortfall in teacher quality becomes significantly larger than it was from 1995 to 2002. Similarly, the second column displays negative coefficients for all three accountability regimes, but the shortfall in teacher quality in the lower grades is significantly larger under NCLB than before any accountability system was implemented or under the ABCs alone. Moreover, the results in the third column indicate that it is the presence of the NCLB pressures, and not the effect of other time-varying factors that account for the greater shortfalls in the post 2002 period.

These patterns provide clear support for the hypothesis that incentives created by the NCLB accountability system increased the strategic placement of the stronger teachers in the upper grades. The fact that teacher placement seems to change more under NCLB than the ABCs suggests that schools respond more strongly to negative sanctions, such as those posed by NCLB, than to the positive rewards offered by the ABCs. Accountability pressures, however, cannot explain the finding that teacher quality falls short in the lower grades in the early years before the introduction of either accountability program. It could be that higher quality teachers simply prefer to teach in the upper elementary grades or parents may be better at judging teacher quality in these grades than in the lower grades.

<Insert Table 3>

The next set of analyses, shown in Tables 4A, 4B and 4C, explore whether accountability systems have different effects on the strategic placement of teachers across the three sets of quintiles introduced in Table 2.

Table 4A reports patterns for school quintiles defined by the percent of minority students in the school. The first column shows that before the introduction of any accountability system the shortfall in teacher quality in the lower grades is only statistically different from zero in the second lowest minority quintile. However, after the introduction of the ABCs system, the shortfall, shown in column two, is larger and statistically significant in all but the highest minority quintile. Under NCLB, the shortfall in teacher quality in the lower grades is statistically significant across all the quintiles.

<Insert Table 4A>

The patterns for quintiles by percent free or reduced price lunch, shown in table 4B, are very similar. The shortfall in teacher quality in the lower grades goes from being statically significant only for the most advantaged schools to being significant for all but the most disadvantaged to being significant in all quintiles. Since performance composites were not calculated prior to 1997, it is not possible to place schools into performance composite quintiles during the pre-accountability period. However, the results shown in Table 4C for the period after the introduction of ABCs and the period after NCLB mirror the patterns found for minority quintiles and free/reduced price lunch quintiles.

<Insert Table 4B>

<Insert Table 4C>

Importantly, the shortfall in teacher quality in the lower grades increased as much or more in response to accountability pressures in the disadvantaged schools than in the advantaged schools. Thus, the principals in all schools, not just those in the advantaged schools responded to accountability pressure in strategic ways. The main difference between the various types of schools reflects the differences in the pre-accountability period when the advantaged schools much more clearly favored the higher grades.

Figure 2 displays the pattern of shortfalls across the three accountability regimes for all schools and for schools with differing proportions of low income students as defined by their eligibility for free or reduced price lunch. The figure clearly illustrates the increased size of the shortfall in teacher test scores under the accountability regimes. It also shows that the increase is larger for the disadvantaged schools. Although Figure 2 depicts results only for the free or reduced lunch quintiles, the patterns are very similar for minority quintiles and performance quintiles.

Movement of Teachers

One of the mechanisms schools can use to place the best teachers strategically is to move teachers between grades. In this section, we use logistic regressions to look at the relationship between a teacher's qualifications and a teacher's probability of moving down from the upper grades to the lower grades or up from the lower grades to the upper grades. The outcome variables in this section are an indicator for whether a teacher who taught in the lower grades in the previous year moved up and an indicator for whether a teacher who taught in the upper grades in the previous year moved down. All results in this section are expressed as odds ratios so that a value below one indicates a lower probability and a value above one a higher probability of the specified move.

Table 5 reports the relationship between teacher licensure test scores and the probability of moving up or down as well as how this relationship differs across accountability regimes. The odds ratio of 0.958 in the first column indicates that a teacher with a licensure score one standard deviation above the mean is only 95.8% as likely as an average teacher to move down to the lower grades. Similarly, the third column reports a significantly increase in the probability of a teacher with high test scores moving up to the upper grades.

The second and fourth columns illustrate the relationship between accountability regimes and the probabilities that teachers with different test scores move up or down between grade sets. The second column shows no statistically significant impacts of the accountability systems on the probability of teachers moving down, but the fourth column shows that the probability of teachers moving up is greater under both accountability systems than in the pre-accountability period. In addition, the probability of moving up to the upper grades is even greater for teachers with high test scores after the introduction of NCLB, such that a teacher with a test score one standard deviation above the mean is 42.7% more likely to move up as an average teacher.[4]

Thus the evidence shows that accountability has increased the probability that schools will move teachers to the upper grades, and that during the NCLB period, the teachers who were moved up were the stronger ones, which is in keeping with strategic behavior by principals. The finding of a positive relationship between accountability and the tendency for schools to move teachers up may reflect a greater willingness of principals to hire new teachers of unknown quality into the untested lower grades than in the tested grades during the accountability period.

