Achievement Trade-Offs and No Child Left Behind

Dale Ballou
Peabody College of Vanderbilt University
dale.ballou@vanderbilt.edu

Matthew G. Springer
Peabody College of Vanderbilt University
matthew.g.springer@vanderbilt.edu

Working Paper

**Please do not quote or cite without permission*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 thank the Northwest Evaluation Association for providing data for this study as well as an anonymous foundation and the federally-funded National Center on School Choice at Vanderbilt University for research support. They also thank Adam Gamoran, Steve Rivkin, and Kim Rueben for their helpful comments and insights in developing this work as well as seminar participants at the *American Education Finance Association*, *American Educational Research Association*, *Association for Public Policy Analysis and Management*, and Amherst College. Special thanks is due to Art (Xiao) Peng for his research assistance on 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

Under the No Child Left Behind Act, states have been required to set minimum proficiency standards that virtually all students must meet by 2014. Sanctions of increasing severity are to be applied to schools that fail to meet interim targets, known as Adequate Yearly Progress (AYP). The authors examine the effect of this legislation using longitudinal, student-level test score data from seven states (N > 2,000,000) between 2002-03 and 2005-06 school years. This paper addresses the following research questions: (1) Has NCLB increased achievement among lower-performing students? ; (2) Have these gains come at the expense of students that are already proficient or that are far below the proficiency target? Identification is achieved by exploiting the fact that in the early years of NCLB, not all grades counted for purposes of determining AYP. The estimate of the NCLB effect is therefore based on a comparison of outcomes in high-stakes vs. low-stakes years. The authors find consistent evidence of an achievement trade-off in the hypothesized direction, though the effects on any given student are not large. Unlike some other researchers, they find mixed evidence at best that students far below the proficient level have been harmed by NCLB; indeed, at higher grade levels they appear to have benefitted. Effects of NCLB on efficiency, while positive, appear to be modest.

1. Introduction

The *No Child Left Behind Act of 2001* (NCLB) is the reauthorization of the nation's omnibus *Elementary and Secondary Education Act of 1965* (ESEA). NCLB represents a major effort by the federal government to improve academic performance among groups of students who have traditionally lagged behind. States have been required to set minimum proficiency standards in reading and mathematics. Sanctions of increasing severity are to be applied to schools that fail to demonstrate Adequate Yearly Progress (AYP), determined by the percentage of students achieving the state-defined performance standard. Over time the percentage of students required to meet this standard is ratcheted upwards, until virtually all students must score proficient or better in 2014.

NCLB targets apply to all of a school's students as a group, as well as to subgroups within the school as long as subgroups meet minimum count requirements. A school fails to make AYP if any of the recognized subgroups within that school fails. The main subgroups are defined on the basis of race/ethnicity, income (eligibility for the free- and reduced-price lunch program), disability (special education students), and English proficiency (English language learners).

NCLB has been criticized for failing to enhance capacity at low-performing schools and for focusing narrowly on a single performance threshold rather than on gains across the spectrum of achievement. In order to bring performance of all students up to the prescribed minimum, it is feared that schools will divert a disproportionate amount of resources to those students who are particularly important to a school's accountability rating. In the short-term this would consist primarily of the group of students near the proficiency threshold but not assured of passing it.[1] In the long-term this will include an ever-larger share of those students below the standard. In schools' effort to raise achievement in this group, traditionally high-performing students may be neglected. In the short-run, students who are far below the performance threshold may also be neglected.

[1] "Near" the proficiency threshold is a relative term, depending on the distribution of ability within the school. We make this concept precise in our definition (below) of a school's marginal student.

It has been argued that achievement trade-offs are an inevitable consequence of the design of NCLB, suggesting that empirical confirmation is not even required. However, the inevitability of trade-offs follows only if schools are operating efficiently, on the production frontier. This should not be taken for granted. In the absence of clear accountability public schools, like other organizations, are apt to perform below their operational capacity. A long-standing debate over "whether money matters" in public education suggests that at a minimum, public schools frequently fail to make efficient use of resources. There may be sufficient slack in the present educational system that raising the achievement of marginally-performing students will not require trade-offs in the form of lower achievement for others, at least in the near term. We ask, therefore, two questions:

- Has NCLB increased achievement among lower-performing students?
- Have achievement gains come at the expense of students that are already proficient or that are far below the proficiency target?

2. Identification Strategies

We are not the first researchers to study the distributional effects of NCLB in public schools, or ask similar questions about accountability systems more generally. However, by their nature accountability systems are typically implemented wholesale, applying to virtually all public schools across the board. Apart from a handful of alternative schools for exceptional needs students, or schools with very few students, there are no schools outside the accountability system. As a result, there is no natural comparison group for estimating the impact of an accountability system on educational outcomes. Researchers have resorted to a variety of identification strategies to make good this deficit.

A. Pre- and post-accountability system comparisons

One strategy relies on pre- and post-accountability comparisons. Neal and Schanzenbach (forthcoming) compared mathematics and reading test scores of Chicago Public School students before and after the implementation of a high-stakes accountability system.[2] They found significant increases in mathematics and reading test scores among those around the accountability system's proficiency threshold, while traditionally low-performing students did not demonstrate increased performance. Effects on the achievement of traditionally high-performing students were mixed.

Pre- and post-accountability comparisons suffer from drawbacks common to interrupted time series designs. Effects of an accountability system can be confounded with other changes occurring at the time the system is implemented. In addition, in many states, the data needed to evaluate the accountability system often does not pre-date the system, as testing on a statewide basis with public disclosure of the results is frequently introduced as part of the accountability program. As a result, either there are no pre-NCLB test data, or the effects of NCLB must be distinguished from those of a state accountability system launched at the same time as the testing regime.

The effect of accountability on student achievement also may be lagged several years, as it takes time for teachers and schools to ascertain how the system affects them. Time is needed for schools and school systems to develop instructional policies to respond to the system, and even more time before their responses have an appreciable impact (if any) on student achievement. Still more time is needed for data to become available to researchers for evaluation purposes. A lagged response is more likely if an accountability system is phased in or if targets are ratcheted up over time, as with NCLB. Thus, early findings that an accountability system does not seem to be working must be taken with a grain of salt—it may be too soon to tell.

B. Exploiting variation in the strength of incentives

NCLB creates incentives that are stronger for some types of schools than others and that affect some students differently than others. In most states, failure to make AYP triggers sanctions only for schools receiving Title I funds. These incentives are weakened to the extent that a failing

[2] A similar approach was implemented in Krieg's (2008) study on the distributional effect of NCLB in Washington.

school itself does not bear the full costs of these sanctions. However, one would expect pressure to be exerted on the schools that are responsible, even when the costs fall on the district. In addition, one would expect that as sanctions become more severe, schools will make greater effort to raise student achievement.

Variation in the level of sanctions is endogenous if there is any serial correlation in the unobserved determinants of achievement, so that additional identification strategies are required to deal with the fact that a school's accountability rating depends on the performance of students in that school or the quality of their teachers. Attempts to remove serial correlation, say by the inclusion of school fixed effects in the model, tend to exacerbate measurement error. With fixed effects in the model, the impact of NCLB is identified from variation in achievement (relative to the school’s mean) that is correlated with variation in sanctions (also relative to the school’s mean). A school that faces sanctions as a consequence of an off year is apt to recover the next year without doing anything differently. This recovery leads to an upward bias on the estimated treatment effect and may be mistakenly interpreted as a positive response to the accountability system.

Several researchers have relied on regression discontinuity techniques to get around the endogeneity problem. Rosaen, Schwartz, and Forbes (2007) detected no impact of NCLB on mathematics achievement in California and only a slight, positive effect in reading. Chakrabarti (2007) and Rouse et al. (2007) reported that public schools graded "F" under Florida's A+ accountability system responded to voucher threats differently from those schools graded "D." “F” schools significantly increased student achievement. The improvement did not come at the expense of high-performing peers.

As a strategy for studying NCLB, regression discontinuity has one notable drawback. Under NCLB, schools that barely made AYP know they will be tested again in the future and judged against a standard that is rising. For such schools to behave significantly differently from schools that barely failed AYP requires a high degree of myopia on the part of the former. As a result, it is likely that regression discontinuity identifies only part of the NCLB effect.

Another identification strategy takes advantage of the quirks in an accountability system that may create the conditions of a natural experiment. Under NCLB schools are accountable for the performance of various subgroups of students only if the number of students in a particular subgroup exceeds a threshold value (minimum N) defined by the state. Fewer subgroups means a lower probability overall of a school failing to make AYP. Thus, the number of subgroups is a credibly exogenous source of variation in the likelihood that a school will face sanctions under NCLB, making it a suitable instrument for sanctions. Sims (2007) takes advantage of this nuance, finding the accountability system California implemented prior to NCLB had no discernible effect on student achievement. He finds counterproductive effects of NCLB sanctions.

A final set of studies capitalizes on the fact that accountability systems, particularly those that resemble NCLB in its emphasis on minimum competency standards, also creates incentives for schools to target instructional resources on the students who count the most.[3] Improvements among students near the proficiency cutscore might therefore be more important to schools than improvements among students who are either well above the cutscore or so far below it that there is no reasonable chance they can make the goal within the current year (or who are not needed, if the school can make AYP without them). Because students near the cutscore might be targeted even in the absence of an accountability system (for example, the cutscore is near the median student), this approach is enhanced by taking into account not only a student's distance from the cutscore but the importance of that student to the school's effort to make AYP (Holmes, 2003; Reback, 2008).

[3] Evidence from case studies in Texas (e.g., Booher-Jennings, 2005) and Chicago (e.g., White and Rosenbaum, 2007) responded to NCLB by expending a disproportionate amount of effort on marginally performing students as a means to avoid sanctions, often to the detriment of the lowest performing students. Diamond and Spillane (2004) report similar tradeoffs in a case study of a pre-NCLB accountability system in Chicago.

The handful of studies pursuing this identification strategy report mixed findings. Evidence reported on the pre-NCLB accountability system in Texas, for example, suggests that test scores improved most among students at or below the passing threshold, while relatively high-performing students performed worse than expected (Deere and Strayer, 2003; Reback, 2008). However, using data from an unidentified western state, Springer (2007) does not detect evidence of such trade-offs. Low-achieving students in schools that failed to make AYP performed better than similar students elsewhere without diminishing test score gains of high-performing students.

C. Identification in this study

Our identification strategy in this study combines elements from several of the studies discussed above. In the early years of NCLB, not all grades counted when determining if a school made AYP. This gave states time to comply with required annual testing in grades three through eight, a requirement that had not been a feature of earlier reauthorizations of the *Elementary and Secondary Education Act*. Until 2005-06 school year, for example, states were required to test in reading/language arts and mathematics once at the elementary level and once at the middle/junior high school level. As NCLB took effect in 2002-03 school year, this resulted in a three-year period during which some grades were high-stakes grades (test results counted toward AYP) while other grades were low-stakes grades (test results did not count toward AYP). States differed with respect to the grades designated as high-stakes as well as when a grade switched from being low-stakes to high-stakes.

