Status versus Growth:
The Distributional Effects of School Accountability Policies*

January 3, 2009
Please do not cite without permission

Helen F. Ladd
Edgar Thompson Professor of Public Policy Studies and
Professor of Economics
Duke University
hladd@pps.duke.edu

Douglas L. Lauen
Assistant Professor
Department of Public Policy
University of North Carolina at Chapel Hill
dlauen@unc.edu

* The authors gratefully acknowledge the research assistance of Steven M. Gaddis, the Spencer Foundation for funding, and the North Carolina Education Research Data Center, housed at Duke University, for providing us with access to the data for this project. Authors listed alphabetically.

Abstract

Although the federal No Child Left Behind program judges the effectiveness of schools based on their students' achievement status, many policy analysts argue that schools should be measured, instead, by their students' achievement growth. Using a ten-year student-level panel dataset from North Carolina, we examine how school-specific pressure associated with the two approaches to school accountability affects student achievement at different points in the prior-year achievement distribution. Achievement gains for students below the proficiency cut point emerge in response to both types of accountability systems. In contrast to prior research highlighting the possibility of educational triage, we find little or no evidence that schools in North Carolina ignore the students far below proficiency under either approach. Importantly, we find that the status, but not the growth, approach reduces the reading achievement of higher performing students, with the losses in the aggregate exceeding the gains at the bottom. Our analysis suggests that the distributional effects of accountability pressure depend not only on the type of pressure for which schools are held accountable (status or growth), but also the tested subject.

High on the U.S. educational policy agenda is how best to hold schools accountable for the performance of their students. One of the goals of any accountability policy is to shorten the feedback loops between policymakers, principals and teachers. With standards based accountability programs, policy makers set clear standards, measure student performance, and use those measures to evaluate the effectiveness of schools (Cohen, 1996; O'Day & Smith, 1993). Successful schools are then typically provided rewards in the form of public recognition, financial bonuses for teachers or some combination of both. Unsuccessful schools may be sanctioned or provided additional support, depending on whether the system is designed to be punitive or constructive. The ultimate goal of a standards based accountability system is to generate greater student achievement consistent with the standards (Figlio & Ladd, 2008; Ladd, 1996).

With the passage of the federal No Child Left Behind Act (NCLB) in early 2002, for better or for worse, student test scores in math and reading have come to represent the outputs of interest, regardless of their relationship to any specific curriculum standard, and schools are judged primarily on the academic status of their students. In particular, NCLB requires every U.S. public school to test all students annually in reading and math in grades 3-8 and once in high school, requires each state to set annual targets for the percentages of students meeting a state-determined proficiency standard in order to reach the goal of 100 percent proficiency by 2013/14, and includes sanctions for schools that fail to make the required adequate yearly progress (AYP) toward that goal. In addition, it holds schools accountable not only for the overall performance of their students but also for that of racial and economic subgroups. Among the many criticisms of NLCB are that the proficiency standards differ across states, the focus on math and reading narrows the curriculum, holding schools accountable for annual progress increases the instability of school performance measures, more diverse schools are more likely to be penalized, and the goal of 100 percent proficiency is unrealistic (Amrien & Berliner, 2002; Balfanz *et al.*, 2007; Figlio, 2005, 2006; Hamilton *et al.*, 2005; Kane & Staiger, 2002; Linn, 2000; Peterson & Hess, 2006).

Despite these criticisms, many people believe that test-based accountability can be a useful strategy for raising student achievement, especially for low-performing students. The theory of action behind educational accountability is that by setting standards and measuring performance relative to standards, teachers will work harder and students will learn more. Increasingly, however, observers have argued for shifting the metric for school accountability away from the achievement status of a school's students, as is the case under NCLB, in favor of a metric based on students' growth in achievement during the year (Hanushek & Raymond, 2005; Ladd & Walsh, 2002; Toch & Harris, 2008).

The argument for using achievement growth rather than achievement status as the basis of school accountability is two-fold. First, because children come to school with different degrees of readiness to learn and prior achievement levels, many people believe it is unfair, and potentially unproductive, to expect schools alone to offset the effects of the background characteristics of their students. Instead, the argument goes, schools should be held accountable for outcomes over which they have more control, such as how much the children learn during the year, typically measured by their gains in test scores. Second, the focus on achievement status, as defined by a proficiency threshold, provides a strong incentive for schools to focus attention on students near the threshold to the potential disadvantage of students far below the threshold and of those above the threshold. This distributional aspect of status based accountability programs has received significant attention in the recent literature (Ballou & Springer, 2008; Booher-Jennings, 2005; Krieg, 2008; Neal & Schanzenbach, Forthcoming; Reback, 2008). At the same time, some growth models have been criticized for lack of transparency and their failure to require students to meet specific standards.

The distributional effects of accountability systems are the focus of this paper. In contrast to recent research, which has focused almost exclusively on the distributional effects of status programs such as NCLB, we compare the distributional effects of a system based on achievement status to one based on achievement growth. We are able to compare the two approaches because our empirical work is based on longitudinal data from North Carolina where schools have been subject to a growth based accountability system since 1996/97 and ever since 2002/03 have also been subject to the status requirements of NCLB. Given the increasing national policy interest in moving to growth models of accountability our comparison of the distributional effects of the two approaches is timely.

Specifically, in this empirical study we use student-level data over time to compare and contrast how school-specific pressure associated with the two approaches to school accountability affects student achievement at different points in the prior-year achievement distribution. The availability of consistent test score data over time allows for careful modeling of student achievement gains, including the use of student fixed effects to account for time-invariant characteristics of students such as their ability.

Consistent with our theoretical predictions about how schools facing accountability pressure are likely to respond, we find evidence of achievement gains for students below the proficiency cut point for both types of accountability systems. The approaches differ, however, in their effects on student performance at the high end of the distribution. In contrast to the gains approach, the status approach generates achievement losses for higher performing students that, in the aggregate, exceed the gains at the bottom. We find that the aggregate gains to students at the bottom of the distribution in the affected schools would have to be weighted at 2-3 times those at the top to justify the losses at the higher end.

Achievement effects of school accountability programs

The most convincing studies of how accountability affects overall achievement emerge from cross-state studies such as Carnoy and Loeb (2002), Hanushek and Raymond (2005) or from careful district-specific studies that permit comparisons to other districts such as Jacob (2005) or Ladd (1999). These and other studies are reviewed in Figlio and Ladd (2008). Emerging from research of this type is that the introduction of a school-based accountability program generally raises achievement when achievement is measured by the high-stakes test used in the accountability system. Some studies also report positive achievement effects when achievement is measured by a low-stakes test, such as the National Assessment of Education Progress (NAEP) as in Carnoy and Loeb (2002) and Hanushek and Raymond (2005), or by a low-stakes state test as in Jacob (2005) but in this latter case only in the higher grades. In general, when achievement gains do emerge, they tend to be larger for math than for reading.

Research has also found, however, that high stakes testing can narrow and fragment the curriculum, promote rote, teacher-directed instruction, and encourage schools to teach test-preparation skills rather than academic content, tendencies that may be stronger in schools with high minority and low income populations (Amrien & Berliner, 2002; Darling-Hammond, 2004; Linn, 2000; Nichols & Berliner, 2007; Orfield & Kornhaber, 2001; Valenzuela, 2005). Moreover, in schools facing accountability pressure, teachers and principals may manipulate the test-taking pool through selective disciplinary practices and reclassifying students as requiring special educational services, thereby making them ineligible for tests. In addition, and of particular relevance for the present study, they may focus instruction and extra resources on those students most likely to improve a school's external standing (Booher-Jennings, 2005; Figlio, 2006; Weitz & Rosenbaum, 2007).

Our main question is how school accountability affects achievement at different points in the achievement distribution in the schools under the most pressure to raise achievement. Four distributional questions are of particular interest. The first, and most basic, is whether there are any within-school distributional effects, that is, whether accountability pressure is associated with greater gains in achievement for students at some points in the prior-year achievement distribution than at others. Unless an accountability system is specifically intended to change the distribution of student outcomes within schools, such distributional effects may well not be desirable. Second, to the extent that there are distributional effects, do the largest benefits accrue to students at the low end of the distribution? Such an outcome would be deemed desirable provided the goal of the accountability system were to raise the achievement of low-achieving students in low-performing schools, but less so if the goal were to raise achievement across the board in such schools. Third, to what extent do trade-offs emerge between gains to students at the bottom of the distribution and achievement losses to those at the top? Making policy judgments about such tradeoffs requires information on how society values achievement gains and losses for students at different points in the achievement distribution. Fourth, to what extent is there evidence of educational triage, in the sense that additional resources are focused on students around a designated threshold to the detriment of those far from the threshold? Of particular concern is that students at the very bottom of the achievement distribution may be so far below the threshold that they are worse off under the accountability system than they otherwise would be.

Several recent studies using different methodologies and data sets address one or more of these questions in the context of status based accountability systems that measure school success by student passing rates. Receiving most attention in the literature is the issue of educational triage. Booher-Jennings (2005) provides qualitative evidence from a single school and its associated urban school district in Texas that teachers do indeed respond to incentives to increase pass rates as one would expect, namely by focusing additional attention on students near the passing rate. Based on careful quantitative analysis, Neal and Schanzenbach (forthcoming) document that the introduction of two separate accountability systems in Chicago induced schools to focus on students near the middle of the achievement distribution to the disadvantage of the students at the two tails of the distribution, while Krieg (2007) reports similar findings for Washington State. In contrast, in their quantitative study of NCLB in 7 states based on test data from the

Northwest Evaluation Association, Ballou and Springer (2008) find little or no evidence of adverse effects for the lowest performers.[1]

More generally, Ballou and Springer (2008) find evidence of gains to students at the bottom of the distribution, but find no consistent evidence that schools facing accountability pressure neglect their high achieving students to focus on low achievers. A study by Reback (2008) based on Texas data during the 1990s also generally finds positive effects on the very low achievers. Contrary to the triage hypothesis, Reback finds that when a school has a realistic chance of improving its accountability rating, the lowest performing students make greater than expected gains, even if they have no chance of passing the exam in that subject. In addition, Reback uncovers some intriguing distributional differences by subject. His evidence suggests that schools respond to incentives related to math in ways that increase the performance of low performing students with at most small adverse effects on higher achieving students. In reading, by contrast, except in certain cases, school-wide incentives to raise student performance on the reading exam appear to harm students who have a moderate to strong probability of passing the exam. These patterns, Reback suggests, may reflect differences in the subject-specific strategies schools used to improve performance. When they are under pressure to raise math scores, they may well improve basic math instruction for all students, but when they are under pressure to improve reading scores, schools may tend to pull students out for individualized or small group instruction.

[1] A study of accountability in England also finds adverse effects of accountability on the lowest performing students (Burgess *et al.*, 2005). In contrast to the studies mentioned in the text, Burgess et al focus on performance at the high school level and introduce the element of accountability through school competition.

Our research makes a three-fold contribution to this literature. First, in addition to testing the triage hypothesis suggested by some of the existing studies in the context of the status approach to accountability, we compare and contrast the distributional effects of the status and growth approaches to accountability. Second, following Reback, we compare distributional effects in both math and reading. Third, the fact that we are able to match the test scores of individual students as they progress through school means that we can use student fixed effects to control for the unmeasurable time-invariant characteristics of students, such as their ability or motivation, that might otherwise confound the analysis.

In the following two sections, we first use a simplified model to predict the distributional effects of stylized versions of the two approaches to accountability and then describe the two programs that form the basis of our empirical work. In the following sections, we describe our data and results and end with a concluding discussion.