<Insert Table 5>

We have focused the analysis so far on teacher licensure test scores as a measure of teacher quality. We now look briefly at teacher value-added as a supplementary measure of quality. The analyses in Table 6 use teacher value-added calculations to further explore the probability of teachers of differing qualities moving down from the upper elementary grades to the lower elementary grades. We look only at teachers moving down because it is not possible to calculate value-added for teachers in the lower grades for which there are no student test scores. Teacher value-added is calculated separately

[4] The probability of a teacher with a score one standard deviation above the mean moving up under NCLB is calculated by adding the original coefficients on teacher test score, NCLB and the interaction term, then converting the sum to an odds ratio, which is equal to 1.427.

for reading and math using a Bayesian shrinkage estimator (for details, see appendix A) and then rescaled to have a mean of zero and a standard deviation of one.

Table 6 shows that the probability of moving down to the lower grades is substantially smaller for teachers with value-added one standard deviation above the mean compared to teachers with average value-added in both reading , 74.5% as likely, and math, 70.1% as likely. The probability of any teacher moving down is lower under both accountability regimes, and the probability of a teacher with high value-added in math moving down to the lower grades is even further reduced after the introduction of NCLB compared to the pre-accountability period. This means that a teacher with a math value-added 1 standard deviation above the mean is only 60.7% as likely to move down as an average teacher under NCLB.[5]

These results are very similar to those shown in Table 5 for teacher test scores and further support the notion that schools strategically move teachers of lower quality to the lower elementary grades, especially under NCLB.

<Insert Table 6>

School Specific Accountability Status

In this final section, we look at the influence of school specific accountability status on the difference in teacher tests scores between lower and upper elementary school grades. Under an accountability regime, all schools, not just those that have previously failed, are under pressure to produce high test scores. This ongoing pressure is particularly true under a system like NCLB where accountability standards rise over time and schools that previously met standards may fail in subsequent years if they do not raise scores. Given this reality, we do not expect school accountability status to be

[5] The probability of a teacher with a score one standard deviation above the mean moving up under NCLB is calculated by adding the original coefficients on teacher value added, NCLB and the interaction term, then converting the sum to an odds ratio, which is equal to 0.607.

as important as the presence of an accountability regime in affecting the placement of teachers. Because other studies have looked at how a school's prior year accountability status has affected strategic behavior, however, we explore it briefly.

Table 7 includes indicator variables for whether the school failed to meet adequately yearly progress (AYP) under NCLB or Expected Growth under the ABCs in each of the previous three years. The table also includes year indicators to control for changes in the pattern of teacher test score differences under the accountability regimes. The second column also includes controls for school characteristics, including the percent of minority race students, the percent of students receiving free or reduced price lunch, and the performance composite of the school, in order to account for the differences in schools that frequently fail accountability standards compared to other schools.

The positive coefficients for failing to meet expected growth during the ABCs regime in the first column appear to suggest that schools that failed to meet expected growth in one of the previous two years increased the quality of their teachers in the lower grades relative to the upper grades in the subsequent years compared to schools that met expected growth. This runs counter to the expected direction of accountability pressure on the distribution of quality teachers. However, the results in the second column indicate that once we control for the characteristics of the school, the unexpected pattern disappears. At the same time, the results for NCLB status in that column show that principals do seem to be reacting to a failure to meet the AYP standards of NCLB in the most immediate prior year. In particular, they have taken actions that reduce teacher quality in the lower grades by 3.4% of a standard deviation relative to the upper grades. Although the coefficients in column two for failure two and three years previous are also negative, they are far from statistically significant.

Thus, we conclude that a failure of a school to meet AYP in a specific year does seem to generate a short term strategic response. Nonetheless, we emphasize once again that any school-specific estimate is likely an underestimate of the effect of the accountability system on the strategic behavior of school principals given that all schools, not just those who fail to meet AYP in a given year, are subject to accountability pressures. For that reason, we believe the results for all schools in the state, as reported in Tables 3 and 4 above provide the most accurate estimate of the strategic responses by North Carolina Schools to the NCLB program.

<Insert Table 7>

Conclusion

This study was motivated by the concern that teachers within elementary schools may be distributed in a manner that disadvantages students in the lower grades and that test-based accountability systems may exacerbate that pattern because the tests are administered only to children in grades 3-5. The results indicate that concern about teacher quality in kindergarten, first and second grades is warranted as teachers in these grades are of lower quality, as measured by their licensure test scores, than those in the upper elementary grades. Moreover, the findings that accountability, especially of the NCLB form, increases the relative shortfalls of teacher quality in the lower grades and also that schools tend to move teachers of higher quality from the lower to the upper grades and teachers of lower quality from the upper down to the lower grades support the conclusion that accountability pressure induces schools to pursue actions that work to the disadvantage of the children in the lower grades.

