

Status versus Growth: The Distributional Effects of School Accountability Policies

Helen F. Ladd
Duke University
hladd@duke.edu

Douglas L. Lauen
University of North Carolina at Chapel Hill
dlauen@unc.edu

Working Paper

**Please do not quote or cite without authors' permission*

Paper presented at the *NCLB: Emerging Findings Research Conference* at the Urban Institute, Washington, D.C. on August 12, 2009. The conference was supported, in part, by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), funded by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education and the National Center for Performance Incentives (NCPI) at Vanderbilt University. The authors gratefully acknowledge the research assistance of Steven M. Gaddis; the advice of Ashu Handa and Randall Reback; the Spencer Foundation for funding; and the North Carolina Education Research Data Center, housed at Duke University, for providing access to the data for this project. The views expressed in the paper are solely those of the authors and may not reflect those of the funders or supporting organizations. Any errors are attributable to the authors.

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 data set from North Carolina, the authors 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, but more clearly in math than in reading. In contrast to prior research highlighting the possibility of educational triage, the authors find little or no evidence that schools in North Carolina ignore the students far below proficiency under either approach. Importantly, they find that the status, but not the growth, approach reduces the reading achievement of higher performing students. The 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 and 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 and 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 NCLB 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 and Berliner 2002; Balfanz et al. 2007; Figlio 2005; Figlio 2006; Hamilton, Berends and Stecher 2005; Kane and Staiger 2002; Linn 2000; Peterson and 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 and Raymond 2005; Ladd and Walsh 2002; Toch and 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 counterproductive, 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 and Springer 2008; Booher-Jennings 2005; Krieg 2008; Neal and 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 these two types 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 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 reading and math test score data over time allows for careful modeling of student achievement gains, including the use of student and school fixed effects to account for time-invariant unobservable characteristics

of students such as their ability and schools such as teacher quality. Consistent with our theoretical predictions for the status program, we find evidence of positive distributional effects for students below the proficiency cut point for both subjects. The response to the growth-based program for these low achieving students is more mixed; clear positive effects emerge in math but not in reading. The approaches differ further with respect to their effects on student performance at the top of the prior-year achievement distribution, with the growth approach generating positive distributional effects for these students and the status approach either zero or negative distributional effects, with larger negative effects in reading than in math. In contrast to many other studies we find no evidence of educational “triage” for either program.

Achievement effects of school accountability programs

The most convincing analysis of how accountability affects overall achievement emerges 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 and Berliner 2002; Darling-Hammond 2004; Linn 2000; Nichols and Berliner 2007; Orfield and 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 and Rosenbaum 2007).

Our main question is how school accountability affects student 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 any gains to students at the bottom of the distribution come at the expense of those at the top in schools under accountability pressure? Fourth, 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 use different methodologies and data sets to 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.”¹ 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

¹ Although many sorts of triage processes are theoretically possible, we will adopt this term to mean “triaging out” students well below and well above grade level and “triaging in” students close to grade level.

test data from the Northwest Evaluation Association, Ballou and Springer (2008) find little or no evidence of adverse effects for the lowest performers.²

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.

²Burgess et al (2005), in a study of accountability in England also finds adverse effects of accountability on the lowest performing students. 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:³

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

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

³ 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 such students; without additional attention, some of them 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 by 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 to 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 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.⁴ 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

⁴ The expected average gains were 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 $TS_t - TS_{t-1} = a + bX_1 + cX_2$ 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 $t-1$, X_1 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 X_2 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.⁵

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

⁵ 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 in 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.⁶ Not surprisingly, a school that performs well under the state’s accountability system may do poorly under the federal system, and vice versa.

Figure 2 illustrates the variation over time in the percentages of schools failing to meet the two types of accountability standards by year. Included in the percentages are schools that failed both standards, a percentage that ranged from four percent in 2003 to 31 percent in 2006. The bottom line is that for the past 11 years, many schools in North Carolina have not met one or both of the standards for student achievement. How a

⁶ 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.

school's failure to meet a specific standard 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.⁷ 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 3 depicts the distribution of students by the number of years each appears in the mathematics analysis sample used to compute mathematics achievement gain.⁸ 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.⁹

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

⁷ 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.

⁸ 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.

⁹ We include retained and double promoted students in the analysis. Supplementary analysis with a retained indicator variable produces identical results to those reported below.

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.¹⁰ 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 score 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 growth targets in the years before AYP and *both* AYP and growth targets in the recent years 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,

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

¹⁰ The state rescaled the reading tests in 2003 and the math tests in both 2001 and 2006.

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.

Distributional variables—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.¹¹ 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 three mutually exclusive types of accountability pressure:

FailAYP*LowMath (or Reading)

FailAYP*HighMath (or Reading)

¹¹ 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.

FailGrowth*LowMath (or Reading)

FailGrowth*HighMath (or Reading)

FailBoth*LowMath (or Reading)

FailBoth*HighMath (or Reading)

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.

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):

$$(2) Ach_{ijt} = \alpha + \beta_1 AP_{jt-1} + \beta_2 Low_{ijt-1} + \beta_3 High_{ijt-1} + \beta_4 AP_{jt-1} * Low_{ijt-1} + \beta_5 AP_{jt-1} * High_{ijt-1}$$

where Ach_{ijt} is the student i 's achievement in year t in reading or math in the current year; Ach_{ijt-1} is the student's achievement in the prior year; AP_{jt-1} is school j 's accountability status vector from the prior year; the vectors Low_{ijt-1} and $High_{ijt-1}$ denote the student's position in the prior test score distribution; X_{ijt} is a vector of student control variables such as gender, race and poverty status; S_{ijt} is a vector of school control variables; and u_{ijt} is an error term.

Among the statistical concerns that arise in the estimation of this model, the two most serious relate to selection. One is the negative selection of students into schools facing accountability pressure that arises because low-ability students are more likely

than other students to attend such schools.¹² One way to address this negative selection is to include in the equation as part of the \mathbf{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 parents 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.

Our longitudinal data, with multiple test scores in each subject for each student, permits us to address this problem by including 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 these 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

¹² In regressions with test score as the dependent variable, the coefficient on the accountability pressure variable becomes less negative as we add student background variables, a pattern that clearly indicates negative selection.

but rather by those who have at least two consecutive test scores and whose school's accountability status changes from year to year.¹³

The other selection concern is that schools are not randomly assigned to accountability pressure at any one point in time or over time. As a result, school-level confounders, such as concentrations of low-performing students or low-quality teachers, could bias the estimates of the accountability pressure and distributional effects. For this reason, we include in the model school fixed effects that account for the time-invariant characteristics of schools as well as other school level variables that change over time. These include the concentrations of minorities and of limited English proficient students, and, as a measure of the diversity of the student body, the number of numerically accountable subgroups according the NCLB guidelines¹⁴ Preliminary analysis indicates that schools with large numbers of accountable subgroups are much more likely to fail AYP than are less diverse schools. In addition, we include year fixed effects to control for the possibility that there may be some year-specific factors correlated with accountability pressures that may affect student achievement.