Predicted Distributional Impacts of the Two Approaches

We examine here the incentives faced by teachers (or other school personnel) in schools subject to each of the two forms of accountability. The status approach, as epitomized by NCLB, sets a target rate of proficiency, where the target is defined as the percentage of students in a particular school and grade level who are deemed proficient. The growth, or value-added approach, sets a target for the average rate growth of student achievement during the year. Under either system, school personnel in schools that reach the specified target in the particular system may receive rewards — financial or reputational or both — and those in schools that fail to reach the target are subject to some form of penalty, whether in the form of naming and shaming, external intervention and loss of autonomy, or potential job loss.

In the absence of either type of accountability pressure, we start with the following very simple achievement model:[2]

(1) $A_{it} = A_{it-1} + u_{it}$

where A_{it} is the student's achievement in year *t* normalized by the mean and standard deviation for all students in that grade in the state. Similarly A_{it-1} is the student's achievement in the prior year, also expressed as a normalized variable for that year and u_{it} is a random error. Thus in the absence of an accountability system the student in this simple model is assumed to remain at the same point in the performance distribution as she was in the previous year, plus or minus a random error.

The Status Approach

With the introduction of a status-based accountability system in which students are expected to reach a specified proficiency standard, say P_H in year *t*, students fall into two main categories — those with expected achievement in year *t* above the standard and those below (see figure 1). A school that does not expect to meet its overall school target rate without additional effort must decide how much additional effort to exert and on behalf of which students.

Provided the school has a relatively large number of students — large enough so that the expected value of the random components of the performance of individual students is close to zero — the school has little or no incentive to invest additional effort

[2] This model is in the spirit of that presented in Neal and Schanzenbach (forthcoming) but differs by its explicit reference to prior year achievement rather than student ability. The use of prior year achievement makes the conceptual model consistent with our empirical specification discussed below.

in students for whom the expected level of A_{it} exceeds P_H. For individual students for whom the expected level of A_{it} falls short of P_H — that is, those students with a prior year test score below P_H — the school has an incentive to invest up to the point at which the extra cost of the additional effort is just equal to the expected benefit to the school in terms of a reduced penalty. As emphasized by Neal and Schanzenbach (forthcoming), there could well be some students at the bottom of the expected performance distribution for whom the additional effort on the part of the school would simply be too costly relative to the benefits for the school to make the additional effort worthwhile. In that case, the school would focus its additional attention on the students expected to be below the proficiency level, but not so far below to make the standard out of reach.

Two factors are particularly relevant for determining which students receive additional attention — and hence are likely to exhibit achievement gains — in the context of this accountability regime. The first is the level of the proficiency standard. The higher is the standard, the more likely it is that students at the bottom of the distribution will be too far below the standard to make it worthwhile for the school to exert greater effort on their behalf. Analogously, a lower proficiency standard, such as P_L in figure 1, provides incentives for the school to focus attention on students in the lower part of the distribution, and the less likely it is that students at the bottom will be "left behind." The second factor is the nature of the educational production function. The easier it is to raise student performance at the bottom of the distribution, perhaps through improved teaching, tutoring programs or grouping strategies, the greater is the incentive for the school to invest additional effort in students whose expected achievement is low relative to the standard.

Thus, the status model generates one clear distributional prediction. Students whose expected achievement is *below* the proficiency level will receive more attention — and hence should achieve at higher levels than they otherwise would have — than those above the proficiency level. Less clear is whether there will be a group of students at the very bottom who are left behind because of the high costs of raising them to the standard.

Also not fully clear is what will happen to the achievement of students whose expected achievement slightly exceeds the proficiency standard. The presence of the error term in expression (1) means that the school has an incentive to devote some additional attention to students whose prior year achievement is slightly above the current year proficiency level; without additional attention, some of those students could well fall below the proficiency level. The more difficult it is for a school to predict how well its students will do, the more likely it is that the school will devote additional attention to students just above as well as to students below the proficiency standard.

For students whose expected achievement is well above the proficiency standard, in contrast, the question becomes whether they will receive less attention — and hence will achieve at lower levels than they otherwise would have — in the presence of the accountability pressure. If additional effort for the students at the bottom is redistributed from students at the top, achievement of the higher performing students would fall. If the school is able to garner additional resources or find ways to use existing resources more effectively than in the absence of the accountability regime, any achievement gains at the bottom of the distribution need not come at the expense of those at the top. Thus, the impact of a status based accountability system on the high achieving students is an empirical question, which depends on how resources are used within the school.

The Growth Approach

The incentives differ when accountability is based on the school's average growth in student achievement. Once again, a school under pressure to improve has an incentive to invest additional effort on behalf of any individual student up to the point that the benefits of that investment in the form of penalties avoided are just equal to the costs of that investment. In this case, however, it is difficult to predict which students will benefit most because differential benefits depend on the relationship between additional effort and student achievement at different points of the achievement distribution.

One possibility is that the additional effort needed to raise student achievement by a given amount is uniform across students defined by their prior-year achievement. In that case, a school under pressure to raise its average achievement growth has no incentive to invest any more in one group of students more than in another. Alternatively if additional effort generates greater gains for low-performing students than for high-performing students — as might be the case, for example, if achievement is measured with a test with ceiling effects (that is, one in which the performance of high achieving students cannot be distinguished) — a growth-based accountability system would give schools an incentive to invest more in the students at the bottom of the distribution than at the top. A third possibility is that, consistent with the observation that students at the high end of the achievement distribution have made greater gains in the past than those at the bottom end, it may be easier for generate larger additional gains at the top of the distribution than at the bottom. In that case, schools under pressure would have an incentive to invest in the higher performing students, with concomitantly larger gains for that group than for those other groups.

Thus, how a growth based accountability system is likely to affect the distribution of achievement gains across students within schools is an empirical question. In general, the *a priori* prediction for large distributional effects is less compelling for a pure growth approach than for a status approach to accountability.

Background on the two accountability programs in North Carolina

North Carolina is a good state in which to examine the distributional effects of these two types of accountability because its schools have been subject to the state's growth-based accountability system since the academic year 1996/97 and the federal No Child Left Behind (NCLB) status-based accountability system since 2002/03. Because the two systems use different methods for judging the effectiveness of schools, some schools that appear to be performing well under one system may do poorly under the other. In addition, in contrast to most other states, North Carolina has long used tests that are aligned with the state's standard course of study, with test scores reported on a developmental scale. As a result, the tests measure what teachers are expected to teach and students to learn, and students in any grade are less likely to reach a ceiling test score than would be the case with a maximum score in each grade.

The North Carolina ABCs Program

The North Carolina accountability program — referred to as the ABCs program — was part of a broader state effort to improve the academic performance of the state's children throughout the 1990s. First implemented in 1996-97, the ABCs program was intended to hold teachers in individual schools accountable for the overall performance of their students. Though the program applies to high schools as well, the present study focuses solely on schools serving students in grades three through eight. Of particular importance for this study, under the ABCs program schools are judged primarily on the annual achievement gains of their students from one year to the next. This growth approach to accountability was feasible because the state had been testing all students in grades three through eight annually in math and reading since the early 1990s — long before it was required to do so under the Federal No Child Left Behind legislation of 2001.

From 1996/97 to 2005, an expected average gain in test scores was predicted for each student, and the school was deemed effective or not depending on how the actual gains of its students compare to their predicted gains.[3] 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. Schools that did not achieve their expected growth were publicly identified and in some cases subject to intervention from the state. The intent of the program was to induce each school to provide its students with a year's worth of learning for a year's worth of education. In 2005, the formula for calculating growth was

[3] The expected average gains are predicted as follows. For each grade and subject (i.e. math and reading), a student's expected score is based on an equation of the form TS_t - TS_{t-1} = a + bX1 + cX2 where TS_t is the test score in either math or reading in year t and TS_{t-1} the test score in the same subject in year t-1, X1 is a proxy for the student's proficiency and is measured as the sum of the student's math and reading scores for the previous year minus the state average, and X2 is designed to account for regression to the mean and is measured as the student's prior year score in the subject of interest minus the state average in that subject. The tests are scored on a developmental scale and the parameter "a" can be interpreted as the statewide average gain in score for students in the specified grade and for the specified subject. The parameters a, b, and c were estimated using 1994 test scores for each grade. Because the b and c coefficients were quite similar across grades for each subject area, the state uses a single pair of b and c coefficients for each subject area to determine the expected growth rates. For further discussion see Ladd and Walsh (2002).

changed, but the focus on holding schools accountable for achievement growth, rather than levels, remained.[4]

In addition to their growth rankings, schools also receive various designations, such as schools of excellence, schools of distinction, and priority schools, based on the percentages of students meeting grade level standards, which carry with them no financial bonuses. In addition, some schools are labeled “low performing” based on their high failure rates as well as their poor growth performance. Thus the ABCs program does not completely ignore achievement status. At the same time, the teachers’ bonuses are based solely on the growth in student achievement. The existence of positive incentives does not alter the predictions of the simple model presented above. A school’s failure to meet its growth standard still imposes costs on its teachers; the cost is simply the bonuses foregone.

No Child Left Behind (NCLB)

The federal government started holding schools accountable for student achievement with the 2001 reauthorization of the federal Elementary and Secondary Education Act, called No Child Left Behind. This law applied to schools in North Carolina and elsewhere starting in the 2002/03 academic year. NCLB requires states to test students annually in reading and mathematics in grades 3-8, and assesses schools on the basis of whether their students are making adequately yearly progress (AYP) toward the ultimate goal of 100 percent proficiency by 2014. Moreover, each school must meet

[4] The new formula no longer is based on changes in students’ developmental scale scores from one year to the next. Instead, it is based on changes test scores normalized based on the mean and standard deviation from the first year a particular test was used in the state. The academic change for an individual student is now calculated as the student’s actual normalized score minus the average of two prior year academic scores, with the average discounted to account for reversion to the mean.

annual proficiency targets not only for the student body as a whole, but also for various subgroups defined by race, socio-economic status, and disability within the school. Failure to meet AYP brings with it consequences, such as the right of children to move to another school and the requirement that districts use their federal Title 1 grants to pay for supplemental services, including those from private providers. After five years of failure, the school is subject to state takeover by the state, an outcome that, to date, has been rare across the country, and is not directly relevant for this study which ends in 2007.

Under NCLB, North Carolina policy makers must set annual proficiency targets — defined in terms of the percentages of students who are at grade level — that will assure that each school is on target toward the 2013/14 goal of 100 percent proficiency. The result is that under the federal law each school faces an annual target defined in terms of achievement *status* rather than in terms of achievement *growth* as under the state accountability system. Not surprisingly, a school that performs well under the state's accountability system may do poorly under the federal system, and vice versa.

Table 1 shows accountability outcomes for all North Carolina elementary and middle schools by year for the period 1997 to 2007. In the first year of NCLB, 44 percent of schools failed AYP, but met the state's growth standard, and only 4 percent of schools failed both AYP and the state's growth standard. After this anomalous year, the disparities were less dramatic, but in no year did the two programs identify precisely the same schools as below accountability standards. For example, in the most recent available year, 2007, 32.6 percent of schools failed AYP (only) and an additional 23.8 percent failed both AYP and the state's growth standard.[5]

The bottom line is that for the past 11 years, schools in North Carolina have been facing pressure to raise student achievement from one or both of these accountability systems. How each of the two approaches has affected students at different points in the prior year achievement distribution is the subject of the following sections.

Data and Methods

We start with data on all students in North Carolina public schools in grades 3-8 from 1996/97 to 2006/07 for whom test scores are available in either math or reading.[6] The total panel data set includes more than 6.8 million student-year observations, with more than 1.9 million unique students and 2,129 unique elementary and middle schools. Because we are interested in changes in student test scores from one year to the next, our models are based on the approximately 4.7 million student-year observations for which we have test score data and lagged school covariates for at least two consecutive years. Figure 2 depicts the distribution of students by the number of years each appears in the

[5] More recently, North Carolina and several other states have been provided a waiver under NCLB to incorporate some elements of the growth model into the federal accountability standards. Under that provision, some students who are on track to meet the proficiency standard within three years now contribute to a school's progress toward the goal. Because the growth is still evaluated in terms of progress toward the absolute standard rather than in relation to a predicted growth standard, however, the system remains essentially a status model, rather than a growth model.