In the pre-accountability period, the quality of teachers in the lower grades fell short of that of teachers in the upper grades by a smaller margin in the disadvantaged schools than in the more advantaged schools. At the same time, however, accountability had a more pronounced effect on the distribution of teachers in the more disadvantaged schools. These findings imply that even where accountability programs appear to generate gains in test scores for the tested students, they may be having a negative effect on important foundational skills taught in the early grades, especially for disadvantaged students.

In light of these findings, policymakers should consider implementing policies to bring higher quality teachers to the critical early elementary grades. Additionally, those designing accountability systems should focus more attention on the unintended consequences of accountability for untested students in the lower elementary school grades. Without actions to improve the quality of teachers in the early grades, many of the potential benefits of federal and state investment in early childhood programs are likely to be unrealized.

References

Barnett, W.S. (2011). “Effectiveness of Early Educational Intervention.” *Science,* 333: 975-978.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff (2008). “The Impact of Assessment and Accountability on Teacher Recruitment and Retention: Are There Unintended Consequences? “ *Public Finance Review* 36 (1): 88-111.

Chetty, R. , J. Friedman, N. Hilger, E. Saez, D.W. Schanzenbach, D. Yagin. (2010). “How does your kindergarten classroom affect your earnings? Evidence from Project Star. ” National Bureau of Economic Research, working paper 16381.

Chingos, M.C. and M.R. West (2011). “Promoting and reassignment in public school districts: How do schools respond to differences in teacher effectiveness?” *Economics of Education Review* 30 (2011): 419-430.

Clotfelter, C. T, H, F. Ladd, and J L. Vigdor (2006). “Teacher-Student Matching and the Assessment of Teacher Effectiveness. Journal of Human Resources. vol. 41, number 4, Fall 2006, pp. 778-820. (Also available as NBER Working Paper 11936, January 2006; <http://www.nber.org/papers/w11936>).

Clotfelter, C. T, H. F. Ladd, and J. L. Vigdor. (2007). “Teacher credentials and student achievement: Longitudinal analysis with student fixed effects” *Economics of Education Review,* Dec. 2007.

Clotfelter, C.T, H. F. Ladd, J.L. Vigdor and J. Wheeler (2007). “High Poverty Schools and the Distribution of Teachers and Principals.” North Carolina Law Review, Vol. 85, no. 5 (June), pp. 1345-1379.

Clotfelter, C. T.,H. F. Ladd, J. L. Vigdor, and R. Aliaga Diaz (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 (Spring), 251-271.

Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor, and Justin Wheeler, “High Poverty Schools and the Distribution of Teachers and Principals,” North Carolina Law Review 85 (June 2007), 1345-1379.

Cohen-Vogel, L. (2011). “ Staffing to the Test: Are Today’s School Personel Practices Evidence Based? “ *Educational Evaluation and Policy Analysis.* December 2011. Vol. 33, no. 4: 483-505.

Cullen, J.B. and R. Reback. 2006. “Tinkering toward accolades: School gaming under a performance accountability system,” *Advances in Applied Microeconomics,* Issue 14: 1-34.

Currie, J. (2006). *The Invisible Safety Net: Protecting the Nation’s Poor Children and Families.* Princeton and Oxford: Princeton University Press.

Currie. J. and D. Thomas (2000). “School Quality and the Longer-Term Effects of Head Start.” *Journal of Human Resources*, 35(4): 755-774.

Dynarski, S. J.M. Hyman, and D.W. Schanzenbach (2011). “Experimental Evidence on the Effect of Childhood Investments on Postsecondary Atainment and Degree Completion.” National Bureau of Economic Research, Working paper, 17533.

Figlio, D. (2006). “Testing, crime, and punishment,” *Journal of Public Economics* 90(4), 837-851.

Figlio D. and L. Getzler, (2006). “Accountability, ability, and disability: Gaming the System,” *Advances in Microeconomics,* Issue 14: 35-49.

Figlio, D. and J. Winicki. (2005). “Food for thought? The effects of school accountability plans on school nutrition,” *Journal of Public Economics* 89(2), 381-394.

Goldhaber, D. (2008). “Teachers Matter, but Effective Teacher Quality Practices are Elusive.” In H.F. Ladd and E. B. Fiske, eds, *Handbook of Research on Education Finance and Policy.* New York and London: Routledge.

