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 Across Grades in Elementary School

SARAH C. FULLER
AND HELEN F. LADD

WORKING PAPER 75 • FEBRUARY 2013

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

Sarah C. Fuller and Helen F. Ladd

CALDER Working Paper No. 75
February 2013

Abstract

We use North Carolina data to explore whether the quality of teachers in the lower elementary grades (K-2) falls short of teacher quality in the upper grades (3-5) and to examine the hypothesis that school accountability pressures contribute to such quality shortfalls. Our concern with the early grades arises from recent studies highlighting how children's experiences in those years have lasting effects on their later outcomes. Using two credentials-based measures of teacher quality, we document within-school quality shortfalls in the lower grades, and show that the shortfalls increased with the introduction of No Child Left Behind. Consistent with that pattern, we find that schools responded to accountability pressures by moving their weaker teachers down to the lower grades and stronger teachers up to the higher grades. These findings support the view that accountability pressure induces schools to pursue actions that work to the disadvantage of children in the lower grades.

Introduction

Many studies have documented differences across schools in the quality of teachers, where quality is typically 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. Lankford, Loeb, and Wyckoff 2002; Clotfelter, Ladd , and Vigdor 2007).[1] 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 poorer 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).

[1] Although it might be tempting to use other measures of teacher quality, such as teacher value added scores, that approach is not feasible because the value added modeling technique requires the analysis to include school fixed effects or other controls for school characteristics.

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, however, their effectiveness is likely to be diminished if the program participants continue on to 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 simply to 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 and 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 who are intent on maximizing the short run 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 possibility 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 two proxy measures of teacher quality: the average licensure test scores of each teacher and a value added based index that incorporates a number of teacher qualifications. 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 straightforward value added measure of teacher effectiveness.

Based on the analysis presented below using both measures of teacher quality, we conclude that teachers in the lower grades are lower quality than those in the upper elementary school grades, although the differences are small in magnitude. 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 but consistent with our conceptual discussion, however, we find that the quality gaps between the lower and upper grades tend to be larger in the more advantaged schools. Importantly, 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 upper grades, and that the increase has been greatest in the schools with average to large proportions of disadvantaged students.

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 affect the distribution of teacher quality within the school only 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 measured success of the school as a whole depends primarily on the effectiveness of the teachers in the upper grades.

In this context, 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. 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. Hence, it is hard to predict how accountability would affect the distribution of teachers across grades if all that mattered were teacher preferences.

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 strong 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 upper 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], and I really need a stronger teacher there, I will switch people around. "(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." (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.[2] Schools that did not meet their expected growth target were identified as such and in some cases subject to intervention from the state. Although the teacher bonuses were 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 adequate 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.

[2] 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."

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

Although it has become increasingly common to measure a teacher's quality by the gains in the test scores of her students, we cannot use that approach in this study because we need measures of teacher quality not only for teachers in the upper grades, for whom student test scores are available, but also for those in the lower grades. As a substitute, we use two proxies for teacher quality based on teacher credentials. One proxy is the teacher's licensure test score and the other is a weighted average of credentials where the weights come from a value-added model. 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 within-school differences are defined as the average across all schools of the value in the lower grades minus the value in the

upper grades within each school. Hence, a negative difference indicates a shortfall in teacher quality in the early grades.

(Table 1 about here)

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.[3] 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 National Board Certification. The final column represents the value-added index of credentials that we describe below. Consistent with most of the individual credentials, this index shows that the lower grades were at a disadvantage relative to the upper grades in 2009, and also to some degree in 1995.

Licensure test scores

The first of our two proxies for teacher quality is a teacher's average licensure test score. 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; Goldhaber 2007). 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

[3] 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.

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).[4] 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 single-item measure of teacher quality for this study.

The great advantage of the teacher test score measure is that it is straightforward and easy to understand. Its major limitation is that it is only one of a larger package of credentials. Unless it is highly correlated with other credentials, low quality on this measure might well be offset by high quality on another measure. Another concern is that the measure might be confounded by changes over time in the types of tests taken by teachers in the early grades relative to those taken by teachers in the higher grades. Although we largely minimize this concern by normalizing all licensure test scores to have a mean of 0 and standard deviation of 1, we recognize that the distributions of test scores by type of test could have changed over time. This possibility is worth considering given that in 1995 the most common test taken by teachers in the lower grades (Early Childhood Education) differed from the most common test taken by teachers in the upper grades (Education in the Elementary School) while by 2009, the most common tests taken by both sets of teachers were two new elementary education tests, one in content area and the other in curriculum and assessment.[5] Fortunately for this study, however, among those

[4] 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).