In sum, our basic model is the value-added model described in equation (2) as augmented with student, school, and year fixed effects. Within this model, the effects of

¹³ Estimating the model using student fixed effects transforms the equation to a within-child estimator of the effect of within-child deviations from student means on the outcome and covariates. This model produces consistent results that adjust for the negative selection of students into low achieving schools under the following assumptions 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 (Todd & Wolpin, 2003).

¹⁴ 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.

accountability pressure are identified from changes in accountability pressure within a school for students who are in that school before and after the change. This focus on within- school effects is consistent with our interest in exploring how accountability pressure affects different groups of students within a school. In a variation of this basic model we also report levels models that are identical to the value added model but exclude the prior year achievement variable. Although we prefer the value-added model because it reflects the cumulative nature of the education process, some readers may object to including the lagged dependent variable as an explanatory variable on the ground that it is correlated with the error term and therefore may bias other coefficients (Todd and Wolpin 2003). The fact that student fixed effects are already included in the model, means that the only substantive loss from excluding the variable is the time varying component of achievement, which is subject to significant noise in any case.¹⁵

In addition to these two forms of the basic value added model, we also report results from an adjusted gains model. Following Reback (2008), for each subject we define a standardized adjusted gain score (AG) as the difference between a student's actual test score in year t and the expected score for students in the same grade who had the exact same score that she had in the previous year, normalized by the standard deviation of the scores in year t for that group of students. In symbols, we have:

¹⁵ Some researchers recommend addressing the problem by using the twice lagged dependent variable as an instrument or moving the dependent variable to the left hand side to make the dependent variable into a gain measure. Our short time periods for many students rule out effective use of the first strategy and we reject the second because it assumes the coefficient of the lagged variable in equation 1 is equal to 1, which would only be the case if there were no decay in knowledge from one year to the next. We do, however, report results based on adjusted gains as discussed below.

$$AdjGainScore = \frac{Score_{t,gs} - E[Score_{t,gs} | Score_{t,g-1,t-1}]}{\sqrt{E[Score_{t,gs}^2 | Score_{t,g-1,t-1}] - E[Score_{t,gs} | Score_{t,g-1,t-1}]^2}}$$

(3)

The use of the adjusted gain score as the dependent variable means that the lagged test score is no longer required as an explanatory variable and any problems related to mean reversion are kept to a minimum.¹⁶ In addition, this alternative dependent variable accounts for the possibility that one-year differences may signify larger or smaller gains at different points in the prior year achievement distribution. Because this specification is in a gains metric, we do not include a lagged test score. As with the specifications discussed above, we include student, school, and year fixed effects.

Complete model

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

$$(4) Stdmath_{ijt} = \alpha + \beta_1 FailAYP_{ijt-1} + \beta_2 FailGrowth_{ijt-1} + \beta_3 FailBoth_{ijt-1} + \beta_4 LowRead_{ijt-1} + \beta_5 HighRead_{ijt-1} + \beta_6 FailAYP * LowRead_{ijt-1} + \beta_7$$

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

¹⁶ In an earlier version of this paper, we addressed the potential problem of mean reversion by substituting the prior-year test score in the other subject for the actual prior year test scores (that is, math for reading, and reading for math). For statistical validity, that approach requires, however, that the measurement error not be correlated across math and reading scores, which was not fully the case in our data, and from a substantive perspective it requires the debatable assumption that schools make no tradeoffs between math and reading. We note, however, that the distributional effects of that approach were quite similar to the results from the basic model.

prior year. The levels version differs only by the exclusion of the prior year test score. The adjusted gain models are also similar to the basic model except as noted earlier.

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 δ_i , η_j and γ_t , represent student, school and year fixed effects, respectively. The final term is the error term.

Descriptive information for all the variables is reported in the appendix. As shown there, 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 the observation that North Carolina's proficiency level is set at a relatively low level.¹⁷ 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 failed both standards.¹⁸ 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).

¹⁷ 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.

¹⁸ 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.

Justification for this estimation strategy

Implicit in this estimation strategy are the assumptions that schools not making an particular accountability standard in a given year have an incentive to alter their behavior in the following year in an effort to do better and that schools that do make the standard have no such incentives. We believe these are reasonable assumptions, especially for the state's growth program. Under that program, there is an immediate and clear adverse impact on schools that fail to make their expected growth, namely their teachers do not receive a bonus. Moreover, the program is specifically designed to make it possible for any school to meet its growth standard in any year, and the evidence suggests that many schools that fail the standard one year succeed in making it the following year. In particular, over the period of this study, 46 percent of schools that failed the growth standard in year t-1 made it the next year with the percentage ranging from 10 to 70 percent, depending on the year. In contrast, 79% of schools that met the growth standard in year t-1 made it in the following year with the range from 67 to 97 percent.

The assumptions could be somewhat less appropriate for the status model because the immediate adverse consequences for schools of not meeting AYP in year t-1 are less clear. Moreover, schools that are far below the standard one year may find it particularly difficult to make AYP the following year and therefore may have little incentive to change their behavior. Consistent with that possibility, between 2004 and 2007, 64 percent of schools that failed AYP in year t-1 also failed AYP in the following year. This percentage ranged from 43% in the 2004 to 75% in 2006. In fact, between 2005 and 2007, this percentage remained above 70%, indicating a high degree of persistence of NCLB sanctions among North Carolina schools.

Another approach to measuring the incentive effects of the status program would be to focus on the margin is to take into account the difficulty of meeting the standard (Reback 2008; Springer 2008). One can view the status-related findings in this paper as the distributional effects of accountability pressure averaged over all the schools that failed the standard in year t-1, regardless of how likely they viewed themselves to make the standard the following year.¹⁹

Basic Results

Table 1 reports both the main effects and the distributional effects of accountability pressure for the value added model (columns 1 and 3) and the levels model (columns 2 and 4). The standardization of the dependent variables by grade level and year means that 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 first row in table 1 indicates that students in the reference category – namely those just above grade level in the previous year – in schools facing pressure from failing the growth standard alone exhibit lower achievement gains in the following year in both math and reading than comparable students in schools not facing such pressure, with the negative coefficient 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

¹⁹ Both the shorter time frame and the less change in success from one year to the next under the status approach, however, does make it more difficult for us to obtain good estimates of the distributional effects of the status approach relative to the growth approach.

correct. The absence of a counterfactual makes it impossible to make any causal statement about the effect of the accountability system from this type of analysis. The most we can do is to identify relationships between one type of school and another. Hence, the negative coefficient of the failed growth variable simply indicates that students in the reference category in schools that meet their growth targets — that is, schools that generate bonuses for teachers — attain higher test score gains in the subsequent year than do comparable students in schools failing to meet their growth targets. Both sets of scores could be higher, lower or the same as they would have been in the absence of the accountability system. Similar caveats apply to the interpretations of the coefficients of NOTAYP and NOTBOTH in the following two lines.