Hannaway, J. , Z. Xu. T. Sass, D. Figlio, L Feng (2010*). “Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools: Implications for Research, Policy, and Management.” Manuscript, October 19, 2010.*

Jacob, B. (2005). *“*Accountability, incentives, and behavior: The impact of high stakes testing in the Chicago Public Schools,” *Journal of Public Economics* vol 89, Issues 5-6 (June), pp. 761-796.

Mervis, J. (2011). “Giving children a head start is possible—But it’s not easy. *Science,* 333: 956-957.

Schweinhart, L. J. MOntie, Z. Xiange, W.S. Barnetts. C.R. Belfield, and M. Nores (2005). *Lifetime Effects: The High Scope/Perry preschool study through age 40.* Ypsilanti: High Scope Press.

Appendix A: Quintile Descriptive Statistics

Average of School Level Means and Standard Deviations across Quintiles

	Average School Mean	Average School Standard Deviation
Overall	-0.018	0.779
Minority Quintiles		
Lowest	0.077	.724
Low	0.079	.756
Medium	0.036	.781
High	-0.031	.799
Highest	-0.256	.833
Free or Reduced Price Lunch Quintiles		
Lowest	0.134	0.760
Low	0.060	0.764
Medium	0.022	0.769
High	-0.065	0.785
Highest	-0.244	0.820
Performance Composite Quintiles		
Lowest	0.139	0.754
Low	0.051	0.761
Medium	0.004	0.768
High	-0.065	0.788
Highest	-0.208	0.824

Appendix B: Value-Added Calculations

We started with regression of standardized student test scores without teacher fixed effects.

$$Y_{ijt}=X_{it}+\theta_j+\lambda_t+\tau+e_{ijt}$$

Y_{ijt}= student i's score with teacher j in year t
X_{it}= vector of student characteristics in year t
θ_j= teacher fixed effect
λ_t= year fixed effect
τ= school fixed effect

The residuals from this regression are composed of 3 parts:

$$e_{ijt}=\theta_j+\eta_{jt}+\varepsilon_{ijt}$$

θ_j= persistent teacher effect
η_{jt}= classroom error
ε_{ijt}= student error

We then calculated average residual for each class which is composed of the teacher effect, classroom effect, and average of student errors which should be equal to zero if students within a classroom are uncorrelated. Next σ_θ^2 is calculated by taking the average of the product of the average classroom residual and the average classroom residual for all other classes taught by the same teacher. Since the student and classroom portion of the error term are uncorrelated across classrooms, this isolates the teacher portion of the error variance

$$\sigma_\theta^2=\frac{\sum_{j=1}^{J}\sum_{t=1}^{T_j}\bar{e}_{jt}\bar{e}_{jt'}\,N}{}$$

J=number of teachers
T_j= number of classes taught by teacher j
N= number of same teacher pairs

We then calculate σ_ε^2, the variance of student residuals, as the variance of the difference from classroom means.

$$\sigma_\varepsilon^2=\mathrm{var}(e_{ijt}-\bar{e}_{jt})$$

The classroom variance, σ_η^2, is calculated as the difference between the variance of residuals and the student and teacher components.

We calculate weights for each classroom based on classroom errors, student errors & classroom size.

$$w_{jt} = \frac{1}{\sigma_n^2 + \frac{\sigma_\varepsilon^2}{n_{jt}}} * \left(\sum_{t=1}^{T_j} \frac{1}{\sigma_\eta^2 + \frac{\sigma_\varepsilon^1}{n_{jt}}} \right)^{-1}$$

For each teacher, a weighted average of classroom-averaged residuals is created. By using classroom weights we are ensuring that small classrooms are not unduly influencing the teacher averages.

$$\tilde{e}_j = \sum_t w_{jt} \bar{e}_{jt}$$

The variance of the teacher average, var($\tilde{e}_j$), is calculated:

$$var(\tilde{e}_j) = \sigma_\theta^2 + \left(\sum_{t=1}^{T_j} \frac{1}{\sigma_\eta^2 + \frac{\sigma_\varepsilon^1}{n_{jt}}} \right)^{-1}$$

Then, we scale $\tilde{e}_j$ by the scaling factor below. This adjustment reduces the teacher average for teachers that have taught few classes or particularly small classes to account for the tendency of small sample sizes of students to lead to more extreme value-added scores.

$$\frac{\sigma_\theta^2}{var(\tilde{e}_j)}$$

Appendix C: Data Description

The data in this study consists of North Carolina schools administrative data provided by the North Carolina Education Research Data Center (NCERDC) housed at Duke University. The administrative data consists of an individual record for each teacher in each year that they taught in the state. The records include teacher qualifications, including years of experience, highest level of education completed, licensure information, undergraduate institution, and national board certification. Licensure test scores were normalized to a mean of zero and a standard deviation of one based on the type of test and the year the test was completed. For teachers with more than one test score, the normalized scores were averaged. The administrative records also include information on the placement of teacher including the school where the teacher was assigned and the type of assignment. These teacher records were combined with administrative records of teachers assigned to specific courses in each school and the characteristics of students in these courses in order to determine the grade levels of students taught by each teacher in each year.