[5] In 1995, 56.2 percent of the teachers in the lower grades had taken the Early Childhood Education test, while 59.3 percent of the teachers in the upper grades had taken the Education in Elementary School test. Teachers also took a variety of other tests but there was essentially no difference in the percentages of each test taken by teachers as the two grade level. In 2009, the percentages of teachers at the lower and upper levels taking the Elementary Education

teachers for whom we could compare test scores, the scores on the two early year tests were quite highly correlated with each other, and each of the two more recent tests exhibited similar correlations with each of the two earlier tests[6]. The similarity of these correlations mitigates, but does not eliminate completely, concerns about the confounding effects of the change in testing regimes over time. Both for that reason and because it is based on a single credential, we supplement this proxy for teacher quality with a broader measure.

Value added index of teacher credentials

This second measure of teacher quality is essentially a weighted average of several teacher credentials, where the weights reflect the contributions to gains in student test scores. To calculate this measure, we first estimated teacher-specific value added measures using a within-school value added model for math achievement described in appendix A. Because it is based on student test scores that are not available in the lower grades, the value- added model was estimated using data only for teachers in the upper grades. We then used the estimated coefficients from a regression designed to explain the variation in value added across teachers as a function of their credentials as weights to calculate a credentials-based index for all teachers, both those in the upper and those in the early grades (See appendix 2 for the equation). We normalized the predicted measures so that our index has a mean of 0 and standard deviation 1 across teachers. It is worth noting that we lose some of the true variation in teacher quality, even within schools, based on this methodology because, following the literature, we used Bayesian shrinkage procedures to estimate our value added measures, which move

–Content area test were 45.5 and 48.6 respectively and taking the Curriculum and Assessment test were 47.0 percent and 50.8 percent, respectively.

[6] The relevant correlations are as follows. The correlations between the Early Childhood Education test (common in lower grades in 1995) and Education in the Elementary School test (common in the upper grades in 1995) is 0 .726. The correlation between Elementary Education – Content Area (common at both levels in 2005) and Early Childhood Education is 0.492 and Education in the Elementary School is 0.434. The correlation between Elementary Education – Curriculum and Assessment (common at both levels in 2005) and Early Childhood Education is 0.729 and Education in the Elementary School is 0.686.

the estimates for teachers with limited data toward the mean. We then calculated average indices of teacher quality for teachers within the early and late grades at each school. Following the convention we used earlier, the indices are subtracted in such a way that a negative coefficient indicates that on average the early grade teachers within schools are of lower quality than the upper grade teachers.

Results

In this section, we explore the patterns for our two quality proxies by type of school and over time, and we examine the extent to which accountability pressures have affected the patterns.

School Characteristics and Accountability Regimes

We begin our analysis by using ordinary least squares (OLS) equations to look at differences across categories of schools and over time. For Tables 2 through 4, the dependent variables are the within-school difference in average standardized teacher licensure test scores or the within-school difference in our index of teacher quality 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 some form of accountability for 13 of the 15 years in the full period. The statistically significant entry of negative 0.083 for the constant in the first column and the comparable entry of negative 0.044 in the third column tell the same story. The quality of the teachers in the lower grades, whether measured by licensure test scores or the broader value added index, falls short of that in the upper grades.

Columns 2 and 4 show how the patterns differ by schools divided into quintiles based on student disadvantage, where disadvantage is measured by the percent of students eligible for free and reduced price lunch. [7] 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 descriptive statistics). This greater variance among teachers in the disadvantaged schools suggests the potential for larger differences between grade sets in those schools. Nonetheless, while the entries are negative and statistically different from 0 for every quintile, the most disadvantaged schools (those in Qunitile 5) exhibit the smallest differences across grades by both measures. For schools in Quintile 5, for example, the teacher test scores in the lower grades fell short of those in the upper grades by only 3% of a standard deviation in contrast to 9.5 % of a standard deviation in the Quintile 1 schools. This pattern suggests that the principals of the less advantaged schools are less strategic or have less market power to place their stronger teachers in selected grades than do the more advantaged schools. Without further analysis by accountability regime, however, one should not attribute the patterns specifically to accountability pressures.

<Insert Table 2>

To address the role of accountability pressures, Table 3 reports the within-school differences in teacher quality for each year in our time period, from 1995 to 2009, and for the three accountability regimes: no accountability, ABCs and NCLB. Although the two measures of quality exhibit somewhat different patterns in the early years, similar patterns emerge for the NCLB period: Starting in 2003, the first year of NCLB, the shortfall of teacher quality in the lower grades becomes significantly larger according to both measures. This pattern is highlighted in columns 2 and 4 which consolidate the years into the three accountability periods. For both quality measures we find a negative coefficient during the NCLB period which is statistically different not only from zero but also from the pre accountability

[7] Similar patterns emerge for other measures of disadvantage including quintiles defined by percent of minority students or by the average test scores of the students.

period. Only in the years prior to 2003, do the two measures generate somewhat different results. According to the licensure score measure of quality, teachers in the lower grades fell short of those in the upper grades in every year prior to 2003 while that was not the case for the value added index.