We therefore identify an NCLB effect by comparing outcomes across low- and high-stakes years within a grade. In this regard our strategy resembles a pre- and post-NCLB comparison, although there may be some contamination if a school's response to NCLB in a high-stakes grade-year combination affected instructional practices and outcomes in low-stakes grade-year combinations. Thus, similar to earlier studies, we are not estimating the full NCLB effect, but rather the difference, if any, between outcomes of students who counted for AYP and students who did not.

3. Data

Our identification strategy requires test results for both low- and high-stakes years within a grade. This is generally problematic, given that our "low-stakes" grades are typically those for which a state-approved test had not yet been developed, a common occurrence during the initial years of NCLB. We make good this deficit by turning to test data from the Northwest Evaluation Association's Growth Research Database. NWEA has contracted with over 3,400 school districts in 45 states to conduct testing primarily for diagnostic and formative purposes. NWEA has developed tests in reading, mathematics, language arts, and, more recently, science. Exams at different grade levels are placed on a single scale to measure student development over time and are constructed to avoid ceiling effects.

Most schools contracting with NWEA test at least twice a year, in the fall and spring, though not all districts contracting with NWEA test all of their students. This study uses data from seven states—Arizona, Colorado, Idaho, Indiana, Michigan, Minnesota, and Wisconsin—where testing rates are comparatively high. We further restrict the sample to schools that tested at least 80 percent of their students, which results in an average test participation rate that exceeds 90 percent for our sample.[4] We also restrict our sample to public school students tested in both fall and spring in the same school, given that students who switch schools mid-year do not count when determining a school's AYP status.

There are several advantages to using NWEA tests as a measure of educational outcomes. First, the availability of fall-to-spring gain scores allows us to avoid problems that arise when only a single score is available for a given year. In many administrative data sets, newcomers to a school will lack prior test scores and must be dropped from the sample. Because the test score in one spring serves as the starting value for calculating next year's gain, gain scores based on annual results

[4] Enrollments were obtained from the National Center on Education Statistics' Common Core of Data.

exhibit negative serial correlation, complicating the assessment of the effects of policy changes. Spring-to-spring gain scores are also confounded by the influence of summer months. With fall and spring testing, we avoid these problems.

Second, as NWEA tests are not used for accountability system purposes, results should be unaffected by attempts by teachers and administrators to game the system by narrowly teaching to the test, coaching students during testing, or altering student answers after students complete the assessment.[5,6] In addition, NWEA uses a state-aligned computer-adaptive testing system in which questions are drawn from a single, large item bank. There is no single test form used in a given year and no concern about effects of changing from one test form to another.

Third, because schools are interested in using the results of NWEA tests to identify students who need to make extra progress in order to achieve proficiency, NWEA has conducted a series of technical studies to create crosswalks between scores on its tests in mathematics and reading and scores on each state's high stakes assessments. These technical studies are posted on the company's web site and information is disseminated to school districts to aid schools in the interpretation of NWEA test results. Furthermore, NWEA provides reports to classroom teachers and schools within three days of completing testing so teachers and principals know which students in their classes and school are on track to meet proficiency standards and which students may require remediation.

NWEA has conducted a technical study of this kind for each of the seven states represented in our sample (see Appendix A: NWEA Score Alignment Studies). While it should not be supposed that the NWEA tests and state high stakes tests are perfectly equated, the interest in using NWEA test results to guide instructional decisions and the effort the company has made to assist schools by providing these technical studies suggests that schools will regard the gap between a student's fall

[5] See, for example, Grissmer and Flanagan (1998), Koretz (2002), Jacob (2005), and Jacob and Levitt (2007).

[6] Idaho is an exception. Through the 2005-06 school year, the state used NWEA exams for its accountability system.

score and the cut-score equivalent provided by NWEA as an indication of the progress the student needs to make in order to reach proficiency. We exploit this information to construct one of the key variables in our model, as explained below.

There are drawbacks to using NWEA data, namely, the mix of schools represented in the data has changed as districts signed new contracts with NWEA or allowed old contracts to elapse. Table 1 displays the number of student-level observations in our sample, by state, grade, and year. High-stakes grade-year combinations are shown in boldface print. The total number of student observations has increased over time, with the largest increases occurring in Michigan, Minnesota, and Wisconsin. Even in other states, however, where totals are comparatively stable, there has been modest turnover in the districts represented.

Although we include both year and state effects in our models to help control for changes in the composition of our sample, these controls will not capture within-state changes over time. We explore the robustness of our findings to these compositional changes by repeating all analyses using a restricted sample that comprises an unchanging (or only slightly changing) set of schools. The restricted sample is made up of schools that were in the data set in all four years, except in Michigan, Minnesota, and Wisconsin, where we require that a school be present in at least three of the four years. (Imposing a four-year requirement on these three states results in the loss of virtually all observations given the small sample size in the 2002-03 school year.) As reported below, findings are largely robust to this restriction. We have also repeated all analyses using a dataset from which schools are excluded unless the school is present in all four years under study. Results differ little from those obtained with our "restricted" sample and are available from the authors upon request.

4. Estimation Strategy

We construct a series of models with a view to testing the four principle hypotheses identified in the literature on the distributional effects of NCLB. We first describe our dependent variable and then discuss each of the four modeling strategies. This section concludes with a description of our approach for estimating the probability a student scores at the proficient level on the next administration of a state's accountability assessment. For ease of reference, we also summarize these models and our hypotheses regarding the direction of effects in Table 2.

Insert Table 2 Here

A. Annualized Gain Score

In all models the dependent variable is a fall-to-spring gain score as measured by NWEA's mathematics test. Because testing dates vary by state and year, we normalize the fall-to-spring gain score by dividing a student's test score gain by the total number of days between fall and spring administration of the NWEA test. We then annualize the score by multiplying the normalized gain score by a standard number of days in a school year (180 days).

B. NCLB main effects

The first of our models focuses on NCLB main effects. In addition to year and state effects, this model contains three explanatory variables: the predicted probability that student *i* in grade *g*, state *s*, and year *t* achieves proficiency when next tested, $\hat{\pi}_{igst}$; a dummy variable indicating whether year *t* was a high-stakes year for grade *g* in state *s* (hs_{gst}); and an interaction of these two variables ($\hat{\pi}_{igst}$ x hs_{gst}).[7] $\hat{\pi}_{igst}$ is a function of the distance between a student's fall score and the proficiency cutscore for his grade and state. (Details on the calculation of $\hat{\pi}_{igst}$ appear at the end of this section.)

[7] The designation of grades as high-stakes (counting for purposes of AYP) was obtained from each state's accountability workbook filed with the U.S. Department of Education. For more information, see http://www.ed.gov/admins/lead/account/stateplans03/index.html.

We use $\hat{\pi}_{igst}$ rather than a student's actual fall score for two reasons. First, $\hat{\pi}_{igst}$ is more likely to correspond to the variable schools and teachers care about than the fall score or the distance between a student's fall score and the state-defined cutscore (cf. Springer, 2008). $\hat{\pi}_{igst}$ is a non-linear transformation of the gap between the cutscore and the fall score. As $\hat{\pi}_{igst}$ approaches its asymptotes, corresponding on the one hand to students who are almost certain to reach proficiency, and on the other to students who have virtually no chance of passing, changes in the difference between the cutscore and fall score that have little effect on $\hat{\pi}_{igst}$ are not likely to be perceived as meaningful by teachers deciding where to focus their efforts. $\hat{\pi}_{igst}$ captures this, while a linear function of the fall score minus the cutscore does not.

Second, $\hat{\pi}_{igst}$ provides us with a metric that facilitates the comparison of each student with a school's marginal student, where the marginal student is defined as the last student from the top who needs to score proficient for their school to make AYP. Because the marginal student might be from a different grade than student *i*, their fall scores may not be directly comparable. This is not a concern when comparing their probabilities of reaching proficiency.

The coefficient on $\hat{\pi}_{igst}$ is expected to be negative as a consequence of regression to the mean. Indeed, the raw data show that students with low fall scores, and therefore low values of $\hat{\pi}_{igst}$, make greater gains on average at every grade level. Compensatory instructional strategies may also be a factor. As a result, the coefficient on $\hat{\pi}_{igst}$ says nothing about NCLB per se. The effects of NCLB are evident in the coefficients on the other two variables (hs_{gst} and $\hat{\pi}_{igst}\ x\ hs_{gst}$). If NCLB enhances a school's operational efficiency, the coefficient on hs_{gst} will be positive. The distributional effect of NCLB is identified through the coefficient on $\hat{\pi}_{igst}\ x\ hs_{gst}$. If NCLB has led schools to shift resources and attention from high-performing students to low-performing students, this coefficient will be negative.

C. Urgency of improvement hypothesis

The second of our models focuses on the fact that not all schools have been under equal pressure to improve student performance. One would expect a greater response to NCLB among schools at greater risk of failing to make AYP. To measure the latter, we calculate the number of students who must reach proficiency for the school to make AYP. We use the distribution of fall test scores to identify the marginal student—the student ranked Mth in a school where M students must score proficient or above.[8] $\hat{\pi}_{Mst}$ denotes the probability that the marginal student passes the spring high-stakes assessment.

Our second model therefore includes all the variables in the first model, plus the following additional explanatory variables: $\hat{\pi}_{Mst}$, $hs_{gst}\, x\, (1\text{-}\hat{\pi}_{Mst})$, and $\hat{\pi}_{Mst}\, x\, hs_{gst}\, x\, (1\text{-}\hat{\pi}_{Mst})$. $(1\text{-}\hat{\pi}_{Mst})$ is a measure of the urgency with which a school needs to improve. If $\hat{\pi}_{Mst}$ is close to 1, the school is relatively assured of making AYP without altering instructional practices. The closer is $\hat{\pi}_{Mst}$ to zero, the greater are the changes required for that school to make AYP. The interaction of $(1\text{-}\hat{\pi}_{Mst})$ with hs_{gst} therefore captures the strength of a school's incentive in to respond to NCLB, during years when the high-stakes exams are given. To the extent this improves overall efficiency, the expected sign is positive. The three-way interaction—$\hat{\pi}_{igst}\, x\, hs_{gst}\, x\, (1\text{-}\hat{\pi}_{Mst})$—captures the impact of this incentive on the distribution of achievement. If schools in which improvement is more urgent focus on lower-performing students at the expense of high-performing students, the sign on $\hat{\pi}_{igst}\, x\, hs_{gst}\, x\, (1\text{-}\hat{\pi}_{Mst})$ will be negative.

