

**Accountability Incentives:
Do Failing Schools Practice Educational Triage?**

Matthew G. Springer¹
matthew.g.springer@vanderbilt.edu
Department of Leadership, Policy, and Organizations
Vanderbilt University's Peabody College

November 2007

Abstract: A growing body of journalistic accounts offers anecdotal evidence that No Child Left Behind (NCLB) creates incentives for schools to practice “educational triage” whereby a disproportionate amount of resources are allocated to marginally performing students to the detriment of traditionally high- and low-performing students. This study uses a unique panel of student level data from one western state to estimate an education production function in which fall-to-spring test score gains are a function of schools’ incentives to focus instruction on marginally performing students. No evidence of failing schools systematically targeting students near the state-defined performance threshold was found. Evidence suggests that failing schools elevated learning opportunities for low-performing students, and that greater-than-expected performance by traditionally low-performing students did not occur at the expense of high-performing students in failing schools. It is important to note that reported findings have a particular meaning in this context because this study cannot know how these students would have fared in the complete absence of NCLB.

¹Matthew G. Springer is a research assistant professor of public policy and education at Vanderbilt University's Peabody College and director of the federally-funded National Center on Performance Incentives. He wishes to express his appreciation to Dale Ballou, James W. Guthrie, Warren Langevin, Derek Neal, Michael J. Podgursky, Randall Reback, Diane Schanzenbach, Jeffrey A. Springer, Martin West, and Kenneth Wong for their helpful comments and insights in developing this work. He is also grateful to the Northwest Evaluation Association for the data used in the analysis. Any errors remain the sole responsibility of the author.

1. Introduction

No Child Left Behind Act of 2001 (NCLB) is the reauthorization of the *Elementary and Secondary Education Act of 1965* (ESEA). The central purpose of NCLB is that all public school students, and defined student subgroups thereof, should reach academic “proficiency” by the 2013–2014 academic year. NCLB monitors progress toward meeting academic “proficiency” through Adequate Yearly Progress (AYP) calculations, a series of minimum competency performance thresholds that must be met by schools and school districts to avoid sanctions of increasing severity. In theory, NCLB’s threat of sanctions increases incentives for schools and school districts to elevate learning opportunities for traditionally low-performing students and student subgroups.

However, two recent qualitative studies report that systemic incentives in minimum competency accountability programs lead to schools practicing “educational triage” whereby a disproportionate amount of resources is allocated to those students who are particularly important to a school’s accountability rating.² Booher-Jennings’ (2005) case study of a single urban elementary school in Texas concluded that, during NCLB’s inaugural year, educators responded to high-stake accountability system by focusing on students close to the performance threshold to the detriment of peers. In a pre-NCLB case study of two low-performing and two high-performing schools in Illinois, Diamond and Spillane (2004) studied the consequences of high-stake testing for schools operating under Chicago Public Schools’ accountability system. They too concluded that low-performing schools expended disproportionate effort on marginally performing students as means to avoid sanctions, often to the detriment of the lowest performing students.

Research on the impact of high-stake accountability programs on the distribution of student test score gains is needed for several reasons. First, even though considerable interest and

² As noted by Booher-Jennings (2005), the term educational triage comes from Gillborn and Youdell’s (2000) study of educational inequality at two British secondary schools during the 1995-1997 school years.

controversy surround achievement tradeoffs in high-stake accountability programs, surprisingly little empirical research has addressed the issue directly.³ With the exception of three studies (Ballou, Liu, and Rolle, 2005; Booher-Jennings, 2005; Neal and Schanzenbach, 2007), previous research on the effect of accountability programs on the distribution of student achievement has focused exclusively on pre-NCLB accountability programs (Deere and Strayer, 2001; Holmes, 2003; Diamond and Spillane, 2004; Chakrabarti, 2006; Reback, 2006).

Second, the limited evidence on the impact of accountability programs on the distribution of student test scores is decidedly mixed. Studies of state accountability programs in North Carolina and Texas suggested the presence of achievement tradeoffs working to the detriment of traditionally high-performing students (Deere and Strayer, 2001; Holmes, 2003; Booher-Jennings, 2005; Reback, 2006), whereas evidence from Florida and Tennessee indicated that elevated achievement of low-performing students did not come at the expense of estimated gains of high-performing peers (Ballou, Liu, and Rolle, 2005; Chakrabarti, 2006). Most recently, Neal and Schanzenbach (2007) found that marginally performing students in Chicago scored higher following the introduction of high-stake accountability reforms, and that test score gains of traditionally low- and high-performing students tended not to improve.