Figures 1 and 2 shed additional light on the trends over time. Both figures depict teacher quality measures separately for teachers in the upper and lower grades. Thus, they do not specifically capture the within-school differences highlighted in the tables. Nonetheless they are useful for understanding and interpreting the trends in the within-school differences. Figure 1 shows that licensure scores were rising throughout the period for both groups of elementary school teachers. This pattern reflects explicit efforts by state policy makers to raise teacher quality, which they accomplished in part by increasing the quality of new teachers needed to meet the needs of a growing student population. Although a close look at the lines indicates a growing divergence in the two lines over time, one might conclude from the graph that the difference in the later years simply reflects a continuation of the trends in the earlier years.

That is not the case for the trends depicted for the value added index in Figure 2. In this figure we see that from 1997 to 2002, namely the ABCs period, teacher quality was increasing at about the same rate at both levels of schooling. Starting in 2003, however, there was a notable increase in the quality of teachers in the late grades and a decline in the early years. It is hard to come up with explanation for this pattern other than that the accountability pressures under NCLB induced principals to alter the way they allocated teachers among grades.

In Table 4, we document that the principals of some schools were apparently far more responsive to the NCLB pressures than were those in other schools. The greatest response emerges for the Quintile 3 schools, that is those with average proportions of disadvantaged students. For that quintile, we observe quality shortfalls in teacher quality in the lower grades during the NCLB period that, at least for the value added proxy for teacher quality, differ statistically from the situation in the pre-

accountability period. In contrast, the changes over time in the more advantaged Quintile 1 schools are far smaller: throughout the entire period, the lower grades were at a quality disadvantage relative to the upper grades and that pattern did not change much with the advent of NCLB. Among the most disadvantaged schools, we find larger quality shortfalls in the NCLB period relative to the pre-accountability period but we cannot rule out the possibility that the differences reflect chance variation.

These patterns provide support for the hypothesis that school principals responded strategically to the incentives created by the NCLB accountability program by assuring that their stronger teachers were in the upper grades, and that the response was especially strong in the schools with average proportions of disadvantaged students. The observation of a much smaller, if any, response to the pressures of the ABCs accountability system could reflect the difference between the negative sanctions associated with NCLB and the positive rewards offered by the ABCs, differences in the form of the accountability requirements, or the confounding effects of state policies during the ABCs years. In any case, accountability pressures 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 and that principals respond to these preferences.

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 the likelihood of moving down from the upper grades to the lower grades or up from the lower grades to the upper grades.

Table 5 reports results based on the licensure test measure of quality alone. All the entries are expressed as marginal effects calculated at mean values. The statistically significant entry of negative 0.003 in the first column implies that a teacher with a higher licensure test score is slightly less likely to move down than a teacher with lower test scores. Similarly, the third column reports a statistically significant higher probability that a teacher with above average test scores will move up to the higher grades compared to a teacher with lower licensure test scores.

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 levels. 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 regimes 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. Translating the marginal effect into a percent increase, we conclude that a teacher with a licensure test score one standard deviation above the mean is 19 percent more likely to move up than an average teacher during the NCLB period.[8]

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 more likely to be 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>

[8] During the NCLB period, a teacher with licensure test scores at the mean has a 5.6% chance of moving up, calculated by adding a constant of 0.049 plus the marginal effect of 0.007 on the NCLB term. A teacher with a test score one standard deviation above the mean has a 6.7% chance of moving up, a 19.6% increase.

For the final analysis in this section, the fact that we focus only on the movement of teachers who start in the upper grades means we can rely directly on their value added measures. As described in Appendix A, teacher value-added is calculated separately for reading and math using a Bayesian shrinkage estimator 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 higher value-added measures than for those with lower measures in both reading and in math. The marginal entries in the table imply that, compared to teachers with average value added measures , teachers with measures one standard deviation above the mean are 25 percent less likely in reading, and 31 percent less likely in math to move down. In addition, 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.

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

<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 quality 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 the prior year accountability status

for a specific school to be 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 reports patterns for differences in teacher quality between lower and upper grades in regression models that include indicator variables for whether the school failed to meet adequate yearly progress (AYP) under NCLB or Expected Growth under the ABCs in each of the previous three years. The regression models also include year indicators to control for changes in the pattern of teacher quality differences under the accountability regimes. Columns 2 and 4 also include 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 columns 1 and 3 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 pattern runs counter to the expected direction of accountability pressure on the distribution of quality teachers. However, the corresponding entries in columns 2 and 4 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 those columns indicate 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 reduced teacher quality in the lower grades by 3.4% to 4.3% percent of a standard deviation relative to the upper grades. Although the coefficients in columns 2 and 4 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 the NCLB AYP requirement in a specific year appears to generate a short term strategic response in the predicted direction. 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. Our results indicate that concern about teacher quality in kindergarten, first and second grades is warranted in that teachers in these grades are of lower quality, as measured either by their licensure test scores or by our broader value-added based index, than those in the upper elementary grades. Moreover, the findings that NCLB accountability increases the relative shortfalls of teacher quality in the lower grades and that schools tend to move teachers of higher quality to the upper grades and teachers of lower quality to the lower grades support the conclusion that accountability pressure in the form of NCLB induced schools to pursue actions that work to the disadvantage of the children in the lower grades. Responses to the state's ABCs accountability program are far more muted if they appear at all.