[6] These data are available through the North Carolina Education Research Data Center, housed at Duke University. To protect the confidentiality of the data, the data center replaced all the original student identifiers with new unique identifiers that allowed us to match student test scores by student over time.

mathematics analysis sample used to compute mathematics achievement gain.[7] About 31 percent of students have six test scores, one for each grade level covered in the study (grades 3-8). The two percent of students with more than six scores reflects the fact that students who were held back take a test more than once.

Test scores—The fact that North Carolina reports test scores on a developmental scale helps address, but does not fully mitigate, the comparability problems that arise from the different tests as students progress through school. In particular, the periodic rescaling of tests makes it difficult to compare scores from, say, a fifth grade math test taken in one year with a fifth grade math test taken in a different year.[8] To make them comparable both across grades and over time, we standardized all scale scores by subject, grade level and year. As a result, our estimates refer to differences in the *relative* position of students in the achievement distribution across years, rather than absolute changes. For the two subjects, math and reading, we define the two variables as follows:

Stdmath = the standardized test score in math

Stdread = the standardized test sore in reading

Accountability pressure—To capture the accountability pressures from the two programs, we define the following three school-level indicator variables and treat schools that made *both* AYP and Growth as the baseline category:

FailAYP = 1 if the school failed to make AYP, and 0 otherwise,

FailGrowth = 1 if the school failed to make its expected growth, and 0 otherwise,

[7] The histogram is based on the estimation sample from model 1 of Table 2. N=4,533,651. Number of unique students: 1,448,258. A histogram for reading achievement gain looks virtually the same and is available from the authors upon request.

[8] The state rescaled the reading tests in 2003 and the math tests in both 2001 and 2006.

FailBoth = 1 if the school failed both AYP and expected growth, and 0 otherwise.

Because NCLB did not exist prior to 2002/03, FailAYP is coded 0 for all schools prior to that year. As shown in table 1, across the post-NCLB years the percentages of elementary and middle schools not meeting AYP ranged from a low of 26 in 2004 to a high of 56 in 2007. The variation in the growth failure rate across years is even greater, in part because of an anomalous outcome in 2003. Due to changes in the state assessments in 2003, only five percent of the schools failed to make their expected growth in that year compared to 27 and 29 percent in the prior and the following years, respectively. The highest failure rate over the entire period was 43 percent in 1997; as of 2007, it was about 28 percent. Figure 3 illustrates the variation over time in the percentages of schools subject to accountability pressure as measured by these two variables. Not shown in the figure (but shown in table 1) are the proportions of schools that failed both standards. This percentage ranged from a low in 2003 of four percent to a high of 31 percent in 2006.

Distributional variables—Given our primary focus on the distributional effects of accountability, we have defined two sets of binary variables to describe a student's position in the prior year test score distributions in math and reading. We define a series of indicator variables for students below and above the proficiency level, with the category of 0 to 0.5 standard deviations (SD) above the proficiency level as the baseline category. The relevant reference point is the cut score for grade level performance because North Carolina policy makers have defined proficiency for the purposes of

NCLB as being at grade level.[9] We use seven indicators variables, defined in terms of 0.5 standard deviation increments, with four below the base category, and three above. Thus, we define the following two vectors of variables for math or reading:

LowMath (or Reading) = a vector with four elements denoting distance below grade level (below 1.5 SD, 1-1.5 SD below, .5-1 SD below, and 0-.5 SD below).

HighMath (or Reading)= a vector with three elements denoting distance above grade level (above 1.5 SD, 1-1.5 SD above, and .5-1 SD above).

Interaction terms—Of most interest for this study is how accountability pressure affects the distribution of test scores within the schools feeling that pressure. To capture these distributional effects, we define for each subject vectors of interaction terms between place in the achievement distribution and the type of accountability pressure:

FailAYP*LowMath (or Reading)
FailAYP*HighMath (or Reading)
FailGrowth*LowMath (or Reading)
FailGrowth*HighMath (or Readng)
FailBoth*LowMath (or Reading)
FailBoth*HighMath (or Readng)

This flexible specification permits us to examine directly any nonlinearities in the distributional effects, and in particular to look for evidence of educational triage, in the schools facing three types of accountability pressure: 1) pressure from the status model only, 2) pressure from the growth model only, and 3) pressure from both status and growth models.

[9] This North Carolina standard of proficiency corresponds roughly to the “Basic” level of performance on NAEP, commonly referred to the nation’s report card, not the higher “Proficient” standard.

Estimation strategy

Following standard practice in the modeling of student achievement, we begin with the following value-added model of the distributional effects of accountability pressure (here denoted by a generic accountability pressure term, *AP*):

$$(2) Ach_{ijt} = \alpha + \beta_1 \boldsymbol{AP}_{jt-1} + \beta_2 \boldsymbol{Low}_{ijt-1} + \beta_3 \boldsymbol{High}_{ijt-1} + \beta_4 \boldsymbol{AP}_{jt-1} * \boldsymbol{Low}_{ijt-1} + \beta_5 \boldsymbol{AP}_{jt-1} * \boldsymbol{High}_{ijt-1} + \beta_6 \boldsymbol{X}_{ijt} + \beta_7 \boldsymbol{S}_{ijt} + \beta_8 Ach_{ijt-1} + u_{ijt}$$

where Ach_{ijt} is the student's achievement in reading or math in the current year; Ach_{ijt-1} is the student's achievement in the prior year; $\boldsymbol{AP}_{jt-1}$ is school *j*'s accountability status vector from the prior year; the vectors $\boldsymbol{Low}_{ijt-1}$ and $\boldsymbol{High}_{ijt-1}$ denote the student's position in the prior test score distribution; $\boldsymbol{X}_{ijt}$ is a vector of student control variables such as gender, race and poverty status; $\boldsymbol{S}_{ijt}$ is a vector of school control variables; and u_{ijt} is an error term.

A number of statistical issues arise in the estimation of this model. First, although this value-added specification is quite common in the literature, including the lagged dependent variable on the right hand side may bias the estimated coefficients because of its correlation with the error term (Todd & Wolpin, 2003). Yet controlling for prior achievement is a useful way to account for the cumulative nature of the education process because it controls for the learning that the student brings to the classroom in year *t*. Its coefficient, β_8, is expected to be less than 1 (usually between 0.6 to 0.8) because there is likely to be some decay of knowledge from one year to the next (Clotfelter *et al.*, 2007). To eliminate problems associated with using a lagged dependent variable as a predictor, some researchers move the lagged term to the left hand side of the equation and estimate a gain-score model in which the dependent variable is the change in test score ($\mathrm{Ach}_{\mathrm{ijt}}$ - $\mathrm{Ach}_{\mathrm{ijt-1}}$) from one year to the next. If the assumptions underlying the original model are correct, however, the gains model is misspecified in that it forces the coefficient of the prior year test score to be 1. So instead of using a gain-score specification in this study we present results from two models: the value added specification in equation two, above, and a levels specification, which is identical except that it excludes the lagged achievement variable and hence avoids the bias associated with that variable. The levels model is reasonable in this context because, as we explain next, we also include student fixed effects.

A second threat to the validity of the model is the problem of negative selection of students into schools facing accountability pressure, which would downwardly bias the main effects of accountability pressure and the interaction terms associated with them. Because accountability pressure is not randomly distributed among schools, one might expect, for example, that achievement levels in year *t* would be *lower* in the schools designated as failing in the prior year than in other schools. The concern here is primarily one of *adverse* (or negative) student selection, namely that the students with low test scores or low expected gains in test scores may be overrepresented in the schools designated as not meeting performance or growth standards. One way to address this problem is to include in the equation as part of the **X** vector a sufficiently large number of student-specific characteristics that the remaining correlation between the error term and the school level accountability variables is kept to a minimum. Such variables might include, for example, the race of the student, characteristics of the student’s family such as their income and education levels, and special characteristics of the students such as their participation in programs for gifted students or for students with special education needs. Even rich student level data, however, are unlikely to fully solve the problem because some of the relevant student characteristics, such as ability and motivation, are typically unobserved.

Longitudinal data permits a reasonable solution to this problem. Because we have observations on multiple test scores in each subject for each student, we can include student fixed effects These fixed effects control for all the time-invariant characteristics of students, both those that otherwise might have been measurable and those that are not. Along with the student fixed effects, we also include student-level variables that change over time. Among these variables are participation in special education programs of various types, and whether the student is new to the school in the particular year. Such a strategy is not without a cost; it means that any effects of accountability pressure are identified not by all students in the sample but rather by those who have at least two consecutive test scores and whose schools' accountability status changes from year to year.

An additional statistical problem is mean reversion, which if unaddressed would distort our estimates of distributional effects. Most likely, some students who performed well below grade level in a particular year did so not because of low true achievement but because random factors, such as a bad night's sleep or a headache, reduced their score. Such students are likely to have *larger* test score gains than other students due to regression to the mean. The converse is true for some students who performed well above grade level in a particular year; such students would have *smaller* gains than other students due to mean reversion. In sum, this bias would lead to a negative relationship between prior achievement and the following year's test score gain. To reduce bias from mean reversion, we define each student's position in the prior year achievement distribution using his or her performance in the other subject. In other words, in the models explaining math achievement, we place students in the prior year distribution based on reading scores rather than math scores, and vice versa. The underlying assumption is that a student's position in the distribution of one subject is highly correlated with position in the other subject, which is reasonable given that reading and math test scores are highly correlated in this sample (r=.73 to .78 depending on the grade level and year). Also underlying this approach is the assumption that the measurement error in math is uncorrelated with the measurement error in reading, an assumption that is not fully satisfied in our data.[10] Nonetheless, we believe this approach goes a long way to minimizing the impact of mean reversion, a conclusion that is supported by the patterns we report below.

The final threat to validity is the possibility that school-level confounders, such as concentrations of low-performing students, could bias the estimates of the accountability pressure and distributional effects. For this reason, we control for the concentration of minorities, limited English proficient students, and a measure of the diversity of the student population, the number of numerically accountable subgroups according the

[10] To test whether measurement error was correlated across reading and math scores, for each subject we computed the three-year average of lagged score, current score, and next year's score and deviated the current year's score from this average to compute the error from the three-year average. The reading and math errors are positively correlated (Pearson's *r* ranges from .18 to .26, depending on the grade level and year).

NCLB guidelines.[11] Preliminary analysis indicates that schools with large numbers of accountable subgroups are much more likely to fail AYP than less diverse schools.

In sum, we present student fixed effects models that adjust for mean reversion, include time varying student covariates and school-level controls for minority composition and diversity, and report results from both *value added* models and *levels* models.

Complete model

The complete value added model for math takes the following form (with a comparable model for reading):

$$
\begin{aligned}
(3) Stdmath_{ijt} \\
= {} & \alpha + \beta_1 FailAYP_{ijt-1} + \beta_2 FailGrowth_{ijt-1} + \beta_3 FailBoth_{ijt-1} \\
& + \beta_4 \boldsymbol{LowRead}_{ijt-1} + \beta_5 \boldsymbol{HighRead}_{ijt-1} \\
& + \beta_6 FailAYP * \boldsymbol{LowRead}_{ijt-1} + \beta_7 FailAYP * \boldsymbol{HighRead}_{ijt-1} \\
& + \beta_8 FailGrowth * \boldsymbol{LowRead}_{ijt-1} \\
& + \beta_9 FailGrowth * \boldsymbol{HighRead}_{ijt-1} \\
& + \beta_{10} FailBoth * \boldsymbol{LowRead}_{ijt-1} \\
& + \beta_{11} FailBoth * \boldsymbol{HighRead}_{ijt-1} + \beta_{12} \boldsymbol{X}_{ijt-1} + \beta_{13} \boldsymbol{S}_{ijt-1} \\
& + \beta_{14} Stdmath_{ijt-1} + \gamma_t + \delta_i + u_{ijt}
\end{aligned}
$$

where the dependent variable is the standardized math score and the one-year lagged standardized math score is included as a covariate. The accountability pressure variables, *FailAYP*, *FailGrowth*, and *FailBoth*, indicate the type of pressure facing the school in the prior year. The levels model is identical except for the exclusion of the lagged dependent variable.