The next set of variables — the lagged achievement indicators — yield the distributional patterns for students in schools at different points of the prior year achievement distribution in schools facing no negative accountability pressure. Note, however, that the interpretations of these coefficients differ between the two models. The coefficients for the value-added model 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 (after controlling for student fixed effects) and higher achieving students tend to have higher test score gains relative to students closer to grade level, especially in reading. In terms of levels of achievement (once again controlling for time invariant student fixed effects), however, the patterns are reversed. This difference in patterns is not consequential for this analysis given our main interest is not in these variables but rather in their interactions with accountability pressure.

As shown in the first two columns of the next three panels of the table and in figures 4-6, the differential distributional effects in schools facing accountability pressure are similar across the value added and the levels models. All the estimates represent differences in outcomes relative to what would have happened in the absence of the negative pressure of the specific accountability system. We compare results of three mutually exclusive types of pressure – failing only the state’s growth standard, failing the federal status standard, and failing both the growth and status standards – to the baseline case of meeting both the growth and status standards.

In schools failing only the growth standard, for example, students with low prior achievement tend to gain more in math than students just above grade level (see figure 4). Students with the lowest levels of prior achievement (below 1.5 SD below grade level) exhibit the largest gains of about 0.06 standard deviations, with the point estimates declining monotonically as the prior-year achievement category approaches the grade level cut point. Among high achievers in math, students in the top category appear to gain somewhat (about 0.04 standard deviations) relative to students just above grade level. This pattern suggests that in response to the pressure arising from failure to meet the growth standard, schools apparently find it easier to raise math test scores at the bottom and at the very top relative to students close to grade level, and hence provides no support for educational triage.

The responses to growth based accountability pressure differ somewhat for reading, but mainly at the low end of the distribution. Again we find no evidence of educational triage. For this subject, however, no statistically significant effects, either positive or negative, emerge for students below grade level. As was true for math,

positive distributional effects emerge at the top of the distribution, with magnitudes of 0.025 to 0.034 depending on the model.²⁰

With respect to status pressure, the table shows that schools failing only AYP generate positive gains for low-performing students in both subjects, with the coefficients far larger in math than in reading. Of interest is that negative effects emerge for high performing students, with the negative effects larger in reading than in math (see figure 5). 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 negative distributional effects at the high end of the distribution indicates that the gains at the bottom in both subjects have come at the expense of higher achieving students in the affected schools.

Finally, in schools failing to meet both the state's growth standard and the federal AYP standard positive and statistically significant coefficients emerge for low achieving students in reading but not in math. Consistent with the findings for status pressure, negative distributional effects emerge for high achieving students in both math and reading, but with the magnitudes far larger for reading (see figure 6).

In summary, two main conclusions emerge from this basic analysis. The first is that, with a few exceptions, students below grade level typically benefit relative to students close to grade level in schools responding to some form of negative pressure. The main exceptions are low achieving students in reading in schools failing to meet the

²⁰ 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.

growth standard alone and low achieving students in math in schools facing pressure from both programs. The second is that negative distributional effects emerge at the high end of the prior year achievement distribution but only in schools facing AYP pressure alone or both AYP and growth pressure. Stated differently, status pressure appears to generate more within-school shifting of resources away from higher achievers to low achievers than is the case for accountability based on a growth model. In response to pressure from the growth approach, schools appear to focus additional attention both on students who are below grade level and on those at the top of the achievement distribution.

Supplemental Results: The Adjusted Gain Specification

Results for the alternative specification with adjusted gains as the dependent variable are reported in Table 2. This specification adjusts for mean reversion accounting for the possibility of differential gains at each point in the achievement distributions across grade levels and years and includes student, school, and year fixed effects. The results using this specification largely mirror those from our basic specification.

Beginning with the distributional effects associated with not making the growth standard, we find patterns that are quite similar to those from the basic model presented above. Specifically, in response to pressures to make the growth standard, the coefficients indicate first that students at the low end of the achievement distribution benefit relative to students just above grade level, with the positive distributional effects measured as adjusted gain scores being larger for math than for reading. Second, again consistent with our prior finding, positive achievement gains also emerge for students at the very top of the prior-year performance distribution. The only potential anomalies are the statistically

significant negative coefficients for high achieving students closer to grade level in math. Thus, these adjusted gains findings reinforce our earlier conclusion that growth-related accountability pressure leads to positive distributional effects at both ends of the achievement distribution, and, importantly, generate no evidence of educational triage. Moreover, a comparison of the magnitudes of the coefficients across the two tables provides no reason to believe that the basic results were biased upward at the very bottom of the performance distribution or downward at the top of the distribution because of regression to the mean.

With respect to the effect of status-based accountability pressure on math test scores, we again find large positive distributional effects at the bottom of the distribution, but now mixed distributional effects at the top of the distribution, with one significant anomalous positive coefficient (1.0 to 1.5 SD above) and one significant negative coefficient (just above grade level). In reading, the distributional effects are weaker, but are positive for the bottom of the distribution and negative for the top of the distribution. As with our basic specification, we find a negative effect of status-based accountability pressure on reading test scores for students in the upmost tail.

For schools simultaneously facing pressure from both approaches, the patterns in math differ somewhat from those in the basic specification. In reading, however, the patterns are similar.

Discussion and 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 data set and value-added models of student achievement with both student and school fixed effects, 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, students below the proficiency standard typically benefit relative to those just above the standard, although that pattern is clearer and more consistent across models for math than for reading. Nonetheless, in the case of the growth approach, the overall distributional effect within the affected schools is to raise student achievement in the aggregate somewhat 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.²¹ This pattern could represent a shortcoming of the policy if the main goal of the program were to close achievement gaps within such schools. 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 is that in reading, and consistent with Reback (2008), status based accountability pressure appears to generate negative within-school distribution effects for students above the proficiency threshold. 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.

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

²¹ During the period of this study, an average of 79% of students achieved above grade level in reading and an average of 80% of student achieved above grade level in math.

schools facing accountability pressure experience positive (and in some cases, zero) achievement gains on average relative to students slightly 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 within-school distributional patterns highlighted in this study are relevant to policy debates about school accountability. One key question is whether the goal of school-based accountability is to make low performing schools better for everyone or to narrow achievement gaps by raising the performance of students at the bottom of the achievement distribution relative to those at the top in the low-performing schools. To the extent that it raises achievement for some students, but lowers it for

others, as appears to be the case in status-based accountability system in reading, there are clear tradeoffs that require additional policy discussion and debate.