The patterns across schools serving different proportions of disadvantaged students are somewhat mixed. In the pre-accountability period, the evidence shows that the schools serving

advantaged students were more likely than those serving low-income students to place their higher quality teachers in the upper grades. We can only speculate about the reasons for that pattern for the more advantaged schools. It may reflect, for example, teacher or parental preferences, but alternatively it could reflect strategic behavior on the part of school principals. We simply do not know. NCLB accountability pressures appear to have had no effect on the distribution of teachers in those advantaged schools. Its biggest impact appears to be in the schools serving average proportions of low income students, perhaps because those schools felt the most pressure from the accountability system. Compared to their more advantaged counterparts, such schools were more likely to be on the borderline of making the required AYP and compared to their more disadvantaged counterparts, such schools may have had a greater chance of meeting the requirements.

Some people may be tempted to question the causal nature of our NCLB findings. It is difficult, however, to come up with an alternative explanation that would account for the within-school patterns that we observe beginning with the first year of the NCLB period, that emerge so clearly for our value added index of teacher quality in figure 2, that show up in our analysis of teacher movement, and that also emerge from our school- specific models. The situation with respect to the state's ABCs accountability program differs. As a state run program, it was part of a broader state effort to improve North Carolina schools that included efforts to increase teacher quality. Hence, during the ABCs period it is conceivable that the state could have pursued policies that would have had differential effects across grades within schools. Thus, had we found evidence that the ABCs program was associated with a change in the distribution of teacher quality across grades in elementary schools, we would not have been comfortable with strong causal claims. With respect to the federal NCLB program, however, we believe that causal claims are warranted.

One might also wonder whether the measured effects are large enough to be policy relevant. The concern arises most clearly for the licensure measure of teacher quality. Other studies, including

Clotfelter, Ladd and, Vigdor (2007) show that a one standard deviation in teacher test scores is associated with an 1.1 percent of a standard deviation change in students achievement, all else held constant. Hence, the small differences in teacher test scores between the upper and lower grades would translate into very small changes in student achievement. To the extent that teacher test scores are simply a proxy for other credentials, both measurable and unmeasurable, however, that also have positive effects on student test scores, the estimated effect understates the total achievement effects of the accountability pressure. Our broader value added proxy for teacher quality partially addresses the package issue, but even here we find that NCLB generates only a 0.05 standard deviation gap in quality, where the reference now is to the distribution of teacher value added scores.

By either measure, the effects are quite small and perhaps not directly a cause for serious concern. If the effects are indeed reflecting strategic behavior by principals, however, we would expect that the same type of sorting would occur for unmeasured components of teacher quality. If principals are sorting teachers according to unmeasured components of quality as well as the measured components, the true effects on the gap in teacher quality between lower and upper elementary grades may be much larger than reflected in this study. More complete measures of teacher quality are needed before a conclusion can be drawn regarding the true size of the effect on the distribution of overall teacher quality across grades.

We also note that the shortfall in the lower grades looms larger when we account for the fact that a child in the lower grades would be affected by this average shortfall in each of her three years in early elementary school. The other side of that coin, though, is that when she reaches the later grades, she will benefit from a favorable differential. The net effect of this redistribution of teacher quality across the elementary school years ultimately depends on the relative contributions to child development of teacher quality at different points in the child's development trajectory. To the extent that low teacher quality in the early years is more detrimental to a child's development than low teacher

quality in the later years, this unintended redistributional component of NCLB would become particularly salient.

In light of the patterns reported in this paper, we believe that policy makers 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 (6045): 975-8.

Boyd, Donald, Hamilton Lankford, Susanna Loeb and James 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, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach and Danny Yagan. 2010. How does your kindergarten classroom affect your earnings? Evidence from project star: National Bureau of Economic Research.

Chingos, Matthew M. and Martin R. West. 2011. Promotion and reassignment in public school districts: How do schools respond to differences in teacher effectiveness? *Economics of Education Review* 30 (3): 419-33.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41 (4): 778-820.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor. 2007. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review* 26 (6): 673-82.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor. 2010. Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects. *Journal of Human Resources* 45 (3): 655-81.

Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor and Roger 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 (2): 251-71.

Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor and Justin Wheeler. 2006. High-poverty schools and the distribution of teachers and principals. *NCL Rev.* 85: 1345.

Cohen-Vogel, Lora. 2011. "Staffing to the test" are today's school personnel practices evidence based? *Educational Evaluation and Policy Analysis* 33 (4): 483-505.

Cullen, Julie Berry and Randall Reback. 2006. Tinkering toward accolades: School gaming under a performance accountability system. *Advances in Applied Microeconomics* (14): 1-31.

Currie, Janet and Duncan Thomas. 1998. School quality and the longer-term effects of head start: National Bureau of Economic Research.

Currie, Janet M. 2008. *The invisible safety net: Protecting the nation's poor children and families*: Princeton University Press.

Dynarski, Susan, Joshua M. Hyman and Diane Whitmore Schanzenbach. 2011. Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion: National Bureau of Economic Research.

Figlio, David N. 2006. Testing, crime and punishment. *Journal of Public Economics* 90 (4): 837-51.

Figlio, David N. and Lawrence S. Getzler. 2006. Accountability, ability and disability: Gaming the system? *Advances in Applied Microeconomics* 14: 35-49.

Goldhaber, Daniel. 2007. Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources* 42 (4): 765-94.

Goldhaber, Daniel. 2008. Teachers matter, but effective teacher quality policies are elusive. *Handbook of research in education finance and policy*: 146-65.

Jacob, Brian A. 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics* 89 (5): 761-96.

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

Mervis, Jeffrey. 2011. Giving children a head start is possible—but it's not easy. *Science* 333 (6045): 956-57.

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, William S. Barnett, Clive R. Belfield and Milagros Nores. 2005. *Lifetime effects: The high/scope perry preschool study through age 40*. Ypsilanti: High/Scope Press.

Table 1 Credentials of North Carolina Teachers in Lower and Upper Elementary Grades, 2009 and 1995.

	N	Experienced Teachers	Master's Degree	Licensure Test Score	National Board Certification	Value Added Index
PANEL A. 2009						
Lower	13,827	90.5%	27.8%	-0.008	8.2%	0.012
Upper	12,350	88.3%	31.7%	0.090	9.9%	0.103
Within School Difference	1,285	2.7%	-3.9%	-0.100	-1.2%	-0.087
PANEL B. 1995						
Lower	9,507	85.7%	25.7%	-0.079	0.0%	-0.125
Upper	8,590	86.5%	28.3%	-0.016	0.0%	-0.111
Within School Difference	1,016	-1.0%	-2.7%	-0.059	0.0%	-0.018

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 and Value Added Index between Lower and Upper Elementary by Free or Reduced Price Lunch Quintile, 1995-2009

	Licensure Test Scores		Value Added Index	
Constant	-0.083***		-0.044***	
	(0.008)		(0.009)	
Quintile 1 (Most Advantaged)		-0.095***		-0.065***
		(0.014)		(0.016)
Quintile 2 (More Advantaged)		-0.093***		-0.041**
		(0.012)		(0.015)
Quintile 3 (Average)		-0.109***		-0.060***
		(0.011)		(0.014)
Quintile 4 (More Disadvantaged)		-0.083***		-0.044*
		(0.014)		(0.018)
Quintile 5 (Most Disadvantaged)		-0.030^{++}		-0.003^{+}
		(0.017)		(0.021)
Observations	16,311	15,349	16,291	15,329
R-squared	0.000	0.040	0.000	0.008

Note: * indicates that the coefficient is significantly different from zero at the *<.05, **<.01, and***<.001. + 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 free/reduced lunch students. Observations are clustered at the school level.

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

	Licensure Test Score		Value Added Index	
Pre-Accountability		-0.056*		-0.021
		(0.013)		(0.015)
ABCs		-0.073*		-0.009
		(0.010)		(0.012)
NCLB		-0.097*$^{++}$		-0.077*$^{++}$
		(0.009)		(0.011)
1995	-0.053*		-0.018	
	(0.015)		(0.017)	
1996	-0.059*		-0.025	
	(0.014)		(0.017)	
1997	-0.060*		-0.012	
	(0.014)		(0.016)	
1998	-0.061*		0.009	
	(0.013)		(0.016)	
1999	-0.077*		-0.008	
	(0.013)		(0.016)	
2000	-0.087*$^{+}$		-0.009	
	(0.013)		(0.017)	
2001	-0.080*		-0.027	
	(0.013)		(0.017)	
2002	-0.071*		-0.005	
	(0.014)		(0.019)	
2003	-0.095*$^{+}$		-0.080*$^{+}$	
	(0.013)		(0.019)	
2004	-0.102*$^{++}$		-0.077*$^{++}$	
	(0.013)		(0.017)	
2005	-0.095*$^{+}$		-0.091*$^{++}$	
	(0.012)		(0.017)	
2006	-0.092*$^{+}$		-0.060*	
	(0.012)		(0.015)	
2007	-0.093*$^{+}$		-0.071*$^{+}$	
	(0.012)		(0.015)	
2008	-0.100*$^{++}$		-0.075*$^{++}$	
	(0.012)		(0.014)	
2009	-0.104*$^{++}$		-0.087*$^{++}$	
	(0.012)		(0.015)	
Observations	16,311	16,311	16,291	16,291
R-squared	0.038	0.037	0.010	0.010

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. Observations are clustered at the school level.