[11] NCLB regulations define nine accountable subgroups (white, black, Hispanic, Native American, Asian, Multiracial, economically disadvantaged, limited English proficient, students with disabilities). In North Carolina, a school must have at least 40 students to be held accountable in AYP calculations. Because student free/reduced priced lunch status was unavailable for 2007, we are unable to compute and control for the fraction of the student population which receives free or reduced priced lunches.

All positional vectors and accountability pressure variables are entered with a one-year lag because school ratings are released in the Spring and Summer prior to the target school year. Of particular interest are the coefficient vectors β_6- β_{11} which represent the distributional effects of the accountability system in the schools facing accountability pressure. As noted above, ***X*** is a vector of time-varying student characteristics and ***S*** is a vector of time-varying school characteristics. The vectors γ_t and δ_i represent year fixed effects and student fixed effects, respectively. The final term is the error term. We have included year effects to control for any year-specific effects on achievement.[12]

The descriptive information in table 2 shows that on average only 22 and 19 percent of students in reading and math, respectively, were below grade level between 1997 and 2007, a finding consistent with that the observation that North Carolina's proficiency level is set at a relatively low level.[13] The valid N for the achievement positional indicators are lower than for other variables because they are entered in lagged form and the prior year test score for a student's first test is missing by definition. On average over the period 1997 to 2006, 22% of students attended a school that failed AYP, 32% attended a school that failed the growth standard and 10% attended a school that

[12] Estimating this model using fixed effects to transform this equation to a within-child estimator of the effect of within-child deviations from student means on the outcome and covariates produces consistent results which adjust for the negative selection of students into low achieving schools under the following assumptions (Todd & Wolpin, 2003): 1) The effect of student fixed characteristics such as ability is independent of age; 2) Future school choices are invariant to prior achievement outcomes; and 3) The effect of school inputs are independent of age. We plan to explore the tenability of these assumptions in future work.

[13] Comparing results from the state assessment to NAEP scores is one way to determine the relative rigor of North Carolina's proficiency levels. On NAEP in 2007, 34% of 8th graders were at or above proficient in math and 28% were at or above proficient in reading. On the NC state assessments in 2007, 63% of 8th graders were above grade level in math and 86% were above grade level in reading.

failed both standards.[14] The sample is 30% black, five percent Hispanic, five percent Other (Asian and American Indian), and 46% received a free or subsidized lunch (not shown).

Results

Table three reports both the main effects and the distributional effects of accountability pressure for the period 1998-2007. Because the dependent variables are standardized by grade level and year, the regression coefficients represent the effect of being in a particular category relative to the base category on the student's test scores, as measured in terms of fractions of a standard deviation. The levels specification estimates the change in student achievement relative to students' own average achievement, as captured by the student fixed effects. The value-added fixed effects specification estimates the change in student achievement relative to the student's own average achievement, controlling for the students' lagged achievement. We use the shorthand of *test score gain* to describe changes in outcomes for estimates from either model.

Because of the interaction terms in the model, the coefficients of the accountability pressure variables indicate the average relationship between students' current year test scores and being in a school facing accountability pressure compared to being in a school not facing accountability pressure only for students with prior year test scores in the base category, namely between the grade level cut score and ½ standard deviation above it. The effects on students elsewhere in the prior achievement distribution can be derived from these main effects of accountability pressure combined

[14] During the period 2003 to 2006, 50% of students attended a school that failed AYP, 42% attended a school that failed the growth standard, and 29% attended a school that failed both standards.

with the coefficients of the interaction terms. Analogously, the coefficients on the prior achievement variables indicate the relationship between students' current test scores and being low or high in the distribution of the prior year test scores relative to being near the grade level cut score only for students in schools not facing accountability pressure.

The first row in table 3 indicates that students in the base, or reference category, in schools facing pressure from failing the growth standard alone exhibit lower student achievement in subsequent years in both math and reading than comparable students in schools not facing such pressure, with the negative coefficients being larger in math than in reading. Although it might be tempting to interpret these negative coefficients as evidence that an accountability system based on a growth approach reduces overall student achievement, that interpretation would not be correct. In fact it is not possible to determine the overall effects of the accountability system from this type of analysis. The most we can do is to identify results for one type of school relative to another. Hence, the negative coefficients of the failed growth variables simply indicate that students in the reference category in schools that meet growth targets — that is, schools that generate bonuses for teachers — attain higher test score growth in the subsequent year than do students in schools failing to meet their growth targets.

Facing pressure under the AYP, or status, standard alone is associated with higher student achievement in subsequent years in reading, and lower achievement in math, though only the negative coefficient in model two is significant. Students in schools facing pressure from both AYP and the state's growth standard perform somewhat less well in subsequent years in reading and roughly the same in subsequent years in math (coefficients are negative but not distinguishable from zero).

The next set of variables — the achievement indicators — yield the distributional patterns for students in schools facing no negative accountability pressure. The coefficients indicate that initial status is reinforced as students progress through such schools. In other words, lower achieving students tend to have lower gains relative to students closer to grade level. Higher achieving students tend to have higher test score gains relative to students closer to grade level. For example, column one indicates that students in the lowest category of the prior performance distribution exhibit subsequent math scores that are 0.05 SD below a student who performs at grade level. The same model indicates that students in the highest category of the prior year distribution exhibit test scores that are 0.06 SD above those of students who performs at grade level. This pattern holds in both reading and math, with the value added coefficients (models two and four) being slightly larger than the levels coefficients (models one and three). Figure 4 plots the coefficients for each of the four models. The key point is that the lines all are sloping upward, a finding that suggests that we have successfully countered much, if not all, of any bias related to regression to the mean.

Of most interest are the coefficients of the interaction terms. As shown in the next three panels of the table and figures 5-7, the differential distributional effects in schools facing accountability pressure follow a very different pattern from those shown in figure 4. All these estimates represent differences in outcomes relative to what would have happened in the absence of the negative pressure of the specific accountability system, that is relative to the patterns displayed in figure 4.

In schools failing only the growth standard, students with low prior achievement tend to gain more in math than students with middle and high prior achievement (see figure 5). Students with the lowest levels of prior achievement (below 1.5 SD below grade level) exhibit the largest gains, with the point estimates declining monotonically as the prior-year achievement category approaches the grade level cut point. Among high achievers, only students in the very top tail gain slightly relative to students just above grade level. This pattern suggests that in response to a growth standard of accountability, schools apparently find it easier to raise test scores at the bottom and at the very top relative to students at grade level, and do not focus additional attention on students close to the proficiency standard. In reading, little or no evidence emerges of significant distributional effects associated with pressure from the state's program alone, although slightly higher reading gains do emerge for high achieving students.[15]

With respect to status pressure, the table shows that schools failing only AYP generate positive gains in both math and reading among low performing students but negative effects in reading among high achieving students (see figure 6). The pattern at the bottom of the distribution is fully consistent with the prediction that under a status approach to accountability, schools facing pressure to raise student achievement to the proficiency standard would focus attention on students below it. The large negative effects in reading at the high end of the distribution indicate that the gains at the bottom in that subject have come at the expense of higher achieving students in the affected schools, a finding that echoes Reback's (2008) analysis of accountability in Texas described above. In summary, for math AYP pressure is associated with test score gains for low achieving students, with no test score losses for high achieving students, while in

[15] Restricting the analysis to the pre NCLB period (1997-2002) does not alter the patterns. The distributional patterns of growth based accountability are similar to those in Table 3, albeit with slightly larger coefficients for the low-performing students. Results are available from authors upon request.

reading AYP pressure is associated with test score gains for low achieving students and rather large test score losses for high achieving students.

Schools failing both AYP and the state's growth standard tend to exhibit positive gains in math for low achieving students in the bottom tail (below one SD below grade level) as well as negative effects in that subject among high achieving students (see figure 7). In reading no statistically significant distributional effects emerge from such schools, except for in the top tail (more than 1.5 SD above grade level) where students gain slightly relative to middle and low achieving students.

In summary, accountability pressures appear to offset at least to some extent the distributional patterns that favor high achievement students in schools not facing negative accountability patterns. Rather than reinforcing initial status, accountability pressure from either the federal status-based program alone or the state growth-based program, alone is associated with test score gains for low achieving students in both math and reading, with the effects stronger and more consistently statistically significant for math than for reading. Only in reading and only in response to status pressure do the gains at the bottom come at the expense of students with high prior year test scores. We speculate that the strikingly different patterns for math and reading under status based accountability reflects the particular challenges that schools face in raising student achievement in reading relative to math given the differentially important contribution of family background to reading achievement. Efforts to raise reading scores among the low performers apparently come not from improved instructional approaches that benefit all students but rather from shifts of resources from one group to another.

The distributional patterns for schools facing both kinds of accountability pressure are somewhat harder to explain. Consistent with the view that it is difficult for schools to raise reading scores, we find no evidence that such schools succeed in raising reading scores more at the bottom end of the distribution relative to the middle and upper ends. Somewhat inconsistent with our speculations in the previous paragraph, however, is that the gains in math at the bottom end appear to come at the expense of the students at the higher end of the distribution in such schools. Our only explanation is that such schools are under such pressure to raise student achievement that even in math, they do so by shifting resources from the higher performing to the lower performing students.

Aggregate Achievement Tradeoffs

The analysis to this point focuses on average achievement effects of accountability pressure by prior achievement category. Table 4 compares the weighted gains or losses to students below grade level to the weighted gains or losses to those above grade level and highlights the nature of the distributional tradeoffs that emerge from the two accountability regimes. The entries in the table are the coefficients of the interaction terms from table 3 weighted by the estimation sample proportion of students in each category and summed for the categories above and below grade level.[16] Because only about 25 percent of the students in schools facing growth pressure were below grade level (based on the prior year achievement distribution) during the study period, in contrast to about 75 percent above grade level, all the entries for low achieving students

[16] The aggregate weighted coefficient for the below category is $\beta_{below} = \sum_{k=1}^{4} \beta_k p_k$, where *k* indexes prior achievement categories and p_k is the proportion of students in category *k*. The aggregated weighted coefficient for the above category is $\beta_{above} = \sum_{k=1}^{3} \beta_k p_k$.

are far smaller (in absolute value) relative to the entries for high achieving students than would be suggested by the relevant coefficients in table 3.

In schools facing growth pressure alone, the combined results for math and reading (see top panel in the final two columns) indicate that both low and high achieving students gain relative to those just above grade level, with the test score gains to students above grade level being somewhat larger than for those below. In math, the gains for those above grade level are approximately equal to the gains for those below grade level, while in reading the gains for those above grade level far exceed those for students below grade level.

A very different pattern emerges for schools facing status pressure. In schools facing status pressure alone, the combined math and reading test score gains of students below grade level are more than offset by the test score losses to students above grade level (see final two columns of the second panel). Depending on the model specification (levels versus value added), the test score losses among high achievers are between 2.4 and 2.9 times the gains among low achievers. Combining math and reading, however, masks the differences between the two subjects noted earlier. In math, no tradeoff emerges between gains at the bottom versus losses at the top. In reading, in contrast, gains for students below grade level are negligible compared to the far larger test score losses among those above grade level. Similarly, in schools facing both growth and status pressure, the combined results indicate that students below grade level gain essentially nothing, while the test scores of students above grade level fall. Hence the patterns in these schools are more similar to those from schools facing status pressure than to those from schools facing growth pressure.