References

- Amrien, Audrey L., and David C. Berliner. 2002. "High-Stakes Testing, Uncertainty, and Student Learning." *Education Policy Analysis Archives* 10:Retrieved 12/23/08 from <http://epaa.asu.edu/epaa/v10n18/>.
- Balfanz, R., N. Legters, T. C. West, and L. M. Weber. 2007. "Are NCLB's measures, incentives, and improvement strategies the right ones for the nation's low-performing high schools?" *American Educational Research Journal* 44:559-593.
- Ballou, Dale, and Matthew G. Springer. 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* 42:231-268.
- Burgess, Simon, Carol Propper, Helen Slater, and Deborah Wilson. 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.
- Carnoy, Martin, and Susanna Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educational Evaluation and Policy Analysis* 24:305-331.
- Cohen, David K. 1996. "Standards-Based Reform: Policy, Practice, and Performance." Pp. 99-127 in *Holding Schools Accountable*, edited by Helen Ladd. Washington, DC: Brookings Institution Press.
- Darling-Hammond, L. 2004. "Standards, accountability, and school reform." *Teachers College Record* 106:1047-1085.
- Figlio, D., and H. Ladd. 2008. "School Accountability and Student Achievement." Pp. 166-182 in *Handbook of Research in Education Finance and Policy*, edited by H. Ladd and E. Fiske. New York and London: Routledge.
- Figlio, David. 2005. "Measuring School Performance: Promise and Pitfalls." in *Measuring School Performance and Efficiency: Implications for Practice and Research*, edited by L Stiefel. Larchmont, NY: Eye on Education.
- . 2006. "Testing, Crime and Punishment." *Journal of Public Economics* 90:837-851.
- Hamilton, L., M Berends, and B Stecher. 2005. "Teachers' Responses to Standards-Based Accountability." Santa Monica, CA: RAND Corporation.
- Hanushek, Eric A., and Margaret A. Raymond. 2005. "Does school accountability lead to improved student performance?" *Journal of Policy Analysis and Management* 24:297-327.
- Jacob, Brian A. 2005. "Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics* 89:761.
- Kane, T. J., and D. O. Staiger. 2002. "The promise and pitfalls of using imprecise school accountability measures." *Journal of Economic Perspectives* 16:91-114.
- Krieg, J.M. 2008. "Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act." *Education Finance and Policy* 3:250-281.
- Ladd, H. F. 1999. "The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes." *Economics of Education Review* 18:1-16.

- Ladd, H. F., and R. P. Walsh. 2002. "Implementing value-added measures of school effectiveness: getting the incentives right." *Economics of Education Review* 21:1-17.
- Ladd, Helen F. 1996. *Holding schools accountable : performance-based reform in education*. Washington, D.C.: Brookings Institution.
- Linn, Robert L. 2000. "Assessments and Accountability." *Educational Researcher* 29:4-16.
- Neal, Derek, and Diane Whitmore Schanzenbach. Forthcoming. "Left Behind By Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics*.
- Nichols, Sharon Lynn, and David C. Berliner. 2007. *Collateral damage : how high-stakes testing corrupts America's schools*. Cambridge, Mass.: Harvard Education Press.
- O'Day, Jennifer, and Marshall Smith. 1993. "Systemic Reform and Educational Opportunity." Pp. 250-312 in *Designing Coherent Educational Policy: Improving the System*, edited by Susan Furman. San Francisco: Jossey-Bass.
- Orfield, Gary, and Mindy L. Kornhaber. 2001. *Raising standards or raising barriers? : inequality and high-stakes testing in public education*. New York: Century Foundation Press.
- Peterson, P, and F Hess. 2006. "Keeping an Eye on State Standards." *Education Next* 6:28-29.
- Reback, R. 2008. "Teaching to the rating: School accountability and the distribution of student achievement." *Journal Of Public Economics* 92:1394-1415.
- Springer, M. G. 2008. "The influence of an NCLB accountability plan on the distribution of student test score gains." *Economics of Education Review* 27:556-563.
- Toch, Thomas, and Douglas Harris. 2008. "Salvaging Accountability." *Education Week* 28:36.
- Todd, P. E., and K. I. Wolpin. 2003. "On the specification and estimation of the production function for cognitive achievement." *Economic Journal* 113:F3-F33.
- Valenzuela, Angela. 2005. *Leaving children behind : how "Texas-style" accountability fails Latino youth*. Albany: State University of New York Press.
- Weitz, K., and J. Rosenbaum. 2007. "Inside the Black Box of Accountability: How High Stakes Accountability Alters School Culture and the Classification and Treatment of Students and Teachers." Pp. 97-116 in *No Child Left Behind and the Reduction of the Achievement Gap*, edited by A. Sadovnik et. al. New York and London: Routledge.

Figures

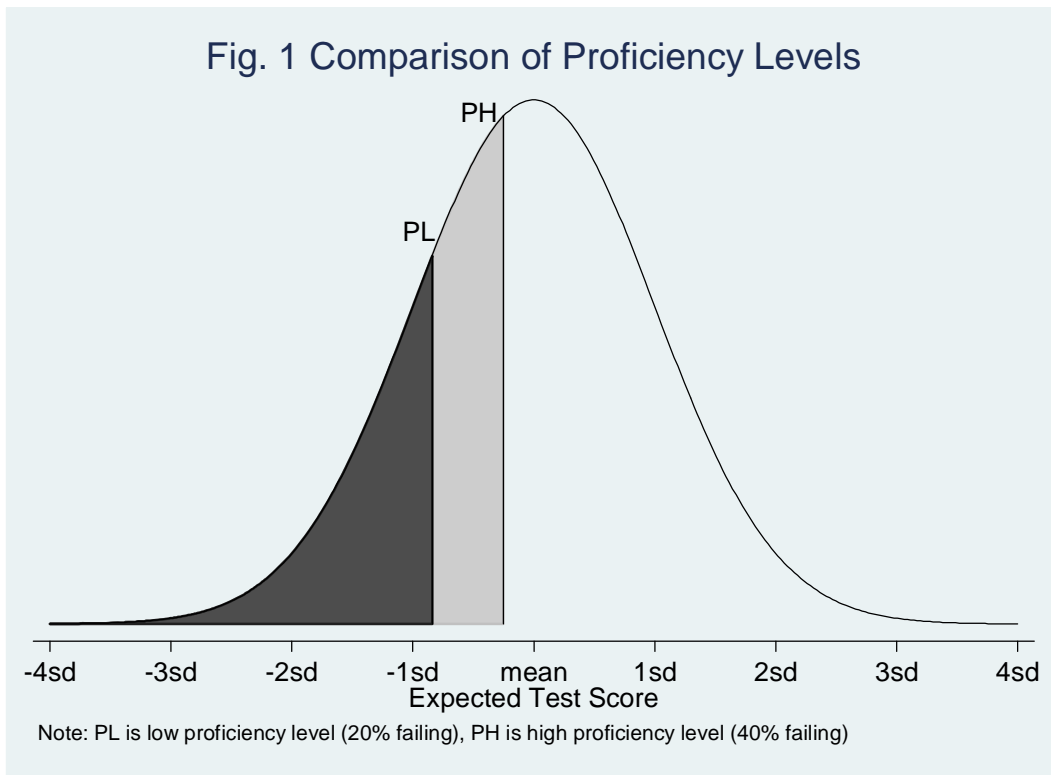


Figure 2.