The administrative data provided by NCERDC also includes testing records for all students who completed state tests in each year. These records were used to match students to math and reading teachers in grades three through five. While the testing records do not identify the teacher for each classroom, they do identify the exam proctor and using a multistep process, this information was used to match at least 75% of students to their teachers in all years except 2005 when match rates were around 64%. The steps in the matching process were : First, if the teacher who proctored the End of Grade test for the student was a valid reading or math teacher in the year of the test, the proctor was assumed to be the teacher of the student in that subject. Second, if a single teacher taught at least 95% of students in reading or math in the relevant grade at the school in the relevant year, that teacher was assigned to the student as the teacher in the relevant subject. Finally, using class composition numbers, the composition of the classes taught by the teacher and the class proctored in the test by the teacher were compared for total enrollment, the number of male students, the number of female students, the number of white students, and the number of nonwhite students. If square root of the sum of squared percentage differences across the five categories was less than or equal to .125, the proctor was assumed to be the correct teacher for the students for whom they proctored the exam.

Table 1 Credentials of North Carolina Teachers in Lower and Upper Elementary Grades, 2005 and 1995.

	N	Experienced Teachers	Master's Degree	Licensure Test Score	Elementary Ed License	National Board Certification
PANEL A. 2009						
Lower	13,827	90.5%	27.8%	-0.008	96.3%	8.2%
Upper	12,350	88.3%	31.7%	0.090	97.4%	9.9%
Difference		2.2%	-3.9%	-0.098	-1.1%	-1.1%
Within School Difference	1,285	2.7%	-3.9%	-0.100	-1.1%	-1.2%
PANEL B. 1995						
Lower	9,507	85.7%	25.7%	-0.079	35.6%	0.0%
Upper	8,590	86.5%	28.3%	-0.016	39.5%	0.0%
Difference		-0.8%	-2.6%	-0.063	-3.9%	0.0%
Within School Difference	1,016	-1.0%	-2.7%	-0.059	-4.3%	0.0%

Note: Experienced teachers are those with three or more years of experience. Teachers' licensure scores are the averages of one or more Praxis tests taken by the teacher, with each test normalized to a mean of 0 and a standard deviation of 1 by year of test based on all teachers who took the test, not just those in our sample.

Table 2. Differences in Licensure Test Scores between Lower and Upper Elementary by School Characteristics, 1995-2009

	Basic	Minority Quintiles	Free/ Reduced Lunch Quintiles	Performance Composite Quintiles
Constant	-0.083*			
	(0.003)			
Lowest Quintile		-0.100*$^{+}$	-0.094*	-0.105*
		(0.007)	(0.008)	(0.008)
Low Quintile		-0.126*	-0.097*	-0.108*
		(0.008)	(0.008)	(0.008)
Medium Quintile		-0.088*	-0.106*	-0.116*
		(0.007)	(0.008)	(0.008)
High Quintile		-0.073*$^{+}$	-0.084*	-0.084*
		(0.007)	(0.008)	(0.008)
Highest Quintile		-0.026*$^{+++}$	-0.033*$^{+++}$	-0.027*$^{+++}$
		(0.008)	(0.008)	(0.008)
Observations	16,311	16,311	15,349	12,820
R-squared	0.000	0.042	0.040	0.047

Note: * indicates that the coefficient is significantly different from zero at the <.01 level. + indicates that coefficient is significantly different from the coefficient on the first quintile at the +<.05, ++<.01, and +++<.001 level. All quintiles run from most advantaged to most disadvantaged with the first quintile having the least minority or free/reduced lunch students and the highest performance composite.

Table3. Differences in Licensure Test Scores between Lower and Upper Elementary over Time and Across Accountability Regimes, 1995-2009