Thus, we conclude that the distributional effects clearly differ between the two types of accountability regimes. Though the achievement results are quite similar for students below grade level, the results differ quite substantively for students above grade level. Schools under pressure to meet the state's growth standard were apparently able to raise achievement at the bottom of the distribution without reducing it at the top. That was not the case for schools under pressure from the NCLB type status pressure, where students at the top of the distribution, especially in reading, were harmed relative to those at the middle and the bottom of the distribution.

Conclusion

Many educational policy makers currently view school accountability as a crucial component of any school reform strategy. As a result, holding educators accountable for student learning is now a part of all state and federal educational policy. There are two main metrics for holding educators responsible for student learning at the school level: status, which measures average achievement or percent of students at grade level, and growth, which measures the average achievement growth of students during the year. In this study, we compare and contrast how these two types of accountability pressure affect student achievement at different points in the achievement distribution in the schools under the most pressure to raise achievement. We conduct the study in North Carolina where schools have been subject to both types of accountability.

Using a ten-year panel dataset and fixed effects models of student achievement gains in reading and math we find that neither type of school-based accountability system generates distributionally neutral effects on student achievement in the schools subject to accountability pressure. Moreover, the distributional effects differ depending on whether the system holds schools accountable for the growth or the status of their students' learning. This first conclusion should not be surprising. It simply reflects the fact that educators do indeed respond to incentives, and that the incentives to pay attention to students at different points of the achievement distribution differ between the two approaches. The policy challenge is to design a system consistent with the goals of the policy.

Second, we find that under both approaches to accountability, the average gain to a student below the proficiency standard is larger than to a student at or above the standard. Nonetheless, in the case of the growth approach, the overall distributional effect is to raise student achievement in the aggregate slightly more at the high end of the distribution than at the low end relative to the students in the reference category. This outcome occurs because of the far larger number of students above grade level than below in the relevant schools. How to evaluate this aggregate distributional pattern depends on the goal of the accountability program. It could represent a shortcoming if the main goal of the program were to close achievement gaps. No such concern arises if the main goal is to raise student achievement throughout the achievement distribution within the low-performing schools; the patterns simply indicate the empirical fact that schools do so by raising achievement at both ends of the distribution.

Third, perhaps the most striking finding to emerge from this study is that status based accountability pressure generates large adverse effects for students above the proficiency threshold in reading. No such negative effects emerge with respect to the growth approach. One possible explanation for the differentially adverse effects in reading relative to math is that schools may try to improve reading scores of low performers by using student-specific strategies that reduce resources available to other students while in math they may use more general strategies, including better instruction for all students. Whether the gains to students below proficiency are worth the costs to students above proficiency is a question of values. If policy makers place equal value on effects at different points of the prior year achievement distribution, the net distributional effects of the program would clearly be negative. Based on the results combined for math and reading in table 5, the net distributional effects of the status based accountability program would be positive only if policy makers weighted the gains to low achievers 2-3 times as much as the costs to high achievers.

Fourth, we find little or no evidence of educational triage in connection with either approach to accountability. In particular, under both types of accountability in North Carolina, even students at the very bottom of the achievement distribution in schools facing accountability pressure experience positive achievement gains on average relative to students at or above grade level. This finding contrasts with much, but not all, of the prior literature that examines how schools respond to status based accountability systems. One possible explanation for the difference is North Carolina's relatively low proficiency standard. It may well be that in this state, raising students up to the proficiency standard is more feasible than in other states with higher standards.

Because this study is specifically designed to focus on the distributional effects of the two types of accountability in schools facing negative accountability pressure, we are not able to make any statements about the overall or average achievement effects of either type of program. Measuring overall effects would require a completely different type of study design, such as those used in the cross-state studies described earlier.

Moreover, we cannot say anything about a variety of other policy relevant considerations that arise with school-based accountability systems, such as the potential narrowing of the curriculum and the tendency for schools to respond to accountability pressure by moving students into special education programs.

Nonetheless, we believe the distributional patterns highlighted in this study should play a central role in policy debates about school accountability. School based accountability is motivated at least in part by a desire to narrow achievement gaps by raising the performance of students at the bottom of the achievement distribution relative to those at the top. It operates, however, by putting more pressure on some schools relative to others, rather than by placing pressure directly on students. There would be no within-school distributional effects if schools responded to accountability pressure by improving outcomes equally for all students. In fact, however, this study shows that schools facing pressure to improve apparently focus more attention on some students than on others, and that the distributional patterns differ both by the type of accountability system and by subject. To the extent that positive distributional effects emerge for students at the bottom of the distribution in the affected schools, with no adverse effects on higher performing students in those schools, an accountability system has the potential to close test score gaps in ways that may be deemed desirable by most observers. To the extent that it raises achievement for some students, but lowers it for others, as in the case for reading in status-based accountability system, in contrast, there are clear tradeoffs that require additional policy discussion and debate.

References

Amrien, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), Retrieved 12/23/08 from http://epaa.asu.edu/epaa/v10n18/.

Balfanz, R., Legters, N., West, T. C., & Weber, L. M. (2007). Are nclb's measures, incentives, and improvement strategies the right ones for the nation's low-performing high schools? *American Educational Research Journal, 44*(3), 559-593.

Ballou, D., & Springer, M. G. (2008). Achievement trade-offs and no child left behind: Peabody College of Vanderbilt University.

Booher-Jennings, J. (2005). Below the bubble: "educational triage" and the texas accountability system. *American Educational Research Journal, 42*(2), 231-268.

Burgess, S., Propper, C., Slater, H., & Wilson, D. (2005). Who wins and who loses from school accountability? The distribution of educational gain in english secondary schools. *The Centre for Market and Public Organisation (series), 05*(128).

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305-331.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Cohen, D. K. (1996). Standards-based reform: Policy, practice, and performance. In H. Ladd (Ed.), *Holding schools accountable* (pp. 99-127). Washington, DC: Brookings Institution Press.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106*(6), 1047-1085.

Figlio, D. (2005). Measuring school performance: Promise and pitfalls. In L. Stiefel (Ed.), *Measuring school performance and efficiency: Implications for practice and research*. Larchmont, NY: Eye on Education.

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

Figlio, D. N., & Ladd, H. (2008). School accountability and student achievement. In H. Ladd & E. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 166-182). New York and London: Routledge.

Hamilton, L., Berends, M., & Stecher, B. (2005). *Teachers' responses to standards-based accountability*. Santa Monica, CA: RAND Corporation.

Hanushek, E. A., & Raymond, M. A. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, *89*(5-6), 761.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*(4), 91-114.

Krieg, J. M. (2008). Are students left behind? The distributional effects of the no child left behind act. *Education Finance and Policy, 3*(2 (Spring)), 250-281.

Ladd, H. F. (1996). *Holding schools accountable: Performance-based reform in education*. Washington, D.C.: Brookings Institution.

Ladd, H. F. (1999). The dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review, 18*(1), 1-16.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review, 21*(1), 1-17.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Neal, D., & Schanzenbach, D. W. (Forthcoming). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts america's schools*. Cambridge, Mass.: Harvard Education Press.

O'Day, J., & Smith, M. (1993). Systemic reform and educational opportunity. In S. Furman (Ed.), *Designing coherent educational policy: Improving the system* (pp. 250-312). San Francisco: Jossey-Bass.

Orfield, G., & Kornhaber, M. L. (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: Century Foundation Press.

Peterson, P., & Hess, F. (2006). Keeping an eye on state standards. *Education Next, 6*(3), 28-29.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal Of Public Economics, 92*(5-6), 1394-1415.

Toch, T., & Harris, D. (2008). Salvaging accountability. *Education Week, 28*(6), 36.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113*(485), F3-F33.

Valenzuela, A. (2005). *Leaving children behind: How "texas-style" accountability fails latino youth*. Albany: State University of New York Press.

Weitz, K., & Rosenbaum, J. (2007). Inside the black box of accountability: How high stakes accountability alters school culture and the classification and treatment of students and teachers. In A. S. e. al. (Ed.), *No child left behind and the reduction of the achievement gap* (pp. 97-116). New York and London: Routledge.

Figures

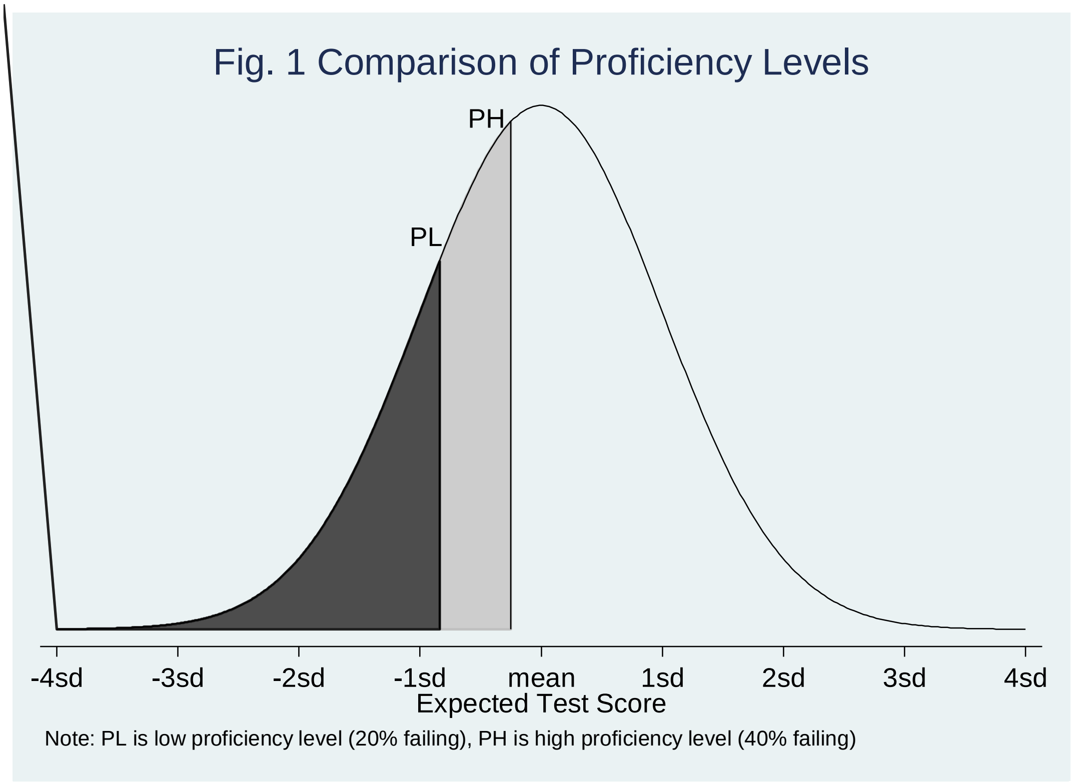
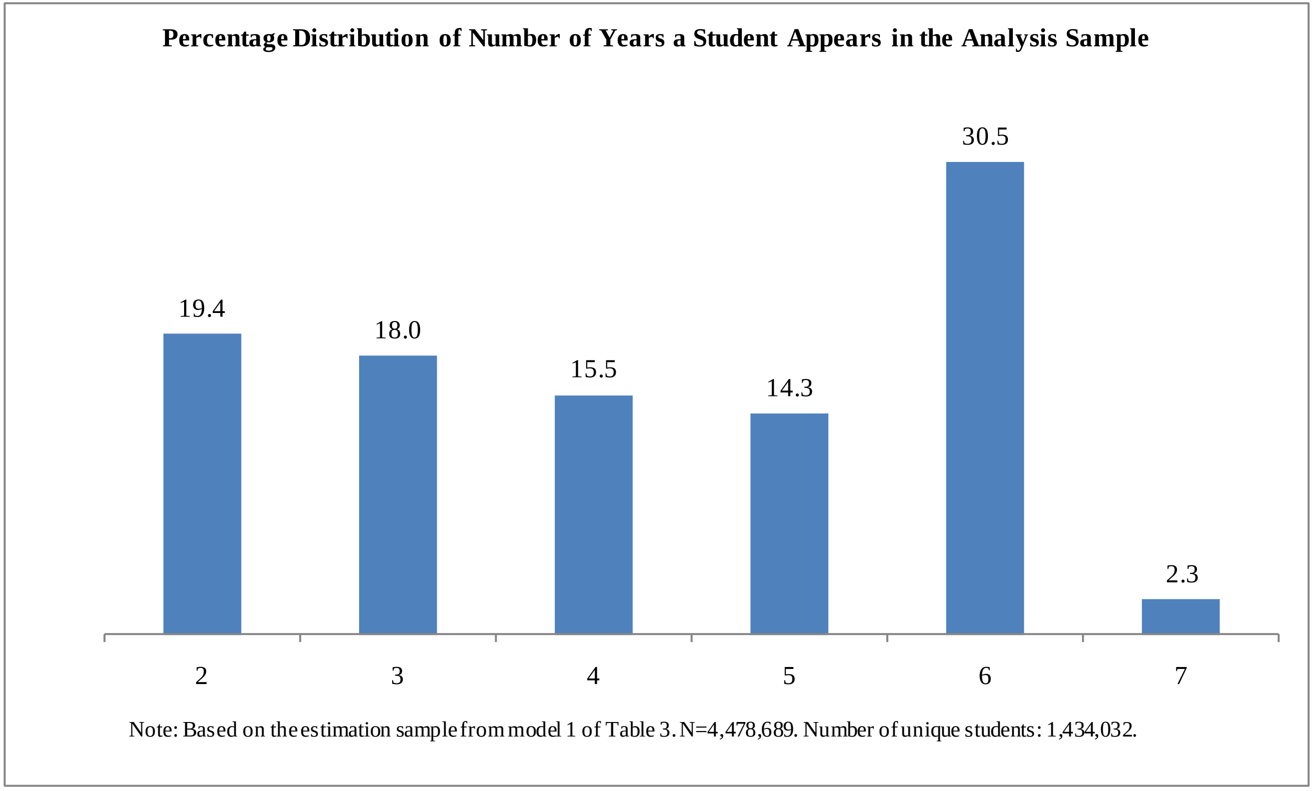