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

	Pre-accountability	ABCs	NCLB
Quintile 1 (Most Advantaged)			
Licensure Test scores	-0.098***	-0.098***	-0.096***
Value Added Index	-0.061**	-0.049*	-0.090***
Quintile 3 (Average)			
Licensure Test scores	-0.049	-0.101***	-0.122***
Value Added Index	-0.007	-0.023	-0.103***$^{+}$
Quintile 5 (Most Disadvantaged)			
Licensure Test scores	0.033	0.025	-0.065***
Value Added Index	0.018	0.098**	-0.063*

Note: * indicates that the coefficient is significantly different from zero at the *<.05, **<.01, and***<.001. + indicates that coefficient is significantly different from the coefficient on the pre-accountability period at the +<.05, ++<.01, and +++<.001 level. Quintiles run from most advantaged to most disadvantaged with the first quintile having the least free/reduced lunch students. Observations are clustered at the school level.

Table 5. Logistic Model of the Probability of a Teacher Moving Up or Down based on Licensure Test Scores and Accountability Regimes, 1995-2009 (Marginal Effects)

	Moving Down	Moving Down	Moving Up	Moving Up
Teacher Test Score	-0.003**	0.000	0.008***	0.000
	(0.001)	(0.004)	(0.001)	(0.003)
ABCs		-0.003		0.015***
		(0.004)		(0.003)
NCLB		-0.002		0.007*
		(0.003)		(0.003)
ABCs*Teacher Test Score		-0.006		0.005
		(0.004)		(0.004)
NCLB*Teacher Test Score		-0.001		0.011**
		(0.004)		(0.004)
Observations	99,957	99,957	122,654	122,654

Note: *** p<0.001, ** p<0.01, * p<0.05; Marginal effects are calculated at the mean of the independent variables.

Table 6. Logistic Model of the Probability of a Teacher Moving Down based on Teacher Value-added and Accountability Regimes, 1995-2009 (Marginal Effects)

	Moving Down	Moving Down	Moving Down	Moving Down
Reading Value-added	-0.015***	-0.012***		
	(0.001)	(0.003)		
Math Value-added			-0.018***	-0.012***
			(0.001)	(0.003)
ABCs		-0.009**		-0.009**
		(0.003)		(0.003)
NCLB		-0.006*		-0.006*
		(0.003)		(0.003)
ABCs*Reading Value-added		-0.003		
		(0.003)		
NCLB*Reading Value-added		-0.003		
		(0.003)		
ABCs*Math Value-added				-0.004
				(0.003)
NCLB*Math Value-added				-0.006*
				(0.003)
Observations	97,618	97,618	97,647	97,647

Note: *** p<0.001, ** p<0.01, * p<0.05; Marginal effects are calculated at the mean of the independent variables.

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

	Licensure Test Scores		Value Added Index	
Failed to Meet AYP 1 Year Previous	-0.010	-0.034**	-0.013	-0.043**
	(0.011)	(0.012)	(0.015)	(0.016)
Failed to Meet AYP 2 Years Previous	-0.006	-0.017	0.000	-0.002
	(0.012)	(0.013)	(0.016)	(0.017)
Failed to Meet AYP 3 Years Previous	-0.014	-0.023	-0.004	-0.018
	(0.014)	(0.015)	(0.017)	(0.018)
Failed to Meet Expected Growth 1 Year Previous	0.030**	0.004	0.037**	0.007
	(0.010)	(0.011)	(0.013)	(0.015)
Failed to Meet Expected Growth 2 Years Previous	0.027**	0.011	0.029*	0.014
	(0.010)	(0.011)	(0.014)	(0.015)
Failed to Meet Expected Growth 3 Years Previous	0.020+	0.008	0.021	0.009
	(0.011)	(0.012)		(0.016)
Percent Minority Students		0.093*		0.026
		(0.041)		(0.050)
Percent Free or Reduced Price Lunch Students		-0.024		-0.011
		(0.050)		(0.065)
Previous Year Performance Composite		-0.002		-0.003*
		(0.001)		(0.001)
Constant	-0.053***	0.025	-0.018	0.175
	(0.014)	(0.065)	(0.017)	(0.132)
Observations	16,311	12,005	16,291	11,989
R-squared	0.003	0.009	0.005	0.009

Note: *** p<0.001, ** p<0.01, * p<0.05; year fixed effects are included in all regressions. Observations are clustered at the school level.