[8] The percentage of students that must reach proficiency for a school to make AYP was obtained from state accountability workbooks. In this calculation we ignore the fact that each subgroup within a school must also make AYP and focus solely on the percentage of students overall who must reach proficiency.

D. Bubble effect hypothesis

Our third model focuses on whether schools devote special attention to those students who start the year near the cutscore and whose improvement is most important if the school is to make AYP. This focus on "bubble kids" has been detected by other researchers in both pre-NCLB and NCLB accountability systems. We test the bubble hypothesis in our third model by adding a variable to our second model that represents the absolute difference of student *i*'s probability of scoring proficient and the marginal student's probability, times our urgency of improvement measure $(|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|(1-\hat{\pi}_{Mst}))$. As the distance between $\hat{\pi}_{igst}$ and $\hat{\pi}_{Mst}$ increases, student *i*'s educational needs are more likely to be neglected, with the degree of neglect a function of a school's need to focus on the marginal student. The effect is therefore greatest in schools where $\hat{\pi}_{Mst}$ is low and disappears as $\hat{\pi}_{Mst}$ approaches one.

To ensure that $(|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|(1-\hat{\pi}_{Mst})$ truly captures an NCLB effect (and not the tendency, say, to focus on students near the middle of the distribution), we enter this variable both alone and in an interaction with hs_{gst}. As with other variables, it is the interaction $(|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|(1-\hat{\pi}_{Mst})\ x\ hs_{gst})$ that identifies the NCLB effect. If correct, the bubble hypothesis implies a negative sign on $(|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|(1-\hat{\pi}_{Mst})\ x\ hs_{gst})$ in high-stakes grade-year combinations.

E. Educational triage hypothesis

Our final model addresses the concern that the focus on "bubble kids" constitutes a form of educational triage that is particularly damaging to the lowest achieving students.[9] To investigate this possibility, we relax the assumption found in our third model that the effect of $|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|$ is symmetric. We do so by estimating separate coefficients for differences in the positive and negative

[9] This phenomenon was first described in Gillborn and Youdell's (1999) study of concomitant increases in average student performance and a growing achievement gap in English schools. It was then popularized in the context of NCLB accountability systems by Booher-Jennings (2005).

directions. We also interact each of these variables with the high-stakes grade indicator (hs_{gst}). Under the educational triage hypothesis, interactions of distance from the marginal student with urgency and the high-stakes indicator will have a negative effect on achievement.

F. Probability student i scores proficient on next test administration ($\hat{\pi}_{igst}$)

From NWEA's technical reports we identify an NWEA equivalent to each state's proficiency cutscore (c_{gst}). We then model the probability that student *i* reaches proficiency on the next state test as the probability that *i*'s fall test score on the NWEA assessment (f_{igst}) plus the expected gain for that student (m_{igst}) exceeds the NWEA cutscore-equivalent. This is expressed as:

$$\hat{\pi}_{igst} = \text{Prob}(c_{gst} - (f_{igst} + m_{igst})).$$

m_{igst} is simply the mean gain for a given state and grade within the sample period, obtained by regressing the observed fall-to-spring gain on student *i*'s fall score and a set of state and grade dummy variables. The variance of $f_{igst} + m_{igst}$ equals the variance of the forecast error from this equation (σ^2_1) plus the variance of the test measurement error (σ^2_2). On the assumption that these errors are normally distributed, we obtain:

$$\hat{\pi}_{igst} = \Phi((c_{gst} - f_{igst} - m_{igst})/\surd\ (\sigma^2_1 + \sigma^2_2)). \quad (1)$$

While we do not suppose that school districts carry out calculations such as these, it does not seem unreasonable to suppose that teachers acquire a sense of the probability that a student with a particular level of performance in the fall will reach proficiency on the next state test, and that our variable, $\hat{\pi}_{igst}$, approximates a teacher's own informal estimates.

We have also compared values of $\hat{\pi}_{igst}$ with a similar set of calculations conducted by NWEA. NWEA published tables of the probability that a student with a particular f_{igst} will pass the high-stakes assessment in state *s*. These tables include scores at five point intervals, requiring interpolation to find pass probabilities for values of f_{igst} that fall between each interval. Although we do not have tables for all states in all years included in our study, the NWEA estimates are highly correlated with our pass probabilities where both exist (ρ = .88). Furthermore, the estimates reported below are qualitatively similar using either $\hat{\pi}_{igst}$ or the set of values reported by NWEA.

5. Results

A. Descriptive Statistics

Table 3 contains descriptive statistics on three key variables: the annualized fall-to-spring gain; students' own probabilities of scoring proficient; and the marginal student's probability of scoring proficient. Annual growth averages about 9 scale score points in the lower elementary grades. It declines with advancing grade until the average is about half that large in grade eight. None of these between-grade differences affect us, as we estimate our model separately for each grade level.

Insert Table 3 here.

The predicted probability that the marginal student achieves proficiency is about 80 percent in the lower elementary grades, declining about ten percentage points by middle school. This again reflects the fact that most schools made AYP. However, there is considerable variation in this probability. A similar pattern is evident in the values of $\hat{\pi}_{igst}$. Indeed, the mean and standard deviation of this probability are similar to the corresponding statistics for the marginal student, suggesting that on average, the marginal student is not far from the average.

Finally, we see there is little difference between the full sample and the restricted sample with respect to these variables.

B. Relationship between annualized gains and fall scores

Before turning to estimates from the four modeling strategies described above, we first present graphical evidence on the effects of NCLB by grade for the complete sample and then for a restricted sample. In contrast to our models, very few assumptions underlie the graphs. The fact that the graphical evidence is consistent with the estimates of our models indicates that our findings are not the result of strained assumptions about functional form or the construction of key variables.

Figures 1a through 1f display the relationship between annualized test score gains and fall test scores, where the latter are centered on each state's proficiency cutscore. The darker set of three curves in each figure represents the relationship for high-stakes years and the lighter set the relationship for low-stakes years. The three curves represent point estimates flanked by a 95 percent confidence interval, which were obtained by fitting the following generalized additive model (GAM):

$$(2)\quad Y_{istg} = \Sigma_{s=1,7}\, \mu_{sg} + \Sigma_{t=1,4}\, \mu_{tg} + \Sigma_{j=1,2}\, \psi_{jg}(F_{istg} - C_{stg}) + \delta\, hs_{gst} + u_{istg} \quad \begin{array}{l} g = \text{grade } (3\ldots7), j = hs, ls \\ s = \text{state } (1,\ldots7) \end{array}$$

where, Y_{istg} is the annualized gain for student *i* in grade *g*, state *s*, and year *t*, μ_{sg} is a state fixed effect, μ_{tg} is a year fixed effect, and $\psi_{jg}(F_{istg} - C_{stg})$ is an unknown function of the fall score/cutscore difference. The ψ_{jg} are approximated by cubic splines, with penalties for departures from smoothness. The smoothing parameter is chosen using generalized cross-validation methods described in Wood (2006). Both functions are identified only up to a location parameter. To estimate the vertical gap between the high-stakes and low-stakes curves, the high-stakes grade indicator (hs_{gst}) is included in (2).

As evidenced in Figures 1a through 1f, third grade is anomalous in that the high-stakes curve lies everywhere below the low-stakes curve while the two curves cross in fourth through eighth grades. In select grades the intersection occurs near the center of the distribution of $F_{istg} - C_{stg}$ (shown at the bottom of each figure), but in others the high- and low-stakes curves cross farther to the left of the distribution. It is also apparent most grades have a more pronounced S-shape in low-stake grade-year combinations, with a bulge near the middle of the distribution, where instruction may have been pitched. This bulge is much attenuated in high-stakes years. Curves in high-stakes years also appear to "straighten out," with smaller gains for above average students and larger gains among below average students. Only in third grade are students at the extreme left of the distribution gaining significantly less, which is also true for the entire distribution of third-graders in high-stakes years.

The neglect of the lowest achievers detected by Neal and Schanzenbach (2007) in Chicago does not appear to characterize the schools in our sample. At the extreme high end of the achievement distribution there is also little evidence that students are harmed by NCLB, though estimates are imprecise at the extremes as can be seen by a fanning out of the three lines. Rather, these figures suggest the redistribution of gains is from middle high to middle low. This is, of course, consistent with the hypothesis that schools are "dumbing down" instruction for the majority of students in order to stress basic skills. However, the magnitude of these effects is small compared to mean grade-level gains.

We have repeated this graphical analysis in using samples restricted in two ways (Figures 2a through 2f). States are retained only if the grade in question was designated as a low-stakes grade in some years and a high-stakes grade in others. States in which the grade was always one or the other (and for which we therefore have no direct contrast between low- and high-stakes regimes) are dropped. Second, within the states that we retain, we keep the set of schools constant. Schools that come in and out of the data set as contracts with NWEA begin or end are dropped.[10] Despite these restrictions, it is apparent the relationship between annualized gains and fall scores as displayed in Figures 2a through 2f are quite similar to those in Figures 1a through 1f. This similarity suggests that

[10] In general this means we kept schools that were present during all four years for which we have data. In some cases, where the number of districts contracting with NWEA rose rapidly between 2002-03 and 2003-04, we have dropped the former year, preferring a sample with fewer years but more schools to a longer panel with a smaller number of schools and students.

findings using the full sample are not unduly affected by changes in the mix of schools contracting with NWEA.[11]

C. NCLB main effects

Our first set of models is devoted to NCLB main effects. Table 4 displays estimates for the three explanatory variables: a student's own probability of achieving proficiency on the next administration of the state test ($\hat{\pi}_{igst}$); an indicator for high-stakes grades (hs_{gst}); and an interaction of these two variables ($\hat{\pi}_{igst}$ x hs_{gst}).[12] The coefficient on the high-stakes indicator represents the shift (outward or inward) of the relationship depicted in Figures 1a through 1f, while the coefficient on interaction represents the change in the slope between high-stakes and low-stakes years. Panel A contains results for the full sample, and Panel B does so for the restricted sample.

Insert Table 4 Here

Except for grade three, we find, as expected, that in high-stakes years there is an outward shift—accountability tends to increase achievement across the board. This is true in both the full and restricted samples. The magnitude of the effect varies by grade, ranging from 10 to 30 percent of the mean gain for the grade level. The relative effect of this shift is especially pronounced in the higher grades, where mean gains are smallest.

The relationship between achievement gains and the probability of achieving proficiency (as a function of the fall score) is negative, again as expected. Much of this doubtless reflects regression to the mean. The distributional effect of NCLB is represented by the coefficient on the interaction of this probability with the high-stakes indicator. As this coefficient is negative (except in grade three),

[11] The greatest dissimilarity arises in grade 8. Here our restrictions on the sample had the greatest effect: only Minnesota designated grade 8 as a low-stakes grade, switching its status to high-stakes in 2005-06. Thus there is only one year of data in one state to estimate the high-stakes curves. As Figure 2f shows, the cross-validation algorithm for the choice of smoothing parameter failed and the program defaulted to a linear model.