Fig. 1 Comparison of Proficiency Levels

Note: PL is low proficiency level (20% failing), PH is high proficiency level (40% failing)

Figure 2.



Note: Based on the estimation sample from model 1 of Table 3. N=4,478,689. Number of unique students: 1,434,032.

Figure 3.

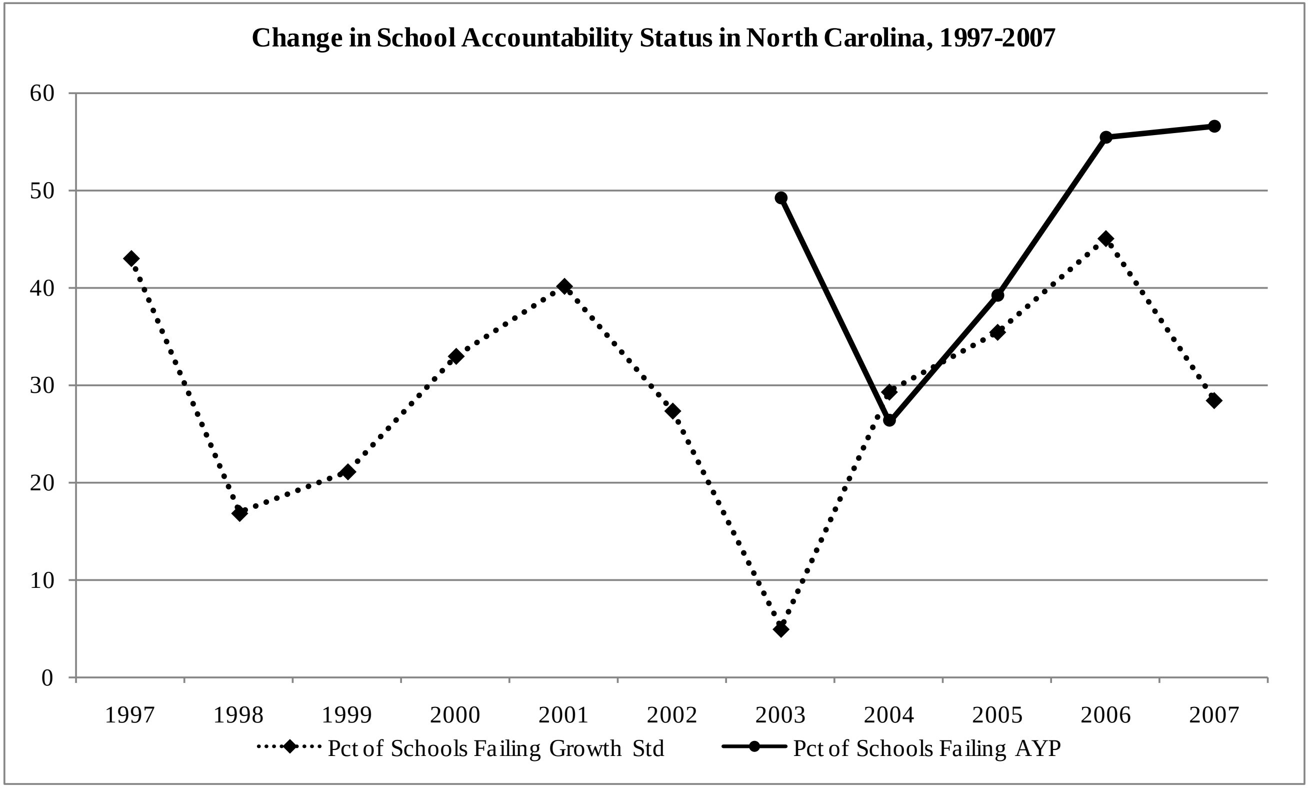


Figure 4.

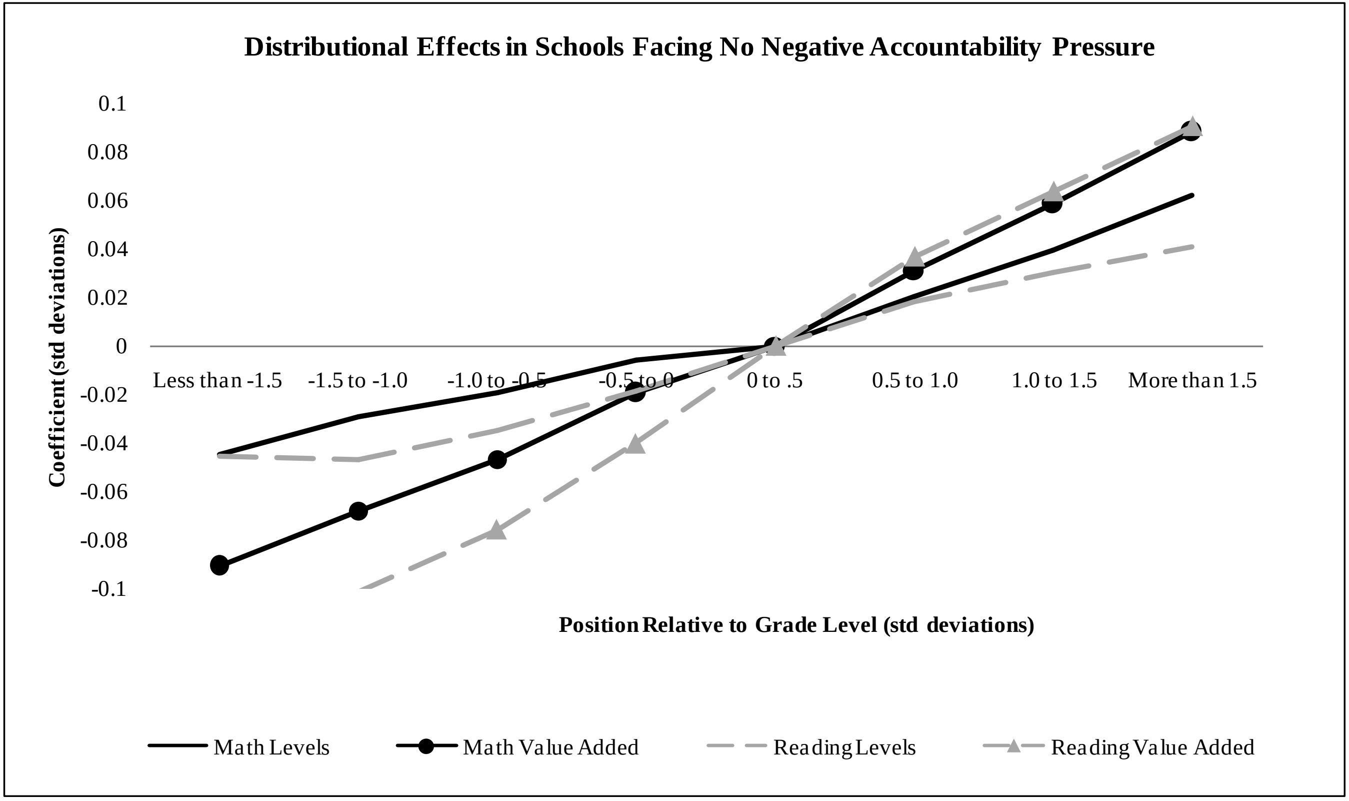


Figure 5.

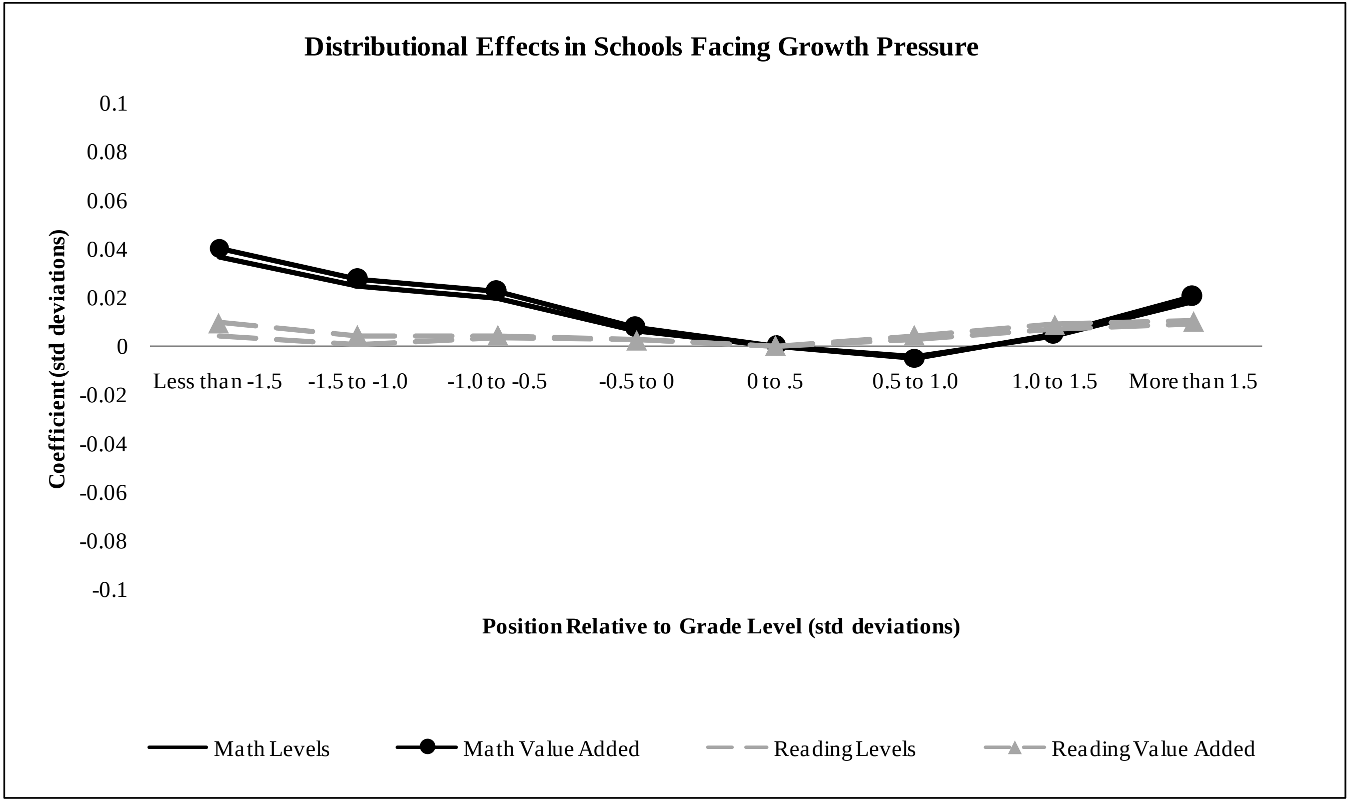


Figure 6.