	(1)	(2)	(3)
Pre-Accountability		-0.056*	-0.059*
		(0.010)	(0.014)
ABCs		-0.073*	-0.060*
		(0.005)	(0.013)
NCLB		-0.097*$^{+++}$	-0.095*$^{++}$
		(0.005)	(0.013)
1995	-0.053*		0.006
	(0.014)		(0.020)
1996	-0.059*		0.000
	(0.014)		(0.000)
1997	-0.060*		0.000
	(0.013)		(0.000)
1998	-0.061*		-0.001
	(0.013)		(0.019)
1999	-0.077*		-0.017
	(0.013)		(0.019)
2000	-0.087*		-0.027
	(0.013)		(0.019)
2001	-0.080*		-0.020
	(0.013)		(0.019)
2002	-0.071*		-0.011
	(0.013)		(0.019)
2003	-0.095*$^{+}$		0.000
	(0.013)		(0.000)
2004	-0.102*$^{++}$		-0.007
	(0.013)		(0.018)
2005	-0.095*$^{+}$		0.000
	(0.013)		(0.018)
2006	-0.092*$^{+}$		0.003
	(0.013)		(0.018)
2007	-0.093*$^{+}$		0.002
	(0.013)		(0.018)
2008	-0.100*$^{+}$		-0.005
	(0.013)		(0.018)
2009	-0.104*$^{++}$		-0.008
	(0.012)		(0.018)
Observations	16,311	16,311	16,311
R-squared	0.038	0.037	0.038

Note: * indicates that the coefficient is significantly different from zero at the <.001 level. + indicates that coefficient is significantly different from the coefficient on the first time period at the +<.05, ++<.01, and +++<.001 level.

Table 4A. Differences in Licensure Test Scores by Minority Quintile and Accountability Regime, 1995-2009

Minority Quintile	Pre-Accountability	ABCs	NCLB
Lowest	-0.054	-0.107***	-0.105***
	(0.029)	(0.011)	(0.011)
Low	-0.114***	-0.119***	-0.134***
	(0.033)	(0.012)	(0.010)
Medium	-0.026	-0.071***	-0.120***
	(0.033)	(0.012)	(0.010)
High	-0.051	-0.063***	-0.086***
	(0.033)	(0.012)	(0.010)
Highest	0.013	0.012	-0.052***
	(0.044)	(0.013)	(0.010)
Observations	970	6,388	8,025
R-squared	0.019	0.037	0.057

Note: *** p<0.001, ** p<0.01, * p<0.05; All quintiles run from most advantaged to most disadvantaged with the first quintile having the least minority students

Table 4B. Differences in Licensure Test Scores by Free/Reduced Price Lunch Quintile and Accountability Regime, 1995-2009

Free/Reduced Lunch Quintile	Pre-Accountability	ABCs	NCLB
Lowest	-0.103***	-0.095***	-0.096***
	(0.025)	(0.012)	(0.013)
Low	-0.043	-0.099***	-0.110***
	(0.028)	(0.011)	(0.012)
Medium	-0.048	-0.101***	-0.119***
	(0.038)	(0.012)	(0.011)
High	0.022	-0.065***	-0.108***
	(0.042)	(0.013)	(0.010)
Highest	0.029	0.020	-0.068***
	(0.056)	(0.013)	(0.010)
Observations	968	6,263	7,191
R-squared	0.022	0.037	0.055

Note: *** p<0.001, ** p<0.01, * p<0.05; All quintiles run from most advantaged to most disadvantaged with the first quintile having the least free/reduced lunch students.

Table 4C. Differences in Licensure Test Scores by Performance Composite Quintile and Accountability Regime, 1995-2009

Performance Composite Quintile	ABCs	NCLB
Highest	-0.099***	-0.108***
	(0.014)	(0.011)
High	-0.113***	-0.106***
	(0.014)	(0.010)
Medium	-0.096***	-0.128***
	(0.014)	(0.010)
Low	-0.071***	-0.092***
	(0.014)	(0.010)
Lowest	0.015	-0.053***
	(0.014)	(0.011)
Observations	4,919	7,901
R-squared	0.039	0.055

Note: *** p<0.001, ** p<0.01, * p<0.05; All quintiles run from most advantaged to most disadvantaged with the first quintile having the highest performance composite.

Table 5. Teachers Moving Up or Down based on Licensure Test Scores and Accountability Regimes, 1995-2009 (Odds Ratios)

	Moving Down	Moving Down	Moving Up	Moving Up
Teacher Test Score	0.958**	1.005	1.166***	1.009
	(0.014)	(0.057)	(0.017)	(0.066)
ABCs		0.960		1.322***
		(0.051)		(0.081)
NCLB		0.964		1.136*
		(0.050)		(0.069)
ABCs*Teacher Test Score		0.919		1.093
		(0.056)		(0.075)
NCLB*Teacher Test Score		0.979		1.245**
		(0.059)		(0.086)
Constant	0.078***	0.080***	0.059***	0.049***
	(0.001)	(0.004)	(0.001)	(0.003)
Observations	99,957	99,957	122,654	122,654

Note: *** p<0.001, ** p<0.01, * p<0.05; Standard errors in this table refer to the original coefficients and not to the odds ratios. An odds ratio of less than 1 should be interpreted as a decrease in the probability of the outcome and an odds ratio of more than 1 should be interpreted as an increase in the probability of the outcome.