[12] All regressions also include binary indicators for year and for state. A complete set of results is available upon request.

the relationship becomes steeper, favoring students at the low-end of the distribution vis-a-vis high achievers. This is true both in the initial and the restricted samples and is consistent with the hypothesis that NCLB has led schools to sacrifice gains by high-performing students in order to promote achievement at the low end of the distribution.

Finally, results for the restricted sample are, in fact, even more in line with our hypotheses than the results for the initial sample, suggesting that these findings are not attributable to the changing make-up of the full sample.

D. Urgency of improvement

Table 5 contains results from models that explored the urgency hypothesis: that NCLB effects would be strongest in schools at greatest risk of failing to make AYP. Recall that urgency is represented by the probability that the marginal student in the school fails to achieve proficiency ($\hat{\pi}_{Mst}$), interacted with the high-stakes indicator ($\hat{\pi}_{Mst}$ *x* hs_{gst}). As with NCLB main effects, the pressure a school faces to improve student performance can affect a shift in the relationship in Figures 1a through 1f, create a tilt in that relationship, or do both. We also include a main effect for the probability that the marginal student achieves proficiency.[13]

Insert Table 5 Here

Estimates reported in Panel A of Table 5 indicate that the main effect of $\hat{\pi}_{Mst}$ is positive. This is no surprise considering $\hat{\pi}_{Mst}$ is a summary measure of overall achievement at a school. However, the interactions generally run counter to expectations. We do not find that the outward shift in achievement is most pronounced in schools where the urgency to improve is greatest: on the contrary, the interaction of $\hat{\pi}_{Mst}$ with the high-stakes indicator is negative. Thus, the positive

[13] These models also include the three variables displayed in Table 4. However, due to the presence of the interactions in Table 5, the coefficients on those variables can no longer be interpreted as main effects. Our interest is therefore in the variables that have been added to the model. For sake of readability we omit the results for other variables. A complete set of results for all models is available from the authors upon request.

coefficient on the high-stakes indicator (hs_{gst}) in Table 4 was due to responses from schools at least risk of failing to make AYP.[14] The tilt in the achievement profile is likewise unexpected—it is the schools least threatened with NCLB sanctions that seem most willing to trade high-performers' gains for low-performers' gains. Estimates in Table 5 further suggest there is virtually no difference in this respect between the estimates obtained using the complete sample and those obtained with the restricted sample.

Model misspecification may play a role in these perverse findings, if our measure of urgency does not in fact capture the likelihood a school will face sanctions under NCLB. We determined the marginal student by ranking all students tested in a school's high-stakes grade and counting down M positions, where M represents the number of students who must score proficient for the school to make AYP. However, as NCLB requires subgroups within a school to reach the same proficiency target (expressed as a percentage of the subgroup), the relevant margin may not be the Mth student overall, but the student occupying the corresponding position within the weakest subgroup.

To further explore this possibility, we identified the marginal student in each subgroup defined by race or ethnicity.[15] Our alternative "marginal student" is the student with the minimum probability of achieving proficiency among this set.[16] Results using this minimum marginal student are displayed in Panels C and D of Table 5. Estimates are somewhat more consistent with our initial hypotheses, though perverse effects remain in evidence. Wrong-signed coefficients are still encountered more frequently than those of the anticipated sign, though their magnitude is smaller

[14] This is evident in the (unshown) coefficient on the high-stakes indicator in this set of models, which becomes even more positive than the values reported in Table 3.

[15] Only subgroups whose membership exceeds the threshold at which the group counts for NCLB purposes are used.

[16] The minimal marginal student is used for the entire school, as opposed to defining a different marginal student within each subgroup for students belonging to that subgroup. Schools are unlikely to organize instruction by racial and ethnic subgroups (for legal reasons, if no other). Thus, if schools are trading off achievement of high-performing students to raise scores among low-performing students (say, through a reorganization of the instructional day, or reassignment of the most effective teachers), we expect all low-performing students, regardless of the subgroup to which they belong, to share in the resulting gains.

than those reported in Panels A and B of Table 5. It remains the case that much of the response to NCLB, as shown in the main effects of Table 4, arises in schools that are not at high risk of facing NCLB sanctions.

E. Bubble students

To test the bubble hypothesis, we added two more variables to the model: the interaction of our urgency measure with each student's absolute distance from the marginal student $(1-\hat{\pi}_{Mst})$, using the probability metric $|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|$; and the interaction of that two-way interaction with the high-stakes indicator $(|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|(1-\hat{\pi}_{Mst})\ x\ hs_{gst})$. We expect the latter, three-way interaction to enter with a negative sign if the bubble hypothesis is true: in high-stakes years, among schools at risk of failing to make AYP, the incentive will presumably be greatest to neglect students far from the margin. We have no strong expectations regarding the two-way interaction: in the absence of accountability, it is not clear how schools target resources (though if the marginal student represents a median or "representative" student, it may be advantageous to be near him).

Estimates displayed in Table 6 indicate there is remarkably little support for the bubble hypothesis in these data. In Panels A and B, most of the coefficients on the three-way interaction are positive. In Panels C and D, where we employ our alternative definition of marginal student, all the coefficients are positive and, with two exceptions, statistically significant. Coefficients on the two-way interaction are mixed, tending to be negative in Panels A and B where significant, but positive in Panels C and D. It should be borne in mind that these models contain all the variables shown in Table 5, so that the bubble effects we are looking for arise in addition to the negative relationship between gains and initial achievement depicted in Figures 1a through 1f. Thus we are looking for evidence of some concavity (a bowing out) in that relationship in the vicinity of the marginal student, among schools are most likely to fail to make AYP. We do not find it: if anything, that curve appears to bow out at the two ends, the further one gets from the marginally performing student.

Insert Table 6 Here

F. Educational triage

The specification of the model in Table 6 may mask differences in the effect of NCLB on students who are above the marginal student and those who are below. One might be especially concerned about harm to students who are far below the marginal student, even if there are no negative consequences of NCLB among high-achieving students. While educational triage would presumably affect both the highest and lowest performing students, parents of the former may take steps to ensure that their children are not neglected in the classroom. No such countervailing influence may be present to protect the weakest students.

To further explore this possibility, we revise our model to distinguish between positive and negative directions in $|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|$. We do so by estimating separate coefficients for differences in the positive $(|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|(1-\hat{\pi}_{Mst})\ x\ \delta_{pos})$ and the negative $(|\hat{\pi}_{igst} - \hat{\pi}_{Mst}|(1-\hat{\pi}_{Mst})\ x\ \delta_{neg})$ directions. As in other models, the effect of NCLB is represented by the interaction of these variables with the high-stakes grade indicator (hs_{gst}). If educational triage is taking place, these interactions will enter with negative coefficients.

Estimates reported in Table 7 indicate results are mixed, depending on grade, sample, and definition of the marginal student. Evidence of educational triage working to the detriment of the lowest performers is strongest in the lower grades in Panel A. However, this finding is not robust. When the restricted sample is used, only low-achieving students in grade four are harmed by the NCLB accountability system. The other coefficients are insignificant or positive, and in the middle school grades, the positive effects are both statistically and substantively significant.

Insert Table 7 Here

By and large this pattern holds up when the alternative definition of marginal student is employed, regardless of sample. Panels C and D of Table 7 points out negative effects are limited to grades four and five. Further, there are substantial positive effects for low achievers in grade six and higher. Results for high-performing students are less consistent.

6. Conclusion

This study has examined whether NCLB has raised achievement of lower-performing students and, if so, whether these gains have come at the expense of students that are already proficient or that are far below the proficiency standard. Identification was achieved by exploiting the fact that in the early years of NCLB, not all grades counted for purposes of determining AYP. Analysis drew upon longitudinal, student level test score data from seven states (N > 2,000,000) between 2002-03 and 2005-06 school years.

Results indicate that NCLB is having an effect on public education in the United States in the expected direction. There has been a tendency for scores to rise across the board, accompanied by a "redistribution" of achievement gains from high-performing to low-performing students. While the redistribution is large enough to make the highest-performing students net losers under NCLB, through most of the achievement range the combined effect is positive. Arguably these effects are of the kind anticipated by proponents of NCLB. To this extent, NCLB appears to be "working." These findings are of greater significance inasmuch as our measures of achievement are not based on high-stakes tests used to determine whether schools are making AYP. Thus, our estimates are not picking up the effects of teaching to the test or other attempts to game the accountability system.

However, the mechanism by which these positive results are produced is far from clear. On several counts, our findings are at variance with the conventional wisdom on NCLB and cast doubt on some of the strategies employed by researchers to identify an NCLB effect. We find no evidence that NCLB effects are largest among schools most likely to face sanctions. Some of our findings show the opposite: that the response to NCLB has been greatest among schools least threatened by sanctions, with little response (or even a perverse response) elsewhere. While these results are somewhat sensitive to alternative specifications of the model and restrictions placed on the regression sample, at best the response to NCLB appears to be at least as strong among schools where the probability of making AYP is high as it is among schools much more likely to fail.

One possible explanation is that a small probability of failure means more to a school with a reputation for success than does a much larger probability of failure in a school inured to poor performance. Capacity may also play a role. Notwithstanding their greater incentive to improve, low-performing schools may find the challenge posed by NCLB overwhelming, while more successful schools with smaller numbers of low-achieving students may find it considerably easier to provide the remedial assistance necessary to boost performance.

NCLB accountability systems provide schools an incentive to focus on marginal students whose success is critical in the attempt to make AYP, and to neglect students certain to score proficient as well as students far below the margin. At least since the work of Booher-Jennings (2005), it has been claimed that schools are responding as one would expect to these incentives. However, we find no evidence that students near the margin are learning more than students far away from it under NCLB. When we break out separate results for students above and below the margin, we find some evidence (though spotty) that schools are neglecting the lowest achievers in the elementary grades, but by the middle school grades this has disappeared as evidenced by the fact that predicted gains are greatest among the lowest-achieving students. This may reflect a greater prevalence of ability grouping in middle schools and junior high schools, possibly in combination with a concerted push to prepare these students for high school. Notably, we do not see such patterns prior to the onset of NCLB; rather, they occur when a grade switches from being low-stakes to high-stakes.

We also fail to find consistent evidence that schools struggling to make AYP neglect their high achievers in order to focus on students below the state-defined proficiency cutscore. NCLB makes high achieving students in such schools a prized commodity. Schools may go to extra lengths to ensure that these students continue to do well and remain enrolled in the school. By contrast, it may be schools with an abundance of such students that feel no particular urgency to promote their gains.