Finally, distributional effects in the context of NCLB's long-term agenda may be different from those elicited in the Florida, North Carolina, and Texas studies. NCLB's long-term goal of all students and student subgroups reaching academic "proficiency" by the 2013–2014 school year is

³ To date, most scholarly research has examined the association between (1) accountability programs and mean achievement growth (Grissmer and Flanigan, 1998; Carnoy and Loeb, 2002; Figlio and Rouse, 2005; Hanushek and Raymond, 2005; Peterson and West, 2006); (2) accountability programs and school, classroom, or teacher behavior and practices indirectly linked to student-level achievement data (Ladd and Zelli, 2002; Diamond and Spillane, 2004; Koretz et al, 1996); or (3) accountability programs and system gaming (Cullen and Reback, 2006; Figlio and Getzler, 2002; Jacob and Levitt, 2003). This deficiency is largely because of data limitations and to the fact that minimum competency accountability programs were not a "potent" feature of state public education systems until federal enactment of NCLB in 2002 (Stecher, 2002). Recently, Nicotera, Teasley, and Berends (2006) examined achievement growth of students who transfer from low-performing schools identified as not meeting AYP.

different than that set by Florida's Opportunity Scholarship Program, North Carolina's ABC Accountability Plan, and Texas's Accountability Program. As such, it is important to investigate the influence of NCLB-induced state accountability programs on the distribution of student test score gains.

Present dialog on "educational triage" is largely informed by a growing body of journalistic accounts. In an effort to fill this knowledge gap, this study estimates an education production function in which test score gains are a function of the incentives schools have to focus instruction on marginally performing students. Longitudinal, student-level test score data from one western state were obtained from the Northwest Evaluation Association's (NWEA) Growth Research Database (GRD). NWEA data were then used to answer the following research questions:

1. Do schools target students expected to fail the spring high-stake assessment?
2. Have schools responded to failing AYP by raising the test score gains of marginally performing students relative to the test score gains of other students?
3. Have schools responded to failing AYP by raising the test score gains of marginally performing students at the expense of other students?

Results contradict what one would expect to find if failing schools were practicing "educational triage." Evidence suggests that failing schools elevated learning opportunities for low-performing students, and did not target marginally performing students to the detriment of traditionally low- and high-performing students. After removing the influence of mean-reverting measurement error, test score gains of nonfailing students enrolled in failing schools met or exceeded what one would expect if schools were concentrating resources on students near the performance threshold of the state accountability system. In total, this pattern of results tend to suggest that failing schools in the state were able to benefit low-performing students in ways that

were consistent with having operational slack, and that the threat of sanctions by the accountability system may stimulate greater operational efficiency within failing schools.

Although this study offers the first statewide analysis of whether schools are responding to NCLB by engaging in “educational triage”, it is important to acknowledge some limitations. Demographic characteristics of the state under study are homogenous and unrepresentative of other states and their respective school systems. NCLB provides states considerable autonomy in the design of their respective accountability programs. Thus, in responses to reform, interstate variations attributable to variation in the standards that each state requires for a student to be considered “proficient” can be expected. Distributional effects also may be different across time because NCLB’s long-term goal is for all students and student subgroups to reach academic “proficiency” by the 2013–2014 school year. Finally, data limitations preclude precise prediction of the impact of the state accountability system on the distribution of student test score gains in the absence of a counterfactual condition.

The subsequent study is divided into seven sections. Section 2 describes the state’s high-stake accountability program and presents selected sample statistics for schools that failed to meet AYP, schools that met AYP, and the universe of students tested over the three-year period under study. Section 3 offers a basic characterization of strategic resource allocation decision making in response to NCLB. Section 4 describes student- and school-level indicators used to measure a school’s short-run incentive to target resources. Section 5 outlines the data source, data development, and identification strategy. Section 6 presents results from empirical analyses of the state’s high-stake accountability program on the distribution of student test score gains. Finally, Section 7 discusses major findings and limitations of this research, provides recommendations regarding future directions for research, and considers general implications for U.S. education policy.

2. State Accountability Plan⁴

Designed to meet federal guidelines and regulations associated with NCLB, State Accountability Plan (SAP) was approved by the United States Department of Education (USDE) in 2003. SAP's genesis was a set of content and achievement standards established in the mid-1990s, and subsequently revisited in early 2000, by the state legislature, state board of education, and a citizen's commission. SAP holds all public schools in the state accountable on the following three dimensions: (1) proficiency scores in math; (2) proficiency scores in reading; and (3) minimum participation rates in testing.