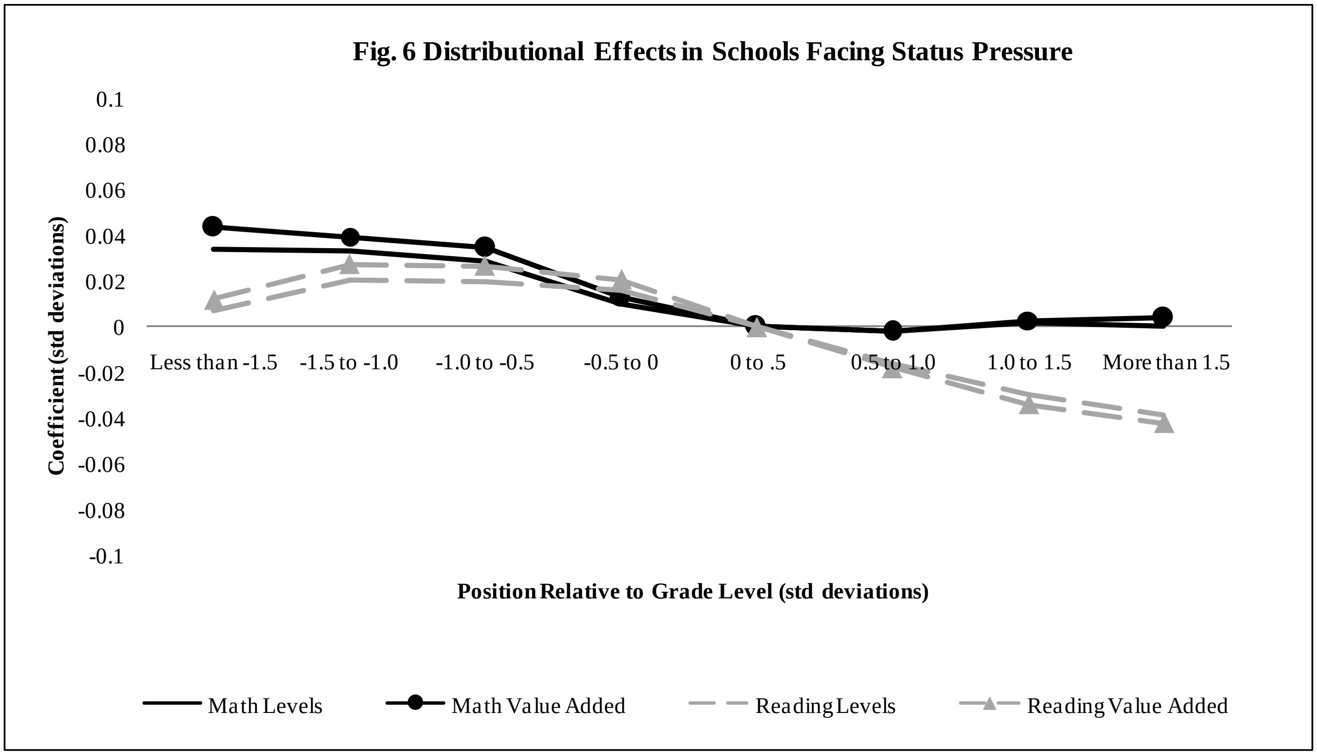
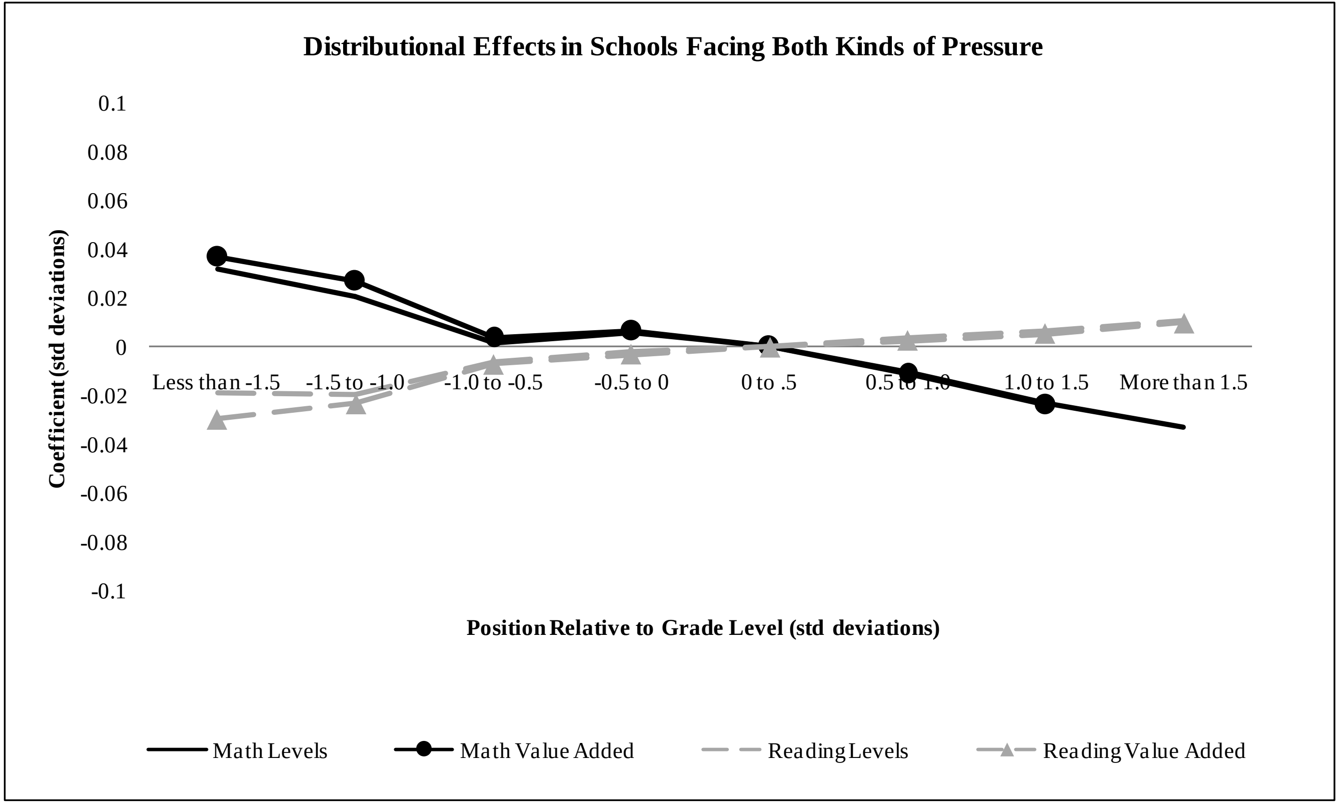


Figure 7.



Tables

Table 1. Number and Percent of Students and Schools by School Accountability Status and Year

	Students - Combined School Accountability Status					Schools - Combined School Accountability Status				
Year	made both	failed ayp	failed growth	failed both	Total	made both	failed ayp	failed growth	failed both	Total
1997	315,865	0	246,600	0	562,465	902	0	682	0	1,584
	56.16	0	43.84	0	100	56.94	0	43.06	0	100
1998	485,733	0	90,867	0	576,600	1,358	0	275	0	1,633
	84.24	0	15.76	0	100	83.16	0	16.84	0	100
1999	487,393	0	102,163	0	589,556	1,292	0	346	0	1,638
	82.67	0	17.33	0	100	78.88	0	21.12	0	100
2000	396,488	0	208,994	0	605,482	1,141	0	562	0	1,703
	65.48	0	34.52	0	100	67	0	33	0	100
2001	350,187	0	272,029	0	622,216	1,068	0	718	0	1,786
	56.28	0	43.72	0	100	59.8	0	40.2	0	100
2002	445,296	0	183,196	0	628,492	1,266	0	477	0	1,743
	70.85	0	29.15	0	100	72.63	0	27.37	0	100
2003	239,361	349,308	5,737	42,235	636,641	909	814	17	73	1,813
	37.6	54.87	0.9	6.63	100	50.14	44.9	0.94	4.03	100
2004	295,833	80,484	104,273	159,555	640,145	1,077	218	276	264	1,835
	46.21	12.57	16.29	24.92	100	58.69	11.88	15.04	14.39	100
2005	252,807	132,325	77,127	180,475	642,734	857	332	265	391	1,845
	39.33	20.59	12	28.08	100	46.45	17.99	14.36	21.19	100
2006	157,980	191,989	57,904	222,002	629,875	572	444	261	584	1,861
	25.08	30.48	9.19	35.25	100	30.74	23.86	14.02	31.38	100
2007	205,930	278,253	19,383	151,352	654,918	728	611	91	446	1,876
	31.44	42.49	2.96	23.11	100	38.81	32.57	4.85	23.77	100
Total	3,632,873	1,032,359	1,368,273	755,619	6,789,124	11,170	2,419	3,970	1,758	19,317
	53.51	15.21	20.15	11.13	100	57.82	12.52	20.55	9.1	100

Table 2. Descriptive Statistics

Variable	Description	Obs (in millions)	Mean	SD
Dependent Variables				
stdmath	Standardized math test score	6.59	0	1
stdread	Standardized reading test score	6.56	0	1
Accountability Pressure				
notayp	School failed AYP std	6.14	0.219	0.414
notgrow	School failed growth std	6.12	0.317	0.465
notboth	School failed both AYP and growth std	6.14	0.096	0.295
Prior Achievement				
lr4	Less than -1.5 SD below grade level in reading	4.70	0.020	0.139
lr3	-1.5 to -1.0 SD below grade level in reading	4.70	0.037	0.188
lr2	-1.0 to -0.5 SD below grade level in reading	4.70	0.059	0.236
lr1	-0.5 to 0 SD below grade level in reading	4.70	0.097	0.296
hr2	0.5 to 1.0 SD above grade level in reading	4.70	0.195	0.396
hr3	1.0 to 1.5 SD above grade level in reading	4.70	0.184	0.387
hr4	More than 1.5 SD above grade level in reading	4.70	0.265	0.441
lm4	Less than -1.5 SD below grade level in math	4.71	0.008	0.090
lm3	-1.5 to -1.0 SD below grade level in math	4.71	0.024	0.154
lm2	-1.0 to -0.5 SD below grade level in math	4.71	0.055	0.228
lm1	-0.5 to 0 SD below grade level in math	4.71	0.100	0.300
hm2	0.5 to 1.0 SD above grade level in math	4.71	0.180	0.384
hm3	1.0 to 1.5 SD above grade level in math	4.71	0.175	0.380
hm4	More than 1.5 SD above grade level in math	4.71	0.309	0.462
Student Background				
gifted	Student was designated gifted	6.82	0.131	0.338
specialed	Student received special education services	6.72	0.138	0.345
currentlylep	Student showed Limited English Proficiency	6.82	0.022	0.148
newtoschool	Student was new to the school	5.23	0.353	0.478
black	Black student	6.80	0.298	0.457
hisp	Hispanic student	6.80	0.055	0.228
other	Other racial/ethnic background	6.80	0.051	0.220
male	Male student	6.80	0.512	0.500
School Background				
pctblack	% of students in school who are black	6.80	0.298	0.236
pcthisp	% of students in school who are Hispanic	6.80	0.055	0.064
pctlep	% of students in school who are LEP	6.82	0.022	0.039
subgroups	Number of subgroups in school	6.81	1.825	2.169

Notes: The ranges for all variables are 0 to 1 except stdmath, -4.66 to 3.65; stdread, -4.11 to 3.13; pcthisp, 0 to 0.720; pctlep, 0 to 0.605; subgroups, 0 to 8. Prior achievement indicator variables and distributional interaction terms require an additional lagged year for calculation, and thus differ in observations from accountability pressure variables. Descriptives for distributional interaction terms shown in table 6.