At least one of our findings should come as good news. Except in grades 3 and 4, NCLB has not promoted a form of educational triage that writes off the lowest-achieving students as too far behind to be helped. Perhaps most reassuring, the biggest turnaround between low- and high-stakes years occurred for eighth graders, suggesting that schools are making extraordinary efforts to reverse years of pre-NCLB neglect, even though these students will soon matriculate to a high school and no longer count toward that school's AYP status. At the same time, this conclusion is tempered by the fact that we have data from only one state in which grade 8 went from being low-stakes to high-stakes.

References

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

Chakrabarti, R. (2007). *Vouchers, public school response, and the role of incentives: Evidence from Florida*. Staff report #306, Federal Reserve Bank of New York.

Deere, D., and Strayer, W. (2001). *Putting schools to the test: School accountability, incentives, and behavior.* Unpublished manuscript, Texas A&M University.

Diamond, J.B. and Spillane, J. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record*, 106(6), 1145-1176.

Gillborn, D. and Youdell, D. (1999). Rationing Education: Policy, Practice, Reform, and Equity. Philadelphia, PA: Open University Press.

Grissmer, D. and Flanagan, A. (1998). *Exploring the rapid achievement gains in North Carolina and Texas*. A report from the National Education Goals Panel.

Holmes, G. M. (2003). *On teacher incentives and student achievement*. Unpublished working paper, East Carolina University, Department of Economics.

Jacob, B. (2005). Accountability, incentives, and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89, 761-796.

Jacob, B. and Levitt, S. (2007). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3), 761-796.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, 37(4), 752-777.

Krieg, J.M. (2008). Are students left behind? The distributional effects of the No Child Left Behind Act. *Education Finance and Policy*, 3(3), 250-281.

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

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

Rosaen, A., Schwartz, N., and Forbes, T. (2007). *The effects of school failure: Using regression discontinuity to measure the impact of California's No Child Left Behind Policies.* Unpublished working paper, University of Michigan, Gerald R. Ford School of Public Policy.

Rouse, C.E., Hannaway, J., Goldhaber, D., and Figlio, D. (2008). *Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure.* CALDER working paper: Urban Institute.

Sims, D.P. (2007). *Can failure succeed? Using racial subgroup rules to analyze vouchers, stigma, and school accountability*. Unpublished working paper. Brigham Young University, Department of Economics.

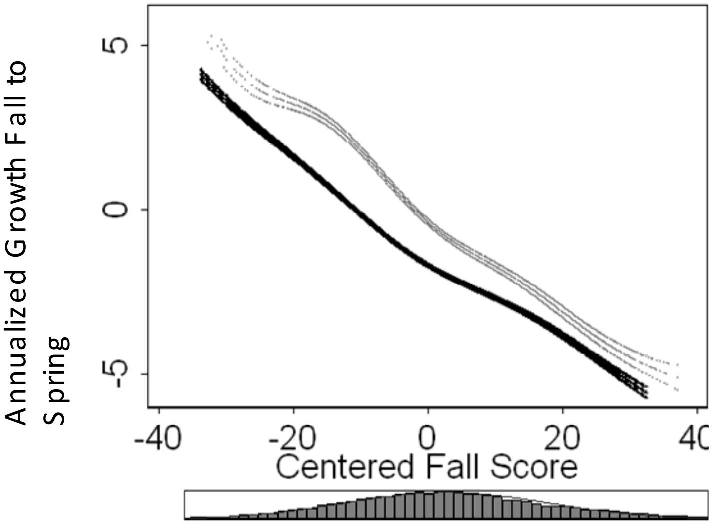
Springer, M.G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, 27(5), 556-563.

Stuit, D. and Cravens, X. (2007). *The impact of NCLB accountability on Washington's schools.* Unpublished working paper, Peabody College of Vanderbilt University, Department of Leadership, Policy, and Organizations.

White, K.W. and Rosenbaum, J.E. (2007). Inside the blackbox of accountability: How high-stakes accountability alters school culture and the classification and treatment of students and teachers. In A. Sadvonik, J. O'Day, G. Bohrnstedt, and K. Borman (Eds.), *No Child Left Behind and the Reduction of the Achievement Gap: Sociological Perspectives on Federal Education Policy*. New York: Routledge.

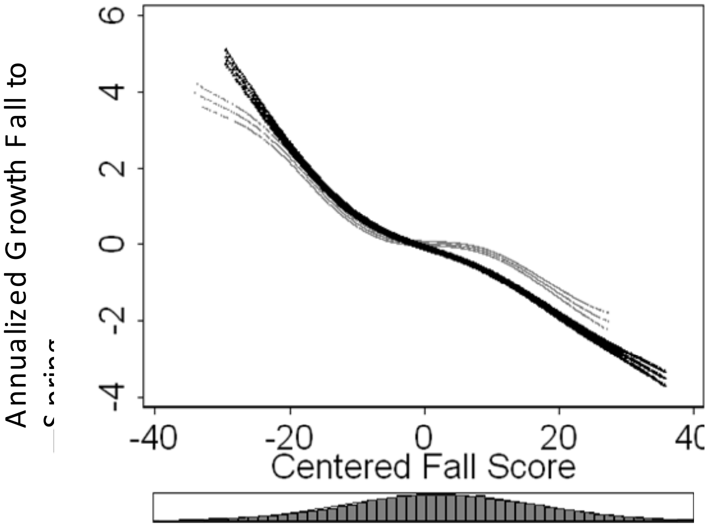
Wood, Simon N. 2006. Generalized Additive Models. An Introduction with R. London: Chapman and Hall.

Figure 1a: Effect of NCLB, All States, Grade 3



3

Figure 1b: Effect of NCLB, All States, Grade 4



4

Figure 1c: Effect of NCLB, All States, Grade 5

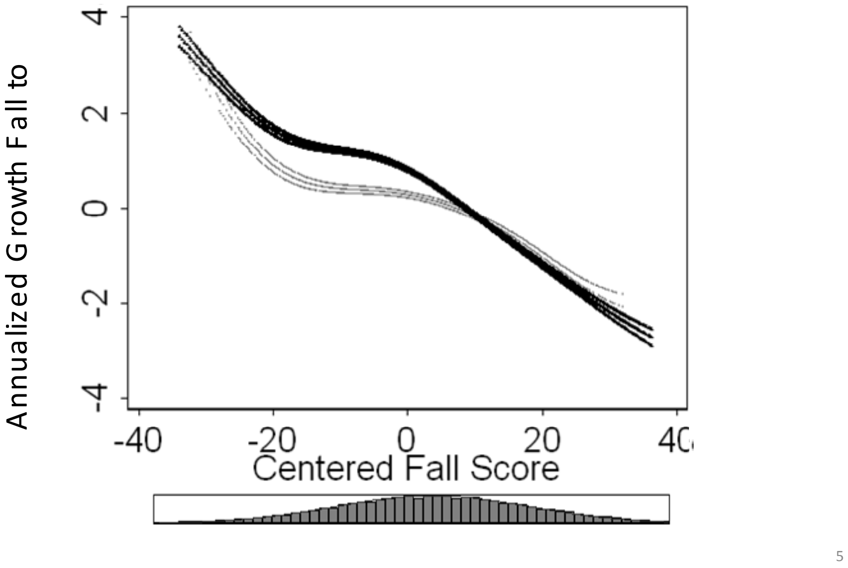


Figure 1d: Effect of NCLB, All States, Grade 6

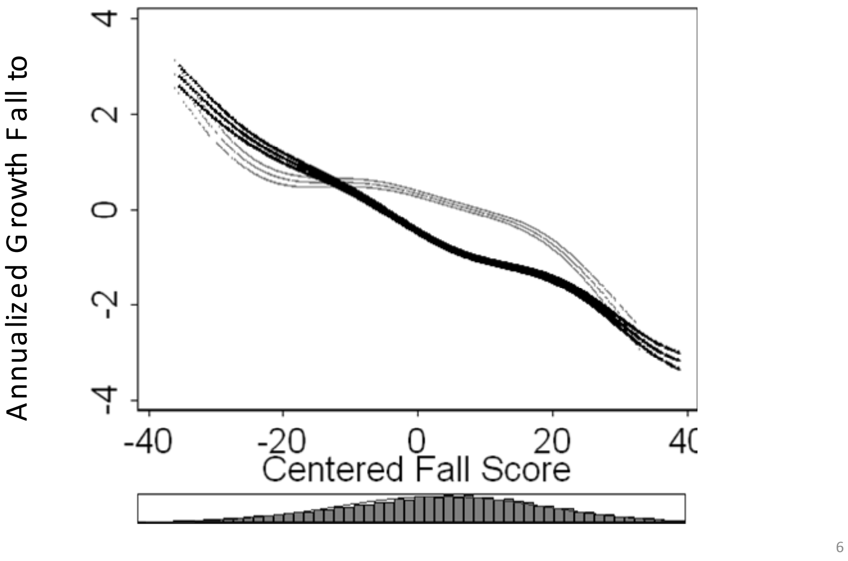
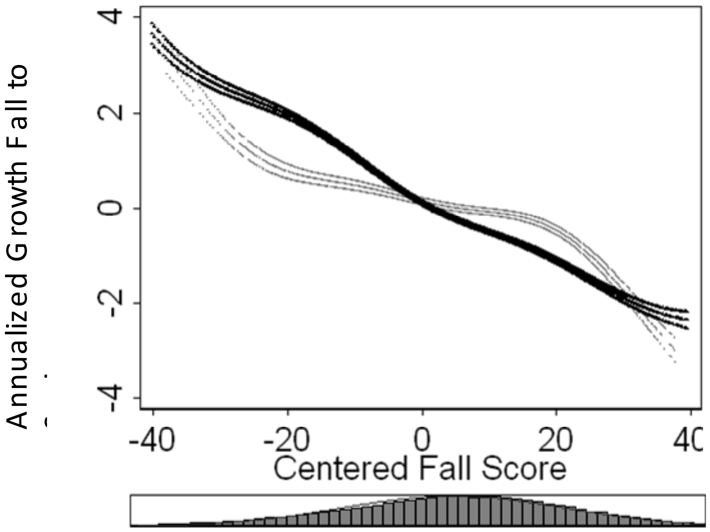
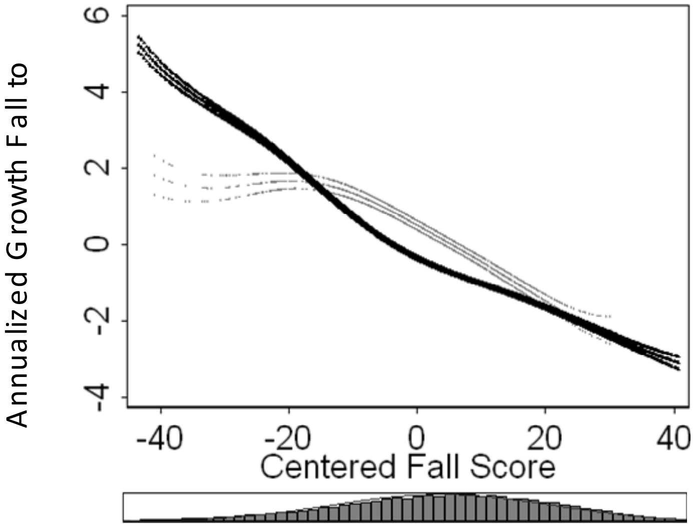