Figure 1. Average Teacher Test Scores over Time in Early and Late Elementary Grades

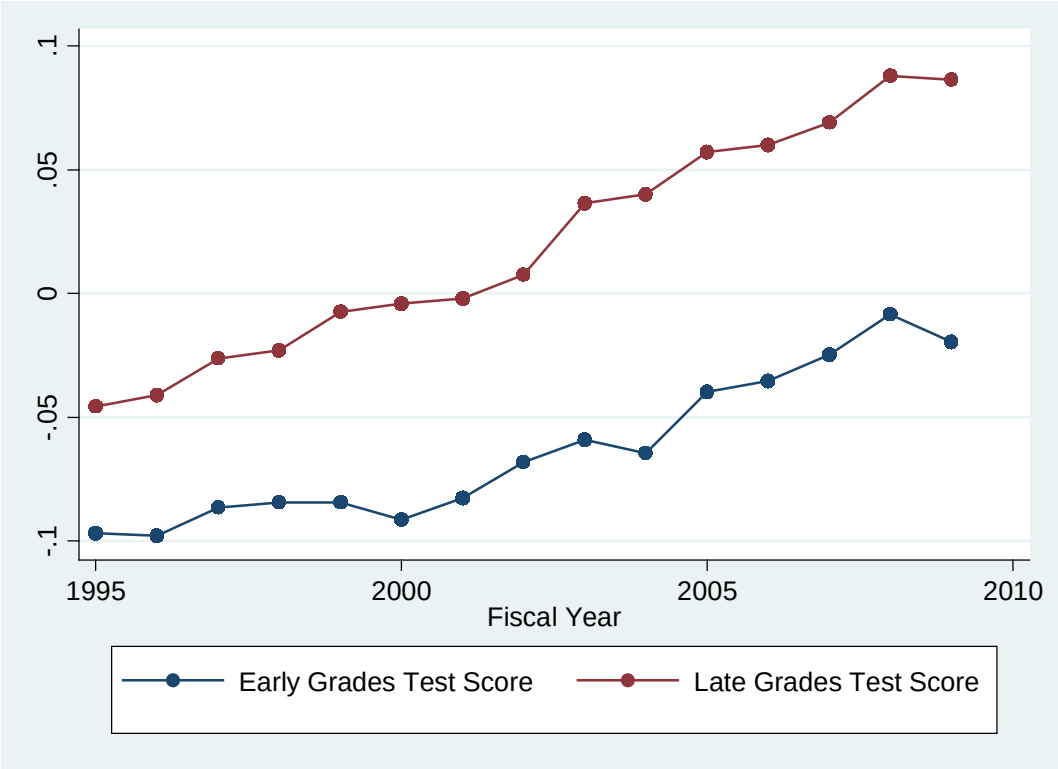
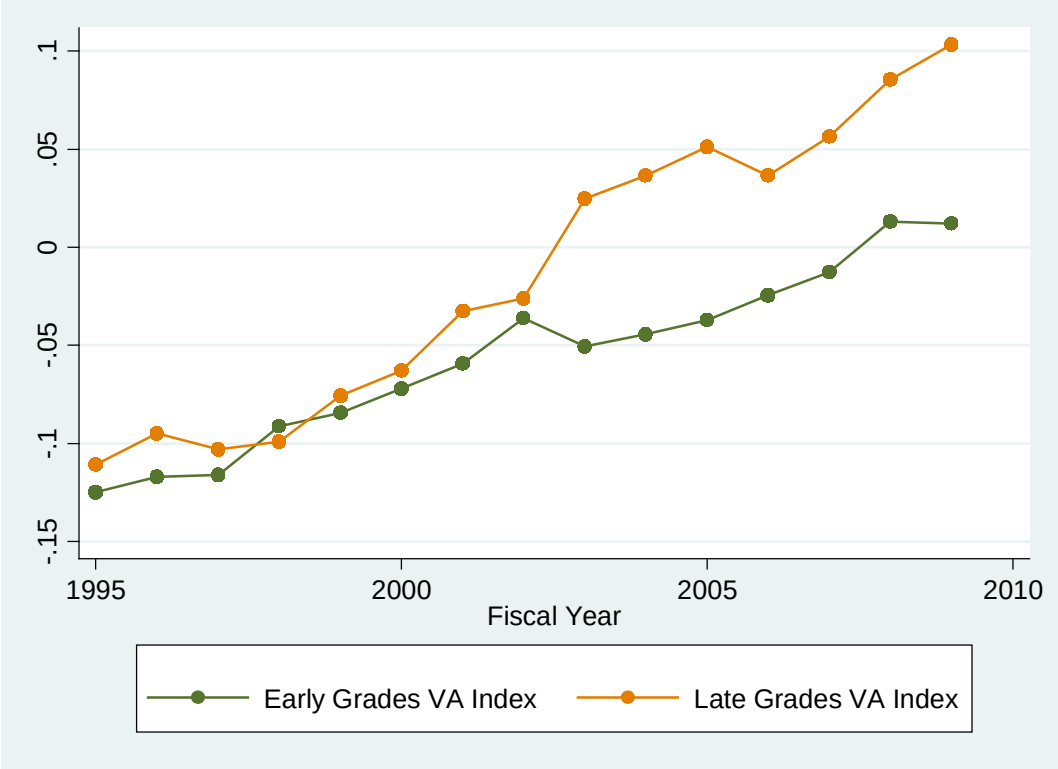