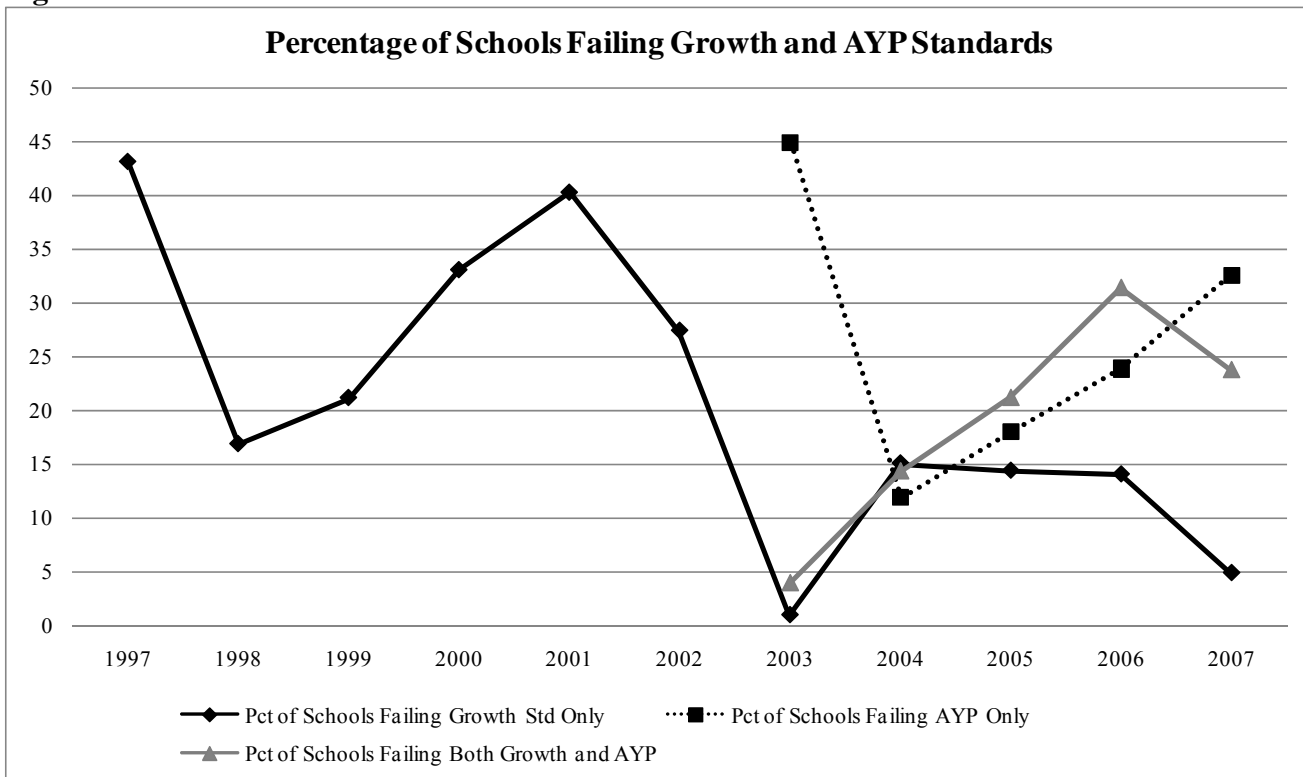


Figure 3.

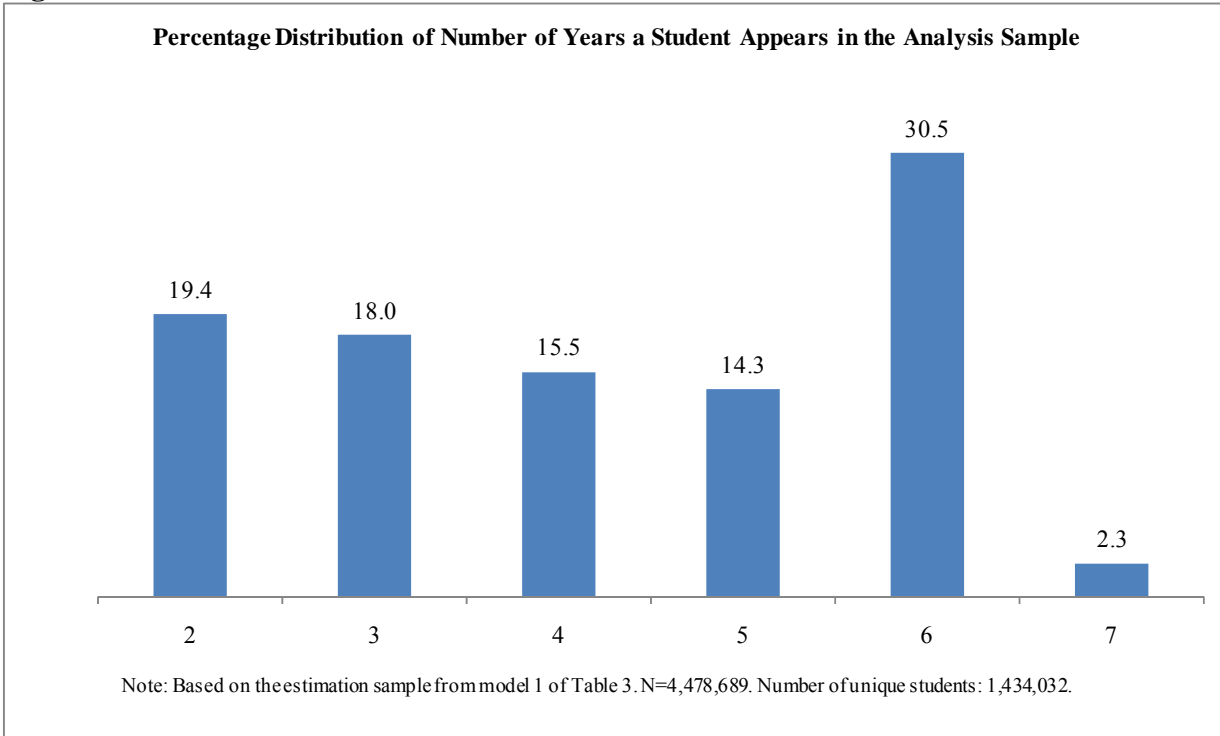


Figure 4.