Figure 1e: Effect of NCLB, All States, Grade 7



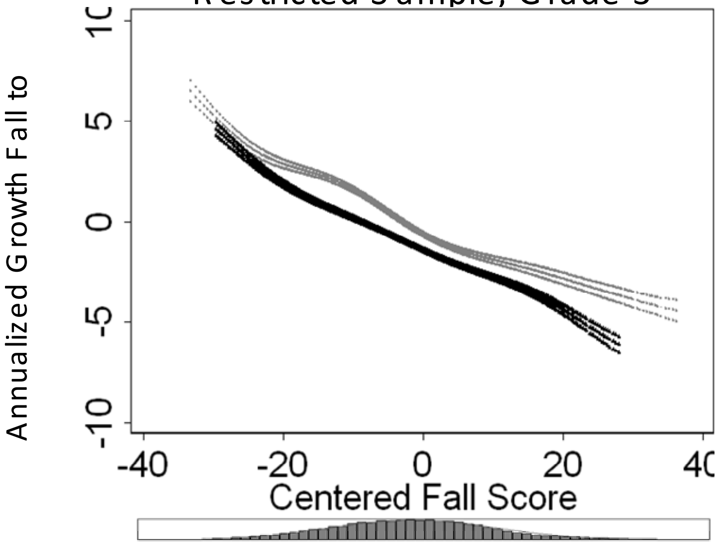
7

Figure 1f: Effect of NCLB, All States, Grade 8



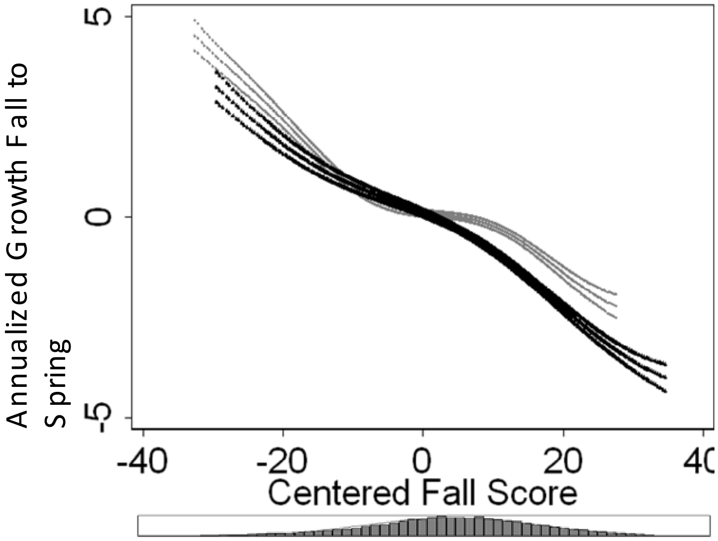
8

Figure 2a: Effect of NCLB, Restricted Sample, Grade 3



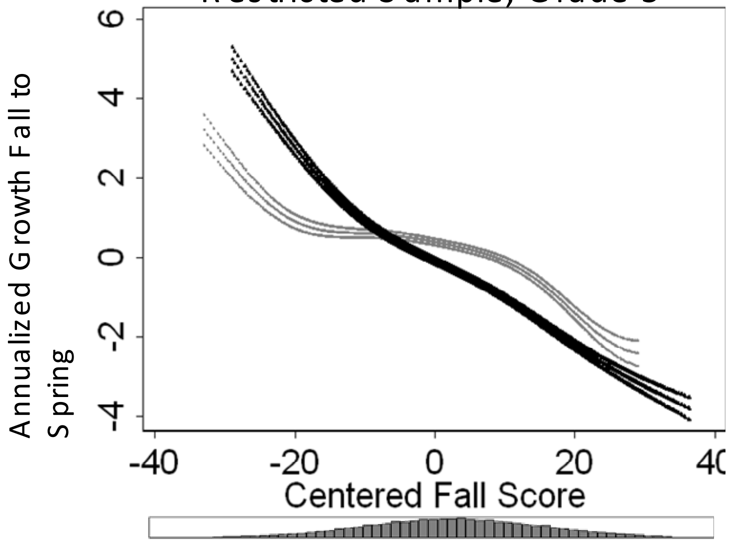
10

Figure 2b: Effect of NCLB, Restricted Sample, Grade 4



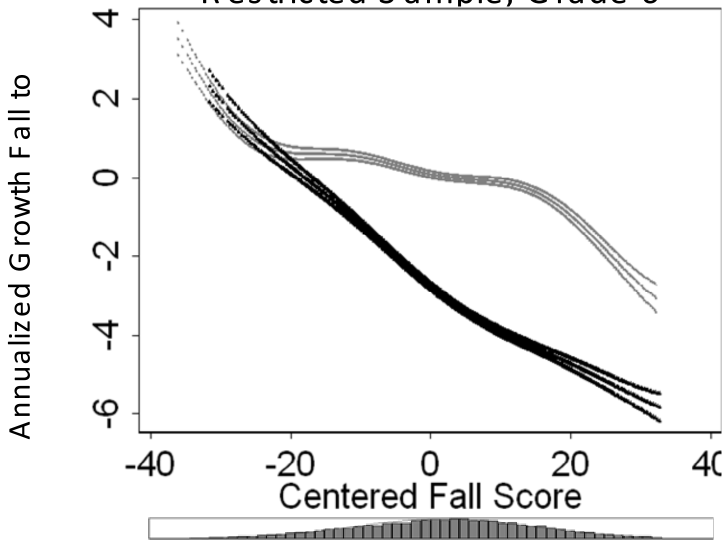
11

Figure 2c: Effect of NCLB, Restricted Sample, Grade 5



12

Figure 2d: Effect of NCLB, Restricted Sample, Grade 6



13

Figure 2e: Effect of NCLB, Restricted Sample, Grade 7

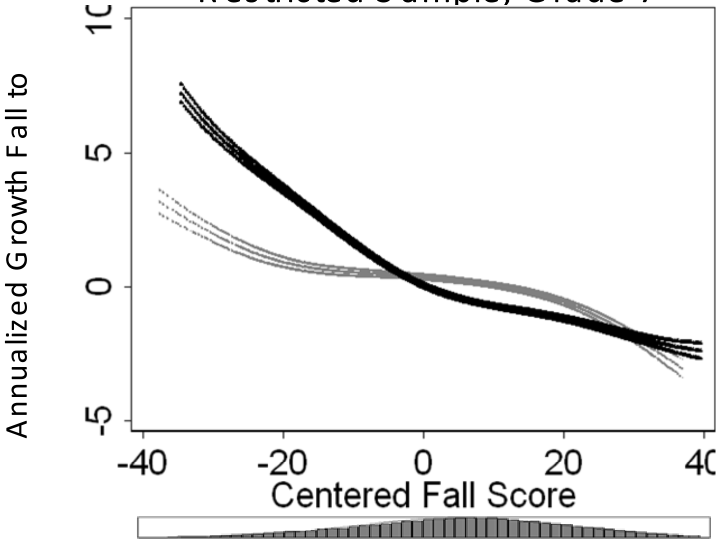


Figure 2f: Effect of NCLB, Restricted Sample, Grade 8

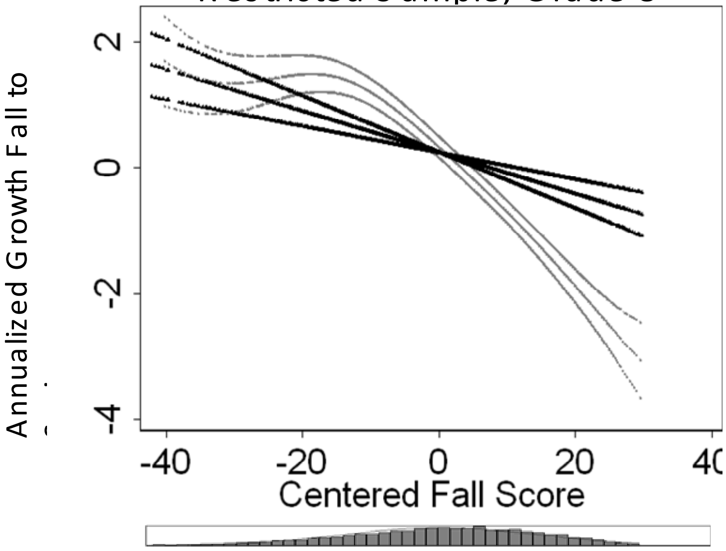


Table 1. Number of Observations by State, Grade, and Year

State	**Grade**	**Year**				
Arizona		02-03	03-04	04-05	05-06	All Years
	3	***2,963***	***3,550***	***5,347***	***3,738***	15,598
	4	2,489	3,447	5,095	***3,900***	14,931
	5	***2,505***	***3,697***	***5,082***	***3,763***	15,047
	6	2,494	3,100	4,602	***3,608***	13,804
	7	2,356	3,091	4,441	***3,645***	13,533
	8	***2,191***	***3,297***	***4,333***	***3,672***	13,493
Colorado						
	3	10,633	8,450	***9,577***	***10,128***	38,788
	4	10,848	8,292	***9,135***	***10,020***	38,295
	5	***11,159***	***8,504***	***9,468***	***10,447***	39,578
	6	***8,581***	***7,530***	***7,710***	***8,827***	32,648
	7	***8,433***	***6,846***	***7,475***	***8,647***	31,401
	8	***8,001***	***6,920***	***7,027***	***8,206***	30,154
Idaho						
	3	14,943	***17,943***	***18,857***	***19,038***	70,781
	4	***15,346***	***18,102***	***18,459***	***19,311***	71,218
	5	15,663	18,558	***18,636***	***18,896***	71,753
	6	15,506	18,898	***18,950***	***19,062***	72,416
	7	16,020	***19,098***	***19,517***	***19,576***	74,211
	8	***16,099***	***19,193***	***19,486***	***19,935***	74,713
Indiana						
	3	***21,570***	***23,259***	***23,106***	***24,391***	92,326
	4	25,931	23,830	***22,914***	***24,257***	96,932
	5	26,249	24,455	***23,212***	***24,548***	98,464
	6	***22,819***	***22,927***	***23,646***	***24,218***	93,610
	7	25,795	23,066	***21,684***	***23,497***	94,042
	8	***21,108***	***20,887***	***21,861***	***22,468***	86,324
Michigan						
	3	1,808	2,753	6,443	***6,327***	17,331
	4	***1,727***	***2,786***	***6,240***	***6,622***	17,375
	5	***1,771***	***2,909***	***6,204***	***6,451***	17,335
	6	1,760	2,857	6,494	***6,246***	17,357
	7	***1,638***	***2,758***	***5,907***	***5,771***	16,074
	8	***1,296***	***2,506***	***5,616***	***5,500***	14,918
Minnesota						
	3	***3,462***	***11,518***	***19,613***	***25,035***	59,628
	4	4,260	13,121	20,824	***23,719***	61,924

	5	***4,021***	***12,640***	***20,636***	***24,382***	61,679
	6	4,251	13,664	22,405	***25,831***	66,151
	7	3,517	***12,473***	***20,374***	***22,935***	59,299
	8	1,801	10,190	17,531	***20,262***	49,784
Wisconsin						
	3	1,253	3,795	6,557	***6,743***	18,348
	4	***709***	***2,313***	***4,464***	***7,037***	14,523
	5	2,071	4,033	7,486	***8,508***	22,098
	6	1,940	6,805	9,878	***9,025***	27,648
	7	1,397	5,984	9,576	***9,881***	26,838
	8	***689***	***4,173***	***6,505***	***9,431***	20,798

Entries in **boldface** and *italics* are high-stakes grades and years.