Figure 2. Average Value Added Index Over Time in Early and Late Elementary Grades



Appendix A: Quintile Descriptive Statistics

Average of School Level Means and Standard Deviations across Free/Reduced Lunch Quintiles

	Average School Mean	Average School Standard Deviation
Licensure Test Scores		
Overall	-0.018	0.779
Quintile 1 (Most Advantaged)	0.077	0.724
Quintile 2 (More Advantaged)	0.079	0.756
Quintile 3 (Average)	0.036	0.781
Quintile 4 (More Disadvantaged)	-0.031	0.799
Quintile 5 (Most Disadvantaged)	-0.256	0.833
Value Added Index		
Overall	-0.003	0.920
Quintile 1 (Most Advantaged)	0.138	0.925
Quintile 2 (More Advantaged)	0.067	0.908
Quintile 3 (Average)	0.028	0.911
Quintile 4 (More Disadvantaged)	-0.068	0.918
Quintile 5 (Most Disadvantaged)	-0.296	0.961

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.

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

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 and 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: Construction of the Value Added Index

Math value added scores for upper grade teachers were regressed on measurable teacher qualifications using the following equation:

$$value\ added\ score\ = \beta_1 * less\ than\ 3\ years\ experience + \beta_2 * test\ score + \beta_3 * test\ score^2 + \beta_4 * college\ selectivity + \beta_5 * masters + \beta_6 * full\ license\ + \beta_7 * lateral\ entry\ + school\ fe$$

Table C1 shows the coefficients for each teacher qualification included in the regression. Less than three years experience is an indicator variable for a teacher who has been teaching for fewer than three years. Test score is the average of the scores received by a given teacher on all licensure tests taken with each test score standardized based on the test type and year of the test. College selectivity levels are indicators of selectivity at the teacher's undergraduate institution. Level 1, the most selective schools, are the omitted category. Levels 2 through 7 indicate decreasing levels of selectivity. College selectivity level 8 indicates teachers who attended college internationally, and level 9 indicates teachers whose undergraduate institution is listed as the central office of a college system. The variable, masters, is an indicator for whether the teacher has a master's degree or higher level of education. Any lateral entry indicates a teacher who is licensed as a lateral entry regardless of how long they have been teaching. Initial provisional lateral entry is an indicator variable for a lateral entry teacher currently licensed on a new provisional license.

Table C1. Regression of Math Value Added Scores on Teacher Credentials

	Math Value Added
Less Than 3 Yrs Experience	-0.137***
	(0.008)
Average Licensure Test Score	0.127***
	(0.003)
Licensure Test Score^2	-0.012***
	(0.003)
College Selectivity Level 2	-0.121***
	(0.028)

College Selectivity Level 3	-0.117***
	(0.027)
College Selectivity Level 4	-0.122***
	(0.026)
College Selectivity Level 5	-0.151***
	(0.026)
College Selectivity Level 6	-0.176***
	(0.032)
College Selectivity Level 7	0.021
	(0.235)
College Selectivity Level 8	0.303***
	(0.052)
College Selectivity Level 9	-0.258***
	(0.046)
Master's Degree	-0.026***
	(0.006)
Fully Licensed	0.308***
	(0.036)
Initial Provisional Lateral Entry	-0.168***
	(0.026)
Any Lateral Entry	0.308***
	(0.065)
Constant	-0.104*
	(0.044)
Observations	173,016
R-squared	0.051

The results of the regression were used to create predicted valued added scores for all teachers in both the lower and upper elementary grades. The predicted value added scores were then normalized to have a mean of zero and a standard deviation of one to create the value added index.

Appendix D: Data Description

The data for this project were provided by the North Carolina Education Research Data Center (NCERDC) housed at Duke University. These administrative data consist of an individual record for every teacher in each year 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 also include 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, we used this information 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, And the numbers of male, female, white students, and 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.