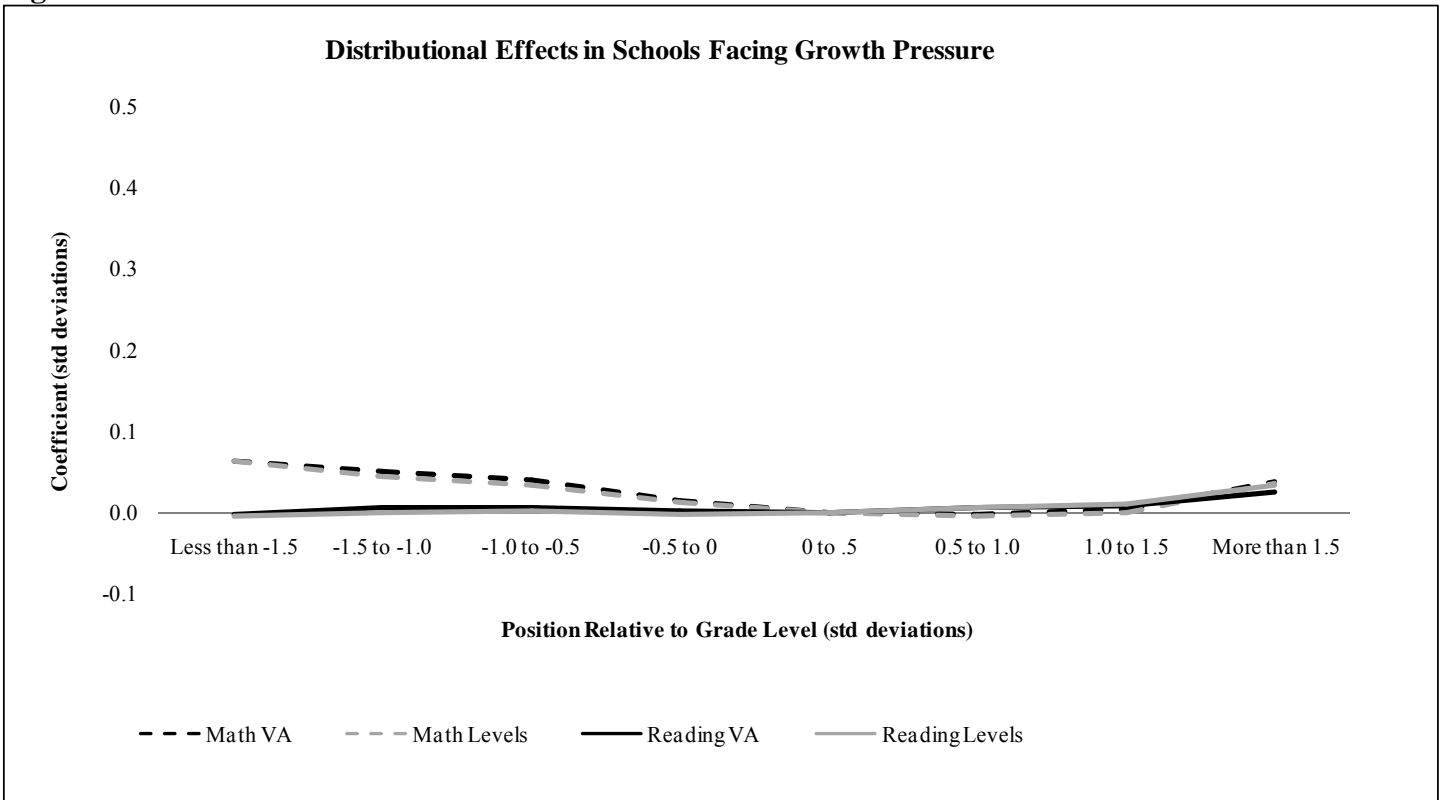


Figure 5.

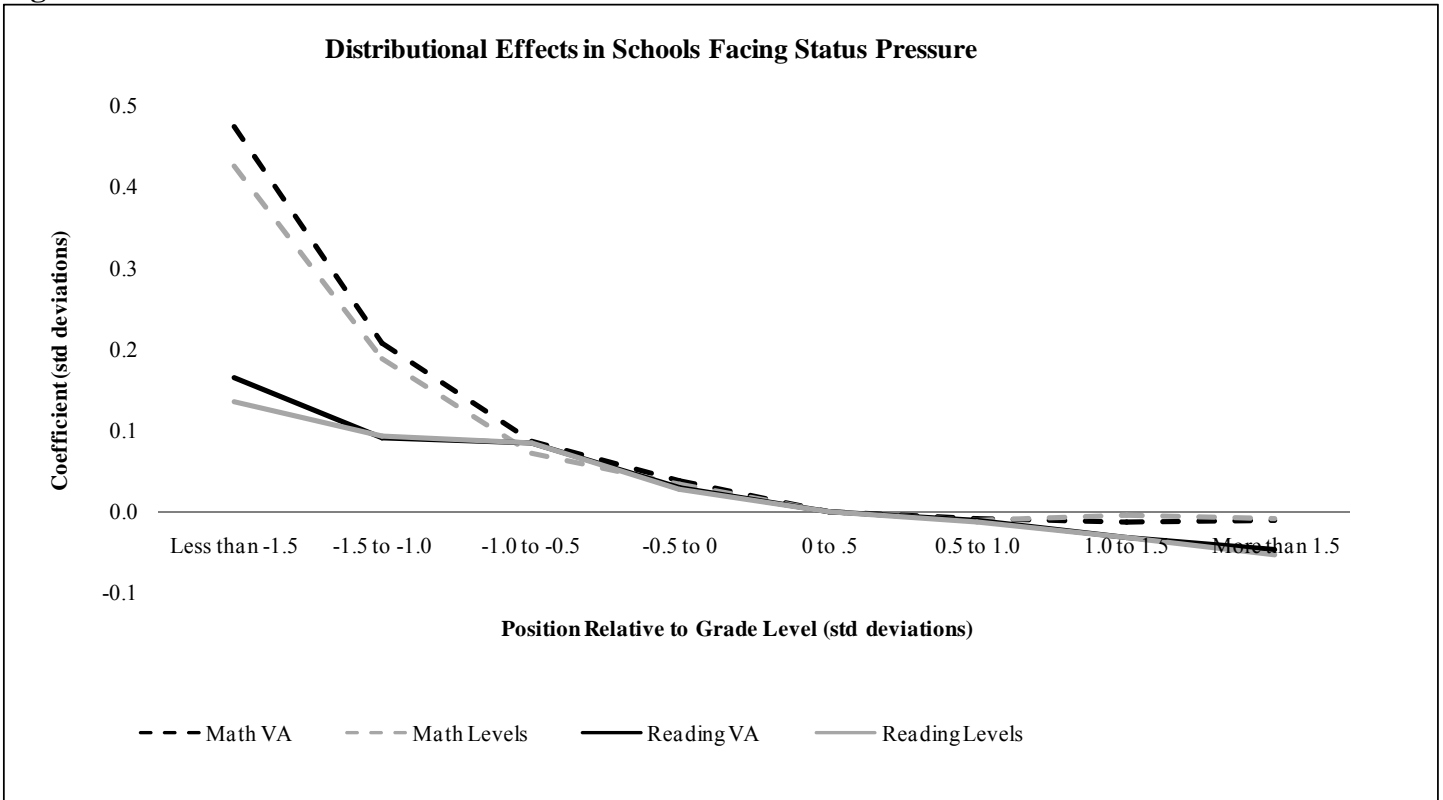


Figure 6.