Table 2. Summary of Models' Explanatory Variables and Hypothesized Direction of Effects

Explanatory Variables	Definitions	
π_i	Estimated probability student i scores at the proficient level on the next administration of the state accountability tests. We expect a negative sign, if only because of regression to the mean.	< 0
hs	Binary indicator of high-stakes grade/year. If accountability improves school efficiency, performance can improve across the board.	> 0
$hs \text{ x } \pi_i$	Interaction of high-stakes indicator with student i's probability of passing. A negative slope would indicate that lower-achieving students are getting more attention when NCLB stakes are high.	< 0
π_m	Estimated probability that the marginal student scores at the proficient level on the next administration of the state accountability tests. When students are ranked with respect to π_i, the marginal student is the last one from the top who needs to score proficient for the school to make AYP. As a high value of π_m is indicative of a high-achieving school, the expected sign is positive.	> 0

hs x $(1-\pi_m)$	Interaction of high-stakes indicator with the probability the marginal student fails to score proficient. $(1-\pi_m)$ is a measure of the urgency with which schools need to focus on raising achievement of students near the margin. If π_m is close to 1, the school is relatively assured of making AYP without altering instructional practices. If π_m is low, greater changes are required. To the extent this is an impetus to improve overall efficiency, the expected sign is positive.	> 0
hs x $(1-\pi_m)$ x π_i	Interaction of the previous variable with student i's probability of scoring proficient. If greater urgency causes schools to focus more on raising achievement of lower-performing students (and the school succeeds), the expected sign is negative.	< 0
$\lvert\pi_i-\pi_m\rvert(1-\pi_m)$	The absolute difference of student i's probability of scoring proficient and the marginal student's probability, times our measure of urgency. This variable tests the bubble hypothesis--that schools under particular pressure to make AYP will focus on students near the margin and ignore those whose probability of passing is much higher or much lower. If correct, the hypothesis implies a negative sign on this variable in high-stakes grades/years. The effect of the variable immediately below is therefore expected to be negative. The sign on this variable (in non-high-stakes grades/years) is not as clear.	?
$\lvert\pi_i-\pi_m\rvert(1-\pi_m)$ x hs	See above.	< 0

$\lvert\pi_i-\pi_m\rvert(1-\pi_m)\ x\ \delta_{pos}$	The next four variables relax the assumption that students well above the marginal student are treated the same as students well below the marginal student. Each of the preceding variables is interacted with binary indicators of whether π_i-π_m is positive or negative. Under the triage hypothesis, interactions of distance from the marginal student with urgency and the high-stakes indicator will have a negative effect on achievement.	?
$\lvert\pi_i-\pi_m\rvert(1-\pi_m)\ x\ \delta_{neg}$	See above.	?
$\lvert\pi_i-\pi_m\rvert(1-\pi_m)\ x\ \delta_{pos}\ x\ hs$	See above.	< 0
$\lvert\pi_i-\pi_m\rvert(1-\pi_m)\ x\ \delta_{neg}\ x\ hs$	See above.	< 0

Table 3: Sample Descriptive Statistics

		Full Sample			Restricted Sample		
Variable	Grade	# obs	Mean	Std. Dev.	Obs	Mean	Std. Dev
Annualized fall-spring growth	3	274806	9.42	6.32	166801	9.69	6.14
Probability of scoring proficient	3	275997	0.77	0.29	167527	0.77	0.28
Marginal student's probability of scoring proficient	3	267552	0.76	0.24	162692	0.76	0.22
Annualized fall-spring growth	4	278833	8.22	6.21	171713	8.58	6.10
Probability of scoring proficient	4	279940	0.82	0.26	172362	0.82	0.26
Marginal student's probability of scoring proficient	4	273203	0.77	0.23	168758	0.76	0.22
Annualized fall-spring growth	5	288254	7.56	6.28	169963	7.78	6.19
Probability of scoring proficient	5	289669	0.80	0.28	170903	0.79	0.28
Marginal student's probability of scoring proficient	5	280597	0.77	0.23	167836	0.77	0.23

Annualized fall-spring growth	6	282088	6.13	6.46	161312	6.36	6.36
Probability of scoring proficient	6	283619	0.77	0.31	162206	0.76	0.31
Marginal student's probability of scoring proficient	6	273133	0.76	0.25	156752	0.76	0.24
Annualized fall-spring growth	7	272918	5.02	6.51	172316	5.16	6.36
Probability of scoring proficient	7	274339	0.73	0.33	173319	0.73	0.33
Marginal student's probability of scoring proficient	7	265726	0.74	0.26	171089	0.73	0.25
Annualized fall-spring growth	8	256039	4.33	6.58	159331	4.57	6.35
Probability of scoring proficient	8	258274	0.69	0.36	160846	0.69	0.35
Marginal student's probability of scoring proficient	8	257832	0.74	0.26	160480	0.73	0.25

Table 4. NCLB Main Effects

		Panel A: Overall Marginal Student, Full Sample, by Grade					
Variable	Hypoth. Sign	3	4	5	6	7	8
Own probability of	-	-7.01	-3.57	-2.08	-1.83	-1.86	-3.04
achieving proficiency (π_i)		(0.11)	(0.07)	(0.08)	(0.07)	(0.08)	(0.11)
High-stakes grade (hs)	+	-1.87	0.93	0.76	0.66	1.00	0.71
		(0.10)	(0.09)	(0.09)	(0.08)	(0.08)	(0.10)
Interaction (hs x π_i)	-	1.19	-1.47	-0.98	-1.30	-1.50	-1.35
		(0.12)	(0.09)	(0.10)	(0.08)	(0.09)	(0.12)

		Panel B: Overall Marginal Student, Restricted Sample					
Own probability of	-	-6.74	-3.76	-2.02	-1.81	-1.92	-3.06
achieving proficiency (π_i)		(0.13)	(0.09)	(0.10)	(0.08)	(0.09)	(0.13)
High-stakes grade (hs)	+	-1.16	1.06	1.43	0.87	1.34	1.52
		(0.12)	(0.11)	(0.11)	(0.10)	(0.10)	(0.13)
Interaction (hs x π_i)	-	0.55	-1.67	-1.55	-1.97	-1.94	-2.34
		(0.14)	(0.12)	(0.12)	(0.11)	(0.11)	(0.14)

Table 5: Urgency of Improvement Hypothesis

		Panel A: Overall Marginal Student, Full Sample, by Grade					
Variable	Hypoth. Sign	3	4	5	6	7	8
Prob. marginal student	+	2.23	1.81	1.40	2.67	3.49	0.61
achieves proficiency (π_m)		(0.15)	(0.11)	(0.12)	(0.13)	(0.17)	(0.25)
Interaction of urgency with	+	-2.22	-2.67	-3.20	-2.84	-1.82	-3.22
high-stakes grade (hs x (1-π_m))		(0.20)	(0.22)	(0.21)	(0.20)	(0.22)	(0.28)
Three-way interaction:	-	2.06	3.86	4.81	2.04	3.12	0.54
(hs x π_i x ((1-π_m))		(0.20)	(0.25)	(0.19)	(0.17)	(0.16)	(0.16)
		Panel B: Overall Marginal Student, Restricted Sample					
Prob. marginal student	+	3.75	2.12	1.45	3.72	3.15	0.70
achieves proficiency (π_m)		(0.18)	(0.14)	(0.15)	(0.17)	(0.21)	(0.28)
Interaction of urgency with	+	-1.15	-1.47	-2.47	-1.97	-3.09	-2.24
high-stakes grade (hs x (1-π_m))		(0.25)	(0.29)	(0.28)	(0.28)	(0.28)	(0.33)
Three-way interaction:	-	3.12	3.33	4.81	2.46	3.42	-0.81
(hs x π_i x ((1-π_m))		(0.27)	(0.33)	(0.25)	(0.23)	(0.22)	(0.21)
		Panel C: Minimum Marginal Student, Full Sample					
Prob. marginal student	+	0.74	1.10	1.22	1.33	2.25	-0.59

achieves proficiency (π_m)		(0.15)	(0.10)	(0.11)	(0.10)	(0.11)	(0.18)
Interaction of urgency with	+	-1.37	-1.19	-0.12	-1.21	1.21	-1.53
high-stakes grade (hs x (1-π_m))		(0.18)	(0.16)	(0.16)	(0.13)	(0.13)	(0.19)
Three-way interaction:	-	-0.06	2.01	2.44	0.37	-0.01	-1.33
(hs x π_i x ((1-π_m))		(0.16)	(0.20)	(0.16)	(0.14)	(0.14)	(0.14)
		Panel D: Minimum Marginal Student, Restricted Sample					
Prob. marginal student	+	2.84	1.25	1.19	1.52	1.90	-0.44
achieves proficiency (π_m)		(0.19)	(0.13)	(0.14)	(0.12)	(0.13)	(0.20)
Interaction of urgency with	+	0.36	-0.47	0.39	-1.11	0.82	-1.51
high-stakes grade (hs x (1-π_m))		(0.23)	(0.21)	(0.20)	(0.17)	(0.15)	(0.21)
Three-way interaction:	-	0.89	1.86	2.58	0.87	-0.67	-1.98
(hs x π_i x ((1-π_m))		(0.21)	(0.26)	(0.21)	(0.20)	(0.17)	(0.18)