Table 3: Achievement Model, 1998-2007

	Math		Reading	
	(1)	(2)	(3)	(4)
	Levels	Value Added	Levels	Value Added
Accountability Pressure				
Failed Growth	-0.0630***	-0.0676***	-0.0325***	-0.0367***
	(0.00180)	(0.00180)	(0.00204)	(0.00204)
Failed AYP	-0.00349	-0.00712*	0.0187***	0.0204***
	(0.00280)	(0.00279)	(0.00299)	(0.00298)
Failed AYP & Growth	-0.00424	-0.00503	-0.0131**	-0.0130**
	(0.00398)	(0.00397)	(0.00411)	(0.00411)
Distributional Effects Among Schools Facing No Negative Pressure				
SD Below Grade Level				
Less than - 1.5	-0.0448***	-0.0912***	-0.0457***	-0.111***
	(0.00411)	(0.00412)	(0.00750)	(0.00763)
- 1.5 to -1.0	-0.0297***	-0.0681***	-0.0471***	-0.102***
	(0.00298)	(0.00298)	(0.00433)	(0.00436)
-1.0 to -0.5	-0.0198***	-0.0469***	-0.0348***	-0.0760***
	(0.00229)	(0.00229)	(0.00283)	(0.00284)
-0.5 to 0	-0.00602***	-0.0194***	-0.0185***	-0.0402***
	(0.00172)	(0.00172)	(0.00205)	(0.00205)
SD Above Grade Level				
0.5 to 1.0	0.0203***	0.0310***	0.0184***	0.0366***
	(0.00138)	(0.00137)	(0.00157)	(0.00157)
1.0 to 1.5	0.0395***	0.0585***	0.0304***	0.0638***
	(0.00153)	(0.00154)	(0.00167)	(0.00167)
More than 1.5	0.0618***	0.0883***	0.0408***	0.0905***
	(0.00168)	(0.00169)	(0.00179)	(0.00180)
Distributional Effects Among Schools Facing Growth Pressure Only				
Failed Growth*Position				
SD Below Grade Level				
Less than - 1.5	0.0365***	0.0405***	0.00410	0.00976
	(0.00578)	(0.00578)	(0.0115)	(0.0116)
- 1.5 to -1.0	0.0246***	0.0274***	0.000910	0.00432
	(0.00430)	(0.00430)	(0.00647)	(0.00649)
-1.0 to -0.5	0.0201***	0.0223***	0.00337	0.00447
	(0.00346)	(0.00345)	(0.00430)	(0.00429)
-0.5 to 0	0.00638*	0.00744**	0.00247	0.00276
	(0.00281)	(0.00280)	(0.00330)	(0.00329)
SD Above Grade Level				
0.5 to 1.0	-0.00445	-0.00470*	0.00292	0.00435
	(0.00231)	(0.00231)	(0.00262)	(0.00261)
1.0 to 1.5	0.00440	0.00483*	0.00695**	0.00888***
	(0.00239)	(0.00239)	(0.00263)	(0.00262)
More than 1.5	0.0183***	0.0203***	0.00936***	0.0105***
	(0.00233)	(0.00233)	(0.00244)	(0.00243)

Distributional Effects Among Schools Facing Status Pressure Only				*Table 3, continued*
Failed AYP*Position				
SD Below Grade Level				
Less than - 1.5	0.0341**	0.0438***	0.00680	0.0119
	(0.0121)	(0.0121)	(0.0172)	(0.0174)
- 1.5 to -1.0	0.0335***	0.0395***	0.0206*	0.0273**
	(0.00727)	(0.00724)	(0.00932)	(0.00935)
-1.0 to -0.5	0.0290***	0.0345***	0.0199**	0.0268***
	(0.00551)	(0.00549)	(0.00618)	(0.00620)
-0.5 to 0	0.00980*	0.0127**	0.0158***	0.0208***
	(0.00433)	(0.00431)	(0.00469)	(0.00469)
SD Above Grade Level				
0.5 to 1.0	-0.00198	-0.00205	-0.0160***	-0.0181***
	(0.00329)	(0.00328)	(0.00362)	(0.00361)
1.0 to 1.5	0.00146	0.00260	-0.0299***	-0.0342***
	(0.00327)	(0.00327)	(0.00355)	(0.00354)
More than 1.5	-0.000122	0.00381	-0.0388***	-0.0424***
	(0.00304)	(0.00304)	(0.00321)	(0.00321)
Distributional Effects Among Schools Facing Growth and Status Pressure				
Failed AYP & Growth*Position				
SD Below Grade Level				
Less than - 1.5	0.0322*	0.0368*	-0.0189	-0.0296
	(0.0164)	(0.0163)	(0.0228)	(0.0231)
- 1.5 to -1.0	0.0210*	0.0268*	-0.0194	-0.0232
	(0.0105)	(0.0104)	(0.0127)	(0.0127)
-1.0 to -0.5	0.00165	0.00388	-0.00591	-0.00690
	(0.00786)	(0.00783)	(0.00861)	(0.00862)
-0.5 to 0	0.00592	0.00639	-0.00184	-0.00307
	(0.00634)	(0.00632)	(0.00667)	(0.00666)
SD Above Grade Level				
0.5 to 1.0	-0.0105*	-0.0111*	0.00358	0.00260
	(0.00496)	(0.00494)	(0.00534)	(0.00532)
1.0 to 1.5	-0.0227***	-0.0241***	0.00694	0.00550
	(0.00497)	(0.00496)	(0.00532)	(0.00529)
More than 1.5	-0.0326***	-0.0345***	0.0111*	0.0101*
	(0.00464)	(0.00463)	(0.00479)	(0.00477)
Intercept	0.0148***	0.0137***	-0.0113***	-0.0141***
	(0.00207)	(0.00211)	(0.00220)	(0.00226)
N	4478689	4475403	4481213	4470088
R^2	0.2321	0.2043	0.076	0.4726

Notes: Student fixed effects models; the value added specification differs from the levels specification by the inclusion of a prior year test score in the specified subject; robust standard errors in parentheses; * p<0.05, ** p<0.01, *** p<0.001; The models also include the following controls: student was gifted, received special education services, had limited English proficiency, was new to the school, school percent black, percent Hispanic, percent limited English proficient, and year fixed effects. Base category: 0 to 0.5 SD above grade level.

Table 4. Weighted Distributional Effects by Accountability Regime

	Math		Reading		Math & Reading	
	Levels	Val Add	Levels	Val Add	Levels	Val Add
Growth Pressure						
Below	0.0041	0.0046	0.0006	0.0009	0.0024	0.0027
Above	0.0039	0.0043	0.0039	0.0047	0.0039	0.0045
Status Pressure						
Below	0.0033	0.0041	0.0034	0.0045	0.0034	0.0043
Above	-0.0002	0.0013	-0.0197	-0.0219	-0.0099	-0.0103
Growth and Status Pressure						
Below	0.0016	0.0019	-0.0016	-0.0020	0.0000	-0.0001
Above	-0.0160	-0.0169	0.0045	0.0038	-0.0057	-0.0065

Note: Cell values are calculated from the coefficients shown in table 3, weighted by the proportion of cases within each prior achievement category and summed across the categories above and below grade level.

Table 5. Descriptive Statistics of Distributional Interaction Terms

Variable	Description	Obs (in millions)	Mean	SD
nayplr4	notayp*lr4	4.61	0.002	0.050
nayplr3	notayp*lr3	4.61	0.006	0.079
nayplr2	notayp*lr2	4.61	0.011	0.106
nayplr1	notayp*lr1	4.61	0.019	0.137
nayphr2	notayp*hr2	4.61	0.047	0.211
nayphr3	notayp*hr3	4.61	0.045	0.208
nayphr4	notayp*hr4	4.61	0.078	0.268
nayplm4	notayp*lm4	4.63	0.002	0.048
nayplm3	notayp*lm3	4.63	0.007	0.083
nayplm2	notayp*lm2	4.63	0.014	0.119
nayplm1	notayp*lm1	4.63	0.025	0.156
nayphm2	notayp*hm2	4.63	0.042	0.201
nayphm3	notayp*hm3	4.63	0.039	0.194
nayphm4	notayp*hm4	4.63	0.072	0.259
nglr4	notgrow *lr4	4.60	0.009	0.092
nglr3	notgrow *lr3	4.60	0.015	0.121
nglr2	notgrow *lr2	4.60	0.024	0.153
nglr1	notgrow *lr1	4.60	0.038	0.190
nghr2	notgrow *hr2	4.60	0.067	0.250
nghr3	notgrow *hr3	4.60	0.056	0.229
nghr4	notgrow *hr4	4.60	0.072	0.259
nglm4	notgrow *lm4	4.61	0.004	0.060
nglm3	notgrow *lm3	4.61	0.011	0.105
nglm2	notgrow *lm2	4.61	0.024	0.153
nglm1	notgrow *lm1	4.61	0.042	0.201
nghm2	notgrow *hm2	4.61	0.063	0.243
nghm3	notgrow *hm3	4.61	0.053	0.225
nghm4	notgrow *hm4	4.61	0.077	0.266
nbothlr4	notboth *lr4	4.61	0.001	0.036
nbothlr3	notboth *lr3	4.61	0.003	0.055
nbothlr2	notboth *lr2	4.61	0.006	0.077
nbothlr1	notboth *lr1	4.61	0.010	0.098
nbothhr2	notboth *hr2	4.61	0.022	0.147
nbothhr3	notboth *hr3	4.61	0.020	0.140
nbothhr4	notboth *hr4	4.63	0.031	0.174
nbothlm4	notboth *lm4	4.63	0.001	0.038
nbothlm3	notboth *lm3	4.63	0.004	0.063
nbothlm2	notboth *lm2	4.63	0.008	0.088
nbothlm1	notboth *lm1	4.63	0.013	0.115
nbothhm2	notboth *hm2	4.63	0.020	0.139
nbothhm3	notboth *hm3	4.63	0.017	0.127
nbothhm4	notboth *hm4	4.63	0.026	0.159

Note: range for all interaction terms is 0 to 1. Entered as one-year lags.