Table 6. Teachers Moving Down based on Teacher Value-added and Accountability Regimes, 1995-2009 (Odds Ratios)

	Moving Down	Moving Down	Moving Down	Moving Down
Reading Value-added	0.745***	0.789***		
	(0.009)	(0.044)		
Math Value-added			0.701***	0.778***
			(0.009)	(0.043)
ABCs		0.839**		0.837**
		(0.049)		(0.049)
NCLB		0.885*		0.887*
		(0.051)		(0.051)
ABCs*Reading Value-added		0.934		
		(0.055)		
NCLB*Reading Value-added		0.946		
		(0.056)		
ABCs*Math Value-added				0.914
				(0.053)
NCLB*Math Value-added				0.879*
				(0.051)
Constant	0.059***	0.067***	0.058***	0.067***
	(0.001)	(0.004)	(0.001)	(0.004)
Observations	97,618	97,618	97,647	97,647

Note: *** p<0.001, ** p<0.01, * p<0.05; Standard errors in this table refer to the original coefficients and not to the odds ratios. An odds ratio of less than 1 should be interpreted as a decrease in the probability of the outcome and an odds ratio of more than 1 should be interpreted as an increase in the probability of the outcome.

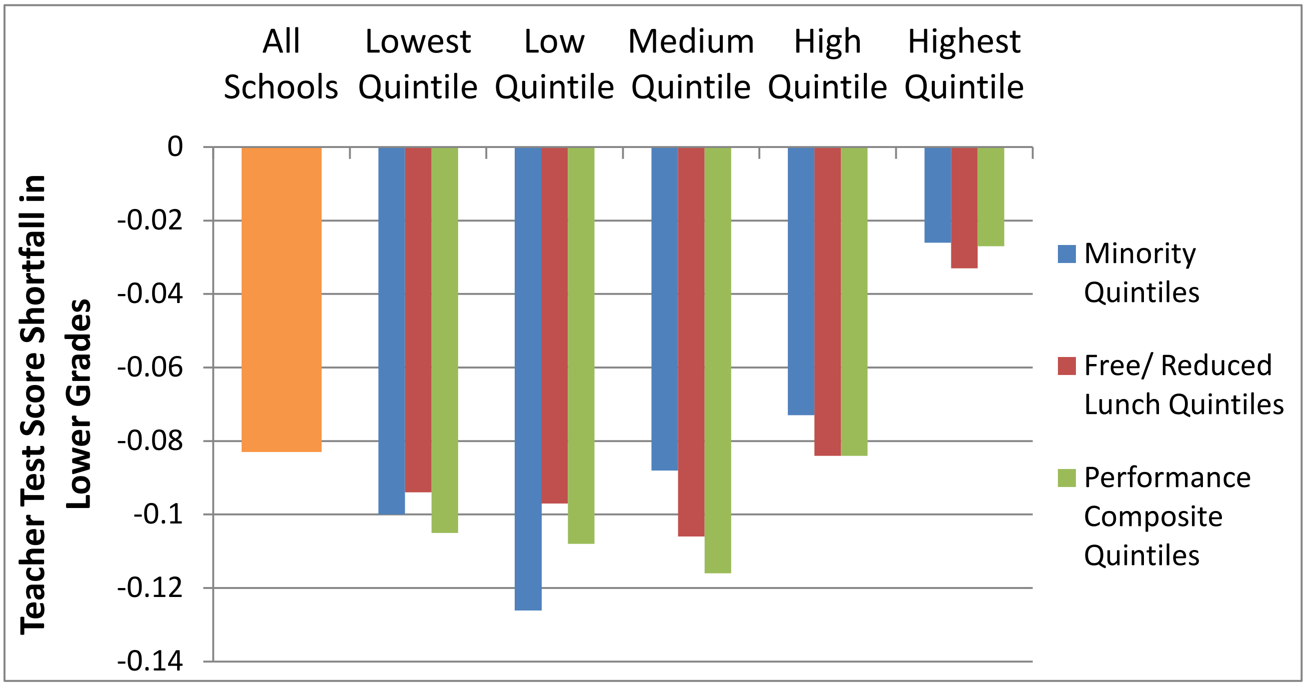
Table 7. Differences in Licensure Test Scores by School Specific Accountability Status, 1995-2009