Table 6: Bubble Hypothesis

Variable	Hypoth. Sign	3	4	5	6	7	8
		Panel A: Overall Marginal Student, Full Sample, by Grade					
Interaction of urgency with distance	**?**	0.17	-0.86	-0.65	-0.97	0.08	-1.64
from marginal student$\|\pi_i-\pi_m\|(1-\pi_m)$		(0.31)	(0.22)	(0.34)	(0.34)	(0.60)	(0.40)
Three-way interaction: foregoing with	**-**	-0.26	1.21	-2.69	4.04	3.36	8.74
high-stakes $\|\pi_i-\pi_m\|(1-\pi_m)$ x hs		(0.38)	(0.37)	(0.42)	(0.42)	(0.64)	(0.46)
		Panel B: Overall Marginal Student, Restricted Sample					
Interaction of urgency with distance	**?**	0.55	-0.87	-0.97	-0.41	-0.23	-1.45
from marginal student$\|\pi_i-\pi_m\|(1-\pi_m)$		(0.34)	(0.27)	(0.44)	(0.44)	(0.66)	(0.44)
Three-way interaction: foregoing with	**-**	-1.21	2.47	-2.88	3.26	4.75	10.32
high-stakes $\|\pi_i-\pi_m\|(1-\pi_m)$ x hs		(0.44)	(0.47)	(0.55)	(0.56)	(0.72)	(0.52)
		Panel C: Minimum Marginal Student, Full Sample					
Interaction of urgency with distance	**?**	1.18	-0.01	-0.81	1.05	1.61	0.35
from marginal student$\|\pi_i-\pi_m\|(1-\pi_m)$		(0.34)	(0.22)	(0.32)	(0.23)	(0.29)	(0.29)
Three-way interaction: foregoing with	**-**	0.62	1.76	1.09	2.08	1.64	3.66

high-stakes $\|\pi_i-\pi_m\|(1-\pi_m)$ x hs		(0.38)	(0.31)	(0.37)	(0.28)	(0.32)	(0.32)
		Panel D: Minimum Marginal Student, Restricted Sample					
Interaction of urgency with distance	**?**	0.69	-0.24	-0.60	2.09	1.41	0.58
from marginal student$\|\pi_i-\pi_m\|(1-\pi_m)$		(0.39)	(0.28)	(0.42)	(0.28)	(0.31)	(0.32)
Three-way interaction: foregoing with	**-**	0.97	3.21	0.79	0.05	2.03	2.84
high-stakes $\|\pi_i-\pi_m\|(1-\pi_m)$ x hs		(0.44)	(0.40)	(0.48)	(0.36)	(0.35)	(0.36)

Table 7: Educational Triage Hypothesis

Panel A: Overall Marginal Student, Full Sample, by Grade

Variable	Hypoth. Sign	3	4	5	6	7	8
Interaction of urgency with positive distance	**0 or ?**	0.09	-3.46	-0.84	-0.84	-0.23	0.65
from marginal student: $\|\pi_i-\pi_m\|(1-\pi_m)$ x δ_{pos}		(0.37)	(0.28)	(0.37)	(0.37)	(0.66)	(0.66)
Interaction of urgency with negative distance	**0 or ?**	0.68	11.08	0.75	0.75	0.98	-7.01
from marginal student: $\|\pi_i-\pi_m\|(1-\pi_m)$ x δ_{neg}		(1.43)	(0.82)	(0.99)	(0.99)	(0.92)	(1.31)
Variable 1, interacted with high-stakes:	**-**	0.98	3.38	-2.88	-2.88	4.60	5.24
$\|\pi_i-\pi_m\|(1-\pi_m)$ x δ_{pos} x hs		(0.48)	(0.44)	(0.48)	(0.48)	(0.73)	(0.75)
Variable 2, interacted with high-stakes:	**-**	-3.45	-9.49	-3.06	-3.06	0.80	15.75
$\|\pi_i-\pi_m\|(1-\pi_m)$ x δ_{neg} x hs		(1.51)	(1.09)	(1.14)	(1.14)	(1.03)	(1.37)

Panel B: Overall Marginal Student, Restricted Sample

Interaction of urgency with positive distance	**0 or ?**	1.46	-3.41	-0.92	-0.04	0.26	1.78
from marginal student: $\|\pi_i-\pi_m\|(1-\pi_m)$ x δ_{pos}		(0.42)	(0.37)	(0.47)	(0.49)	(0.72)	(0.75)
Interaction of urgency with negative distance	**0 or ?**	-5.53	9.85	-1.36	-1.98	-1.46	-9.12
from marginal student: $\|\pi_i-\pi_m\|(1-\pi_m)$ x δ_{neg}		(1.62)	(1.07)	(1.20)	(0.88)	(1.05)	(1.50)
Variable 1, interacted with high-stakes:	**-**	-1.33	4.48	-3.72	1.30	5.63	6.03
$\|\pi_i-\pi_m\|(1-\pi_m)$ x δ_{pos} x hs		(0.56)	(0.58)	(0.63)	(0.64)	(0.84)	(0.91)
Variable 2, interacted with high-stakes:	**-**	3.20	-6.91	-0.29	9.44	3.74	19.27

$\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{neg} x hs		(1.74)	(1.41)	(1.44)	(1.19)	(1.22)	(1.61)
		Panel C: Minimal Marginal Student, Full Sample					
Interaction of urgency with positive distance	**0 or ?**	1.98	-1.61	-1.01	1.32	1.69	2.82
from marginal student: $\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{pos}		(0.39)	(0.27)	(0.34)	(0.24)	(0.30)	(0.48)
Interaction of urgency with negative distance	**0 or ?**	-4.62	8.05	0.74	-1.11	0.84	-8.28
from marginal student: $\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{neg}		(1.47)	(0.85)	(0.98)	(0.68)	(0.82)	(1.36)
Variable 1, interacted with high-stakes:	**-**	0.91	3.16	1.75	1.57	1.69	0.39
$\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{pos} x hs		(0.44)	(0.37)	(0.40)	(0.30)	(0.33)	(0.50)
Variable 2, interacted with high-stakes:	**-**	2.02	-4.79	-2.98	5.74	1.75	16.46
$\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{neg} x hs		(1.56)	(1.09)	(1.12)	(0.89)	(0.94)	(1.41)
		Panel D: Minimal Marginal Student, Restricted Sample					
Interaction of urgency with positive distance	**0 or ?**	1.79	-1.53	-0.50	2.71	1.74	3.68
from marginal student: $\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{pos}		(0.45)	(0.36)	(0.44)	(0.30)	(0.33)	(0.53)
Interaction of urgency with negative distance	**0 or ?**	-7.39	5.93	-1.30	-2.53	-1.52	-10.21
from marginal student: $\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{neg}		(1.69)	(1.11)	(1.16)	(0.85)	(0.96)	(1.53)
Variable 1, interacted with high-stakes:	**-**	0.86	4.28	1.09	-1.34	1.90	-1.80
$\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{pos} x hs		(0.52)	(0.49)	(0.52)	(0.39)	(0.38)	(0.57)
Variable 2, interacted with high-stakes:	**-**	5.23	-1.52	-0.65	9.06	3.97	21.19
$\vert\pi_i-\pi_m\vert(1-\pi_m)$ x δ_{neg} x hs		(1.81)	(1.40)	(1.39)	(1.15)	(1.11)	(1.60)

Appendix A: NWEA Score Alignment Studies

Arizona

Cronin, J. (2003). Aligning the NWEA RIT Scale with Arizona's Instrument to Measure Standards (AIMS). Lake Oswego, OR: Northwest Evaluation Association Research Report 2003.3. http://www.nwea.org/assets/research/state/Arizona%20complete%20report.pdf

Cronin, J. and Bowe, B. (2003). A Study of the Ongoing Alignment of the NWEA RIT Scale with the Arizona Instrument to Measure Standards (AIMS). Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Arizona%20Final%20Draft.pdf

Cronin, J. and Dahlin, M. (2007). A Study of the Alignment of the NWEA RIT Scale with the Arizona Assessment System. Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Arizona%20Alignment%20Report%204.18.07.pdf

Colorado

Cronin, J. (2003). Aligning the NWEA RIT Scale with the Colorado Student Assessment Program (CSAP) Tests. Lake Oswego, OR: Northwest Evaluation Association Research Report 2003.2. http://www.nwea.org/assets/research/state/Colorado%20complete%20report.pdf

Bowe, B. and Cronin, J. (2006). A Study of the Ongoing Alignment of the NWEA RIT Scale with the Colorado Student Assessment Program (CSAP). Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Colorado%20study%20document%20revised.pdf

Cronin, J. (2007). A Study of the Alignment of the NWEA RIT Scale with the Colorado Assessment System. Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Colorado%20Alignment%20Report%204.18.07.pdf

Idaho

NWEA test is the high-stakes assessment during the period under study.

Indiana

Cronin, J. (2003). Aligning the NWEA RIT Scale with the Indiana Statewide Testing for Educational Progress Plus (ISTEP+). Lake Oswego, OR: Northwest Evaluation Association Research Report 2003.3. http://www.nwea.org/assets/research/state/Indiana%20complete%20report.pdf

Cronin, J. and Bowe, B. (2005). A Study of the Ongoing Alignment of the NWEA RIT Scale with the Indiana Statewide Testing for Educational Progress Plus (ISTEP+). Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Indiana%202005.pdf

Cronin, J. (2007). A Study of the Alignment of the NWEA RIT Scale with the Indiana Assessment System. Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Indiana%20Alignment%20Report%205.21.07.pdf

Michigan

Bowe, B. (2006). Aligning the NWEA RIT Scale with the Michigan Educational Assessment Program. Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Michigan%20Study%20documentv5.pdf

Cronin, J. (2007). A Study of the Alignment of the NWEA RIT Scale with the Michigan Assessment System. Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Michigan%20Alignment%20Report%205.22.07.pdf

Minnesota

Cronin, J. (2004). Adjustments made to the Results of the NWEA RIT Scale Minnesota Comprehensive Assessment Alignment Study. Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Michigan%20Study%20documentv5.pdf

Cronin, J. (2007). A Study of the Alignment of the NWEA RIT Scale with the Minnesota Assessment System. Lake Oswego, OR: Northwest Evaluation Association Research Report. http://www.nwea.org/assets/research/state/Minnesota%20Alignment%20Report%204.18.07.pdf

Wisconsin

Cronin, J. (2004). Aligning the NWEA RIT Scale with the Wisconsin Knowledge and Concepts Exams. Lake Oswego, OR: Northwest Evaluation Association Research Report.

http://www.nwea.org/assets/research/state/Wisconsin%20executive%20summary.pdf

Adkins, D. (2007). A Study of the Alignment of the NWEA RIT Scale with the Wisconsin Assessment System. Lake Oswego, OR: Northwest Evaluation Association Research Report.
http://www.nwea.org/assets/research/state/Wisconsin%20Alignment%20Report%205.22.07.pdf

CALDER