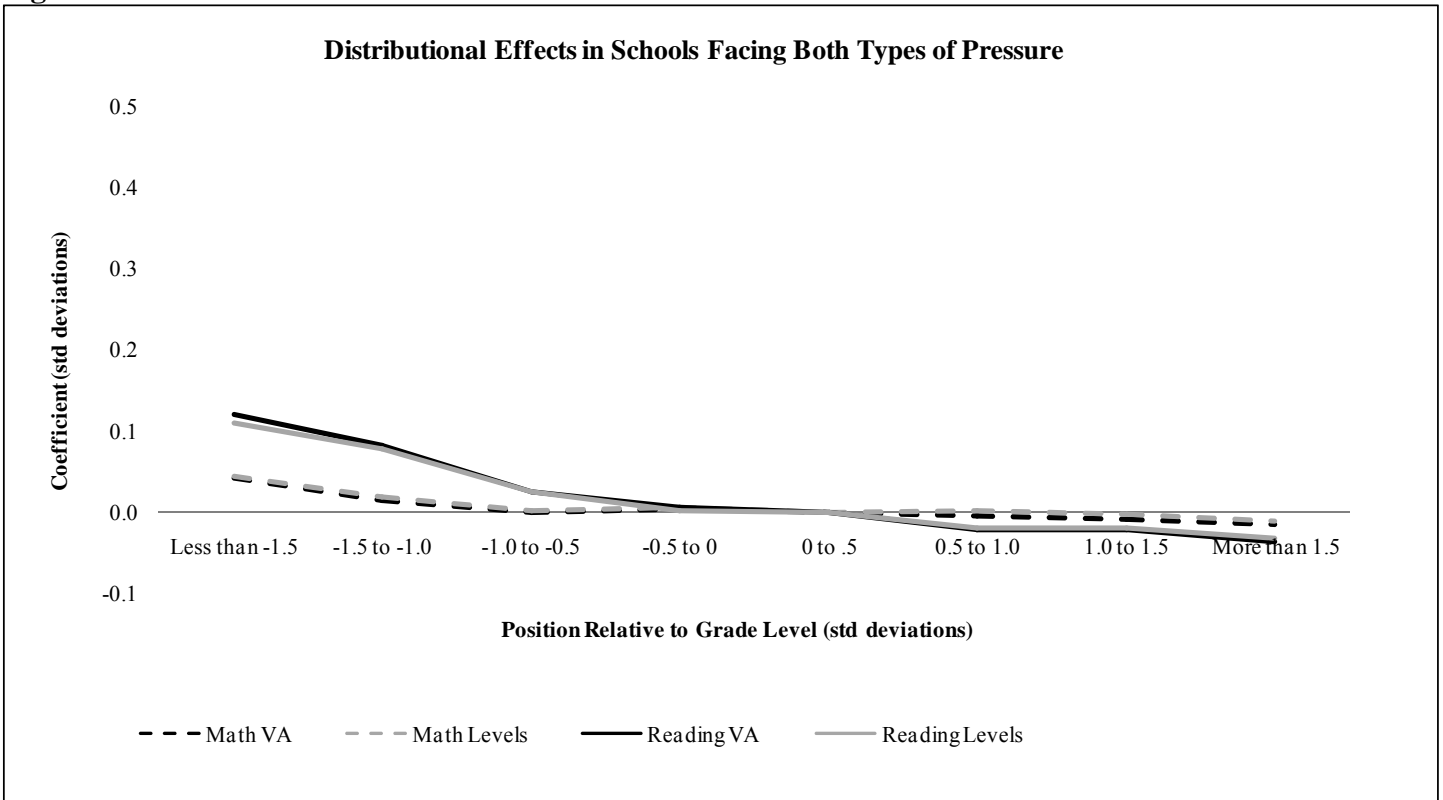


Table 1. Fixed Effects Models Predicting Standardized Test Score Achievement

	(1)	(2)	(3)	(4)
	Math VA	Math Levels	Reading VA	Reading Levels
<i>Accountability Pressure</i>				
Failed Growth	-0.0343*** (0.00169)	-0.0329*** (0.00171)	-0.0193*** (0.00181)	-0.0236*** (0.00182)
Failed AYP	0.00935*** (0.00253)	-0.000491 (0.00255)	0.0220*** (0.00290)	0.0274*** (0.00292)
Failed AYP & Growth	-0.0101** (0.00348)	-0.00884* (0.00351)	0.00123 (0.00413)	0.00456 (0.00416)
<i>Distributional Effects Among Schools Facing No Negative Pressure</i>				
SD Below Grade Level ¹				
Less than - 1.5	-0.239*** (0.00771)	0.199*** (0.00766)	-0.125*** (0.00545)	0.485*** (0.00505)
- 1.5 to -1.0	-0.0850*** (0.00409)	0.229*** (0.00402)	-0.0992*** (0.00382)	0.345*** (0.00350)
-1.0 to -0.5	-0.0322*** (0.00254)	0.180*** (0.00247)	-0.0644*** (0.00277)	0.233*** (0.00258)
-0.5 to 0	-0.00577*** (0.00175)	0.0952*** (0.00174)	-0.0177*** (0.00186)	0.128*** (0.00180)
SD Above Grade Level				
0.5 to 1.0	0.00779*** (0.00134)	-0.0802*** (0.00132)	0.0224*** (0.00145)	-0.112*** (0.00137)
1.0 to 1.5	0.0137*** (0.00151)	-0.159*** (0.00141)	0.0369*** (0.00179)	-0.224*** (0.00153)
More than 1.5	0.00541** (0.00180)	-0.281*** (0.00153)	0.0342*** (0.00227)	-0.389*** (0.00169)
<i>Distributional Effects Among Schools Facing Growth Pressure Only</i>				
Failed Growth*Position				
SD Below Grade Level				
Less than - 1.5	0.0630*** (0.0119)	0.0621*** (0.0121)	-0.00323 (0.00678)	-0.00473 (0.00687)
- 1.5 to -1.0	0.0499*** (0.00586)	0.0439*** (0.00598)	0.00502 (0.00488)	-0.000977 (0.00492)
-1.0 to -0.5	0.0391*** (0.00362)	0.0338*** (0.00368)	0.00689 (0.00378)	0.00191 (0.00380)
-0.5 to 0	0.0136*** (0.00270)	0.0129*** (0.00273)	0.00241 (0.00289)	-0.00269 (0.00291)
SD Above Grade Level				
0.5 to 1.0	-0.00317 (0.00216)	-0.00494* (0.00218)	0.00502* (0.00228)	0.00627** (0.00230)
1.0 to 1.5	0.00393 (0.00219)	-0.00109 (0.00221)	0.00854*** (0.00234)	0.00944*** (0.00236)
More than 1.5	0.0381*** (0.00206)	0.0358*** (0.00209)	0.0253*** (0.00232)	0.0338*** (0.00236)