	(1)	(2)
Failed to Meet AYP 1 Year Previous	-0.010	-0.034**
	(0.012)	(0.013)
Failed to Meet AYP 2 Years Previous	-0.006	-0.017
	(0.013)	(0.014)
Failed to Meet AYP 3 Years Previous	-0.014	-0.023
	(0.015)	(0.016)
Failed to Meet Expected Growth 1 Year Previous	0.030**	0.004
	(0.010)	(0.011)
Failed to Meet Expected Growth 2 Years Previous	0.027**	0.011
	(0.010)	(0.011)
Failed to Meet Expected Growth 3 Years Previous	0.020	0.008
	(0.010)	(0.011)
Percent Minority Students		0.093***
		(0.020)
Percent Free or Reduced Price Lunch Students		-0.024
		(0.027)
Previous Year Performance Composite		-0.002**
		(0.001)
1996	-0.006	0.000
	(0.020)	(0.000)
1997	-0.007	0.000
	(0.019)	(0.000)
1998	-0.019	0.011
	(0.019)	(0.023)
1999	-0.038	-0.010
	(0.019)	(0.021)
2000	-0.051**	-0.015
	(0.019)	(0.021)
2001	-0.042*	-0.005
	(0.019)	(0.020)
2002	-0.037	0.016
	(0.019)	(0.020)
2003	-0.060**	0.000
	(0.019)	(0.000)
2004	-0.057**	0.004
	(0.019)	(0.020)
2005	-0.048*	0.010
	(0.019)	(0.020)
2006	-0.044*	0.016
	(0.020)	(0.021)
2007	-0.057**	-0.010
	(0.020)	(0.020)

2008	-0.063**	-0.014
	(0.020)	(0.022)
2009	-0.055**	-0.017
	(0.021)	(0.024)
Constant	-0.053***	0.025
	(0.014)	(0.065)
Observations	16,311	12,005
R-squared	0.003	0.009

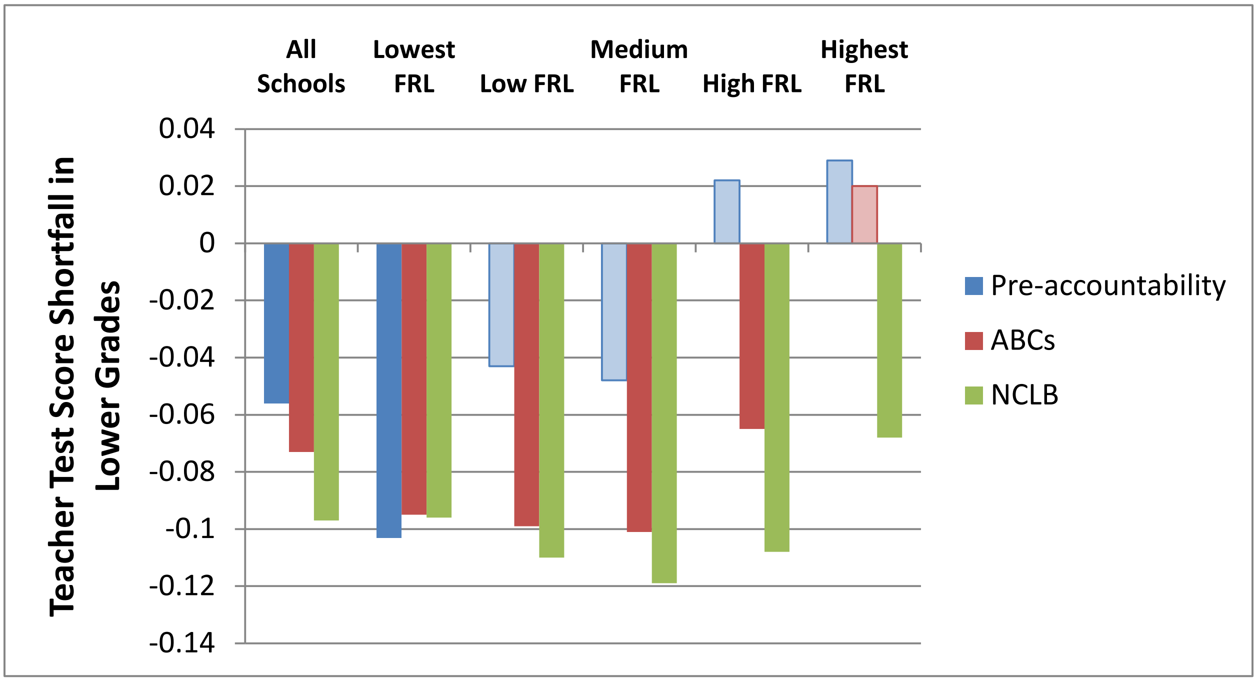
Note: *** p<0.001, ** p<0.01, * p<0.05

Figure 1. Shortfall in Teacher Test Scores in Lower Elementary by School Characteristics, 1995-2009



Note: All coefficients displayed in the table are statistically different from zero.

Figure 2. Shortfall in Teacher Test Scores in Lower Elementary by Accountability Regime, 1995-2009



Note: Lighter colored bars represent coefficients that are not statistically different from zero.