(continued)

	<i>Distributional Effects Among Schools Facing Status Pressure Only</i>			
Failed AYP*Position				
SD Below Grade Level				
Less than - 1.5	0.474*** (0.0184)	0.425*** (0.0184)	0.165*** (0.0147)	0.135*** (0.0148)
- 1.5 to -1.0	0.208*** (0.00911)	0.188*** (0.00919)	0.0905*** (0.00863)	0.0928*** (0.00871)
-1.0 to -0.5	0.0868*** (0.00551)	0.0716*** (0.00556)	0.0852*** (0.00620)	0.0839*** (0.00627)
-0.5 to 0	0.0382*** (0.00405)	0.0334*** (0.00409)	0.0295*** (0.00462)	0.0279*** (0.00466)
SD Above Grade Level				
0.5 to 1.0	-0.00804** (0.00301)	-0.0105*** (0.00303)	-0.0101** (0.00333)	-0.0141*** (0.00336)
1.0 to 1.5	-0.0130*** (0.00295)	-0.00521 (0.00298)	-0.0312*** (0.00330)	-0.0332*** (0.00333)
More than 1.5	-0.0102*** (0.00272)	-0.00886** (0.00276)	-0.0463*** (0.00314)	-0.0541*** (0.00319)
	<i>Distributional Effects Among Schools Facing Growth and Status Pressure</i>			
Failed AYP & Growth*Position				
SD Below Grade Level				
Less than - 1.5	0.0423 (0.0233)	0.0439 (0.0234)	0.121*** (0.0191)	0.110*** (0.0193)
- 1.5 to -1.0	0.0150 (0.0119)	0.0188 (0.0120)	0.0821*** (0.0121)	0.0787*** (0.0122)
-1.0 to -0.5	-0.000335 (0.00750)	0.00132 (0.00757)	0.0251** (0.00872)	0.0243** (0.00879)
-0.5 to 0	0.00451 (0.00566)	0.00561 (0.00572)	0.00596 (0.00670)	0.00216 (0.00676)
SD Above Grade Level				
0.5 to 1.0	-0.00489 (0.00444)	0.000724 (0.00448)	-0.0224*** (0.00500)	-0.0205*** (0.00504)
1.0 to 1.5	-0.00892* (0.00441)	-0.00173 (0.00445)	-0.0211*** (0.00500)	-0.0185*** (0.00504)
More than 1.5	-0.0162*** (0.00403)	-0.0106** (0.00408)	-0.0371*** (0.00476)	-0.0331*** (0.00482)
Intercept	-0.000643 (0.00459)	0.0175*** (0.00460)	-0.0637*** (0.00507)	-0.0460*** (0.00508)
N	4492524	4492524	4472971	4472971
R ² within	0.087	0.053	0.105	0.075
R ² between	0.739	0.540	0.737	0.668
R ² overall	0.644	0.453	0.621	0.563
¹ Base 0 to 0.5 SD above grade level				
Note: Value added (VA) models control for prior test score. All models control for gifted, special education, limited English proficiency, being new to the school, percent black, percent Hispanic, percent LEP, number of accountable subgroups, and student, school, and year fixed effects. Cluster corrected standard errors in parentheses; * p<0.05, ** p<0.01, *** p<0.001.				

Table 2. Fixed Effects Models Predicting Adjusted Gain in Test Score Achievement

	(1)	(2)	
	Math Student and Sch FE	Reading Student and Sch FE	
	<i>Distributional Effects Among Schools Facing Growth Pressure Only</i>		
Failed Growth*Position			
SD Below Grade Level ¹			
Less than - 1.5	0.0668** (0.0228)	0.0270* (0.0116)	
- 1.5 to -1.0	0.0228 (0.0120)	0.0242** (0.00822)	
-1.0 to -0.5	0.0357*** (0.00771)	0.0151* (0.00663)	
-0.5 to 0	0.0347*** (0.00593)	-0.0119* (0.00544)	
SD Above Grade Level			
0.5 to 1.0	-0.0338*** (0.00503)	0.00919 (0.00472)	
1.0 to 1.5	-0.0449*** (0.00526)	0.0169*** (0.00496)	
More than 1.5	0.0290*** (0.00519)	0.0944*** (0.00515)	
	<i>Distributional Effects Among Schools Facing Status Pressure Only</i>		
Failed AYP*Position			
SD Below Grade Level			
Less than - 1.5	0.772*** (0.0375)	0.00821 (0.0237)	
- 1.5 to -1.0	0.368*** (0.0185)	0.0257 (0.0141)	
-1.0 to -0.5	0.103*** (0.0112)	0.0462*** (0.0104)	
-0.5 to 0	0.0540*** (0.00855)	0.00419 (0.00817)	
SD Above Grade Level			
0.5 to 1.0	-0.0350*** (0.00690)	-0.0116 (0.00647)	
1.0 to 1.5	0.0521*** (0.00702)	-0.0115 (0.00656)	
More than 1.5	-0.00322 (0.00660)	-0.0863*** (0.00642)	

(continued)

	<i>Distributional Effects Among Schools Facing Growth and Levels Pressure</i>			
Failed AYP & Growth*Position				
SD Below Grade Level				
Less than - 1.5	0.179*** (0.0472)	0.113*** (0.0309)		
- 1.5 to -1.0	0.126*** (0.0240)	0.107*** (0.0196)		
-1.0 to -0.5	0.0342* (0.0153)	0.0308* (0.0146)		
-0.5 to 0	0.0106 (0.0119)	-0.00585 (0.0119)		
SD Above Grade Level				
0.5 to 1.0	0.0270** (0.0101)	-0.0109 (0.00973)		
1.0 to 1.5	0.0175 (0.0104)	-0.0149 (0.0100)		
More than 1.5	0.0107 (0.00981)	-0.0364*** (0.00983)		
Intercept	0.295*** (0.0108)	0.198*** (0.0101)		
N	4491489	4472076		
R ² within	0.308	0.411		
R ² between	0.020	0.027		
R ² overall	0.000	0.000		
¹ Base 0 to 0.5 SD above grade level				
Note: The dependent variable is test score gain adjusted for the typical gain and variance for students at each level of test score (see equation 3). Models control for the main effects of accountability pressure and prior test score position, gifted, special education, limited English proficiency, being new to the school, percent black, percent Hispanic, percent LEP, number of accountable subgroups, and student, school, and year fixed effects. Cluster corrected standard errors in parentheses; * p<0.05, ** p<0.01, *** p<0.001.				

Appendix

Table A1. 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
pethisp	% 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; pethisp, 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 A2. The valid number of observations 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.				

Table A2. 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.