Under the system, schools must meet or exceed math and reading performance thresholds and minimum participation rates in and across ten student subgroups to avoid sanctions.⁵ Table 1 displays SAP's annual proficiency standards and the intermediate incremental percent increase in the aggregate required of schools to reach 100% proficiency by the 2013–2014 school year. These annual proficiency standards for schools are calculated using a uniform averaging procedure across grade levels in a school. Schools are not held accountable for student and student subgroup performance in individual grades. Furthermore, schools must test a minimum of 95% of students in each subgroup to avoid sanctions.

Insert Table 1 Here

State Accountability Plan measures student content knowledge and skills using an Internet-enabled, high-stake testing system developed by NWEA. The high-stake assessment evaluates students in reading, math, and language arts and is scored on a single cross-grade and equal-interval

⁴ This is a pseudonym for the state's accountability plan.

⁵ The ten student subgroups include the following: (1) All Students; (2) African American; (3) Asian; (4) American Indian/Alaskan Native; (5) Hispanic; (6) White; (7) Hawaiian/Other Pacific Islander; (8) Students with Disabilities; (9) Limited English Proficient; and (10) Economically Disadvantaged. SAP's minimum "n" for accountability purposes is 34 students.

scale ranging from 150 to 300 using a single parameter Rasch Unit (RIU) methodology (Cronin et al., 2005; Kingsbury, 2003; Northwest Evaluation Association, 2004). It is administered to all public school students in grades 2 through 10 in the fall and spring. Scores are equated across grades, not subjects.

Spring assessment results are compared to grade-specific benchmarks to gauge whether a student, student subgroup, and school met SAP's minimum proficiency standards. Table 2 delineates the state's assessment proficiency scores by content, grade, and performance thresholds, including both the test score required of students to be considered proficient and the percentage of proficient students required of a school to avoid SAP sanctions. In math, the statewide percentage of proficient and advanced students in grades three through eight has ranged from a low of 53% for eighth graders in 2003 to a high of 90% for fourth graders in 2005. When pooling across grades and years, approximately 75% of students scored proficient or better on the spring math test.

Insert Table 2 Here

Table 3 presents select sample statistics for schools that met AYP, schools that failed AYP, and all schools in the sample. These statistics indicate that approximately 84% of students are White, 12% are Hispanic/Latinos, and the remaining 4% are Black/African American, Asian/Pacific Islander, or American Indian/Native Alaskan. Approximately 41% of students were identified as economically disadvantaged during the spring semester as defined by free and reduced price lunch status. Schools that did not meet AYP have a higher percentage of Hispanic/Latino students (18% vs. 10%), free and reduced price lunch eligible students (48% vs. 39%), and fewer White, non-Hispanic students (77% vs. 86%). Average fall-to-spring gain scores were greater in schools that met AYP (9.25 vs. 8.28). Demographic differences among failing and nonfailing schools are unlikely to bias findings because a school fixed effects estimator is used to compare schools to themselves over time.

Insert Table 3 Here

Generalizing any analysis from this state is problematic because the state's education system is demographically unrepresentative of other states and their respective educational systems. The state is disproportionately white and rural and has much smaller than typical schools and districts. Even so, the state's testing regime is unique in that it is administered to public school students twice per year, allowing for construction of a fall-to-spring gain score for each individual student in the 2002–2003, 2003–2004, and 2004–2005 school years. Using a fall-to-spring gain score as the dependent variable is advantageous because spring-to-spring gain scores are subject to the confounding influence of the summer months. This means that a school's effect on any gain (or potential loss) in a student's test score cannot be disentangled easily from how much gain (or loss) occurred as a result of summer activities.⁶

3. A framework for understanding strategic decision making of resource allocation

Investigating strategic decision making of resource allocation by schools seeking to avoid SAP sanctions requires a measure that captures systematic shifts in intraschool resource distribution.⁷ Recognizing that no formal accounting system tracks allocation of resources at the student or classroom level, distributional inequities in student test score gains among students above and below the SAP-defined proficiency standard are used to infer a reprioritization of intraschool resources. “Educational triage” as a specific manifestation of SAP-induced resource allocation decision making is detected if a greater-than-expected increase in the test score gains of marginally performing students occurs in tandem with a less-than-expected increase in the test score gains of

⁶ Alexander, Entwisle, and Olson (2001) examine summer learning.

⁷ This tradeoff framework builds upon the work of Deere and Strayer (2001), Holmes (2003), Reback (2006), and Chakrabarti (2006).

traditionally high- and/or low-performing students. The following vignette from Booher-Jennings (2006) provides a fine illustration of “educational triage” in action:

“Take out your classes’ latest benchmark scores,’ the consultant told them, ‘and divide your students into three groups. Color the ‘safe cases,’ or kids who will definitely pass, green. Now, here’s the most important part: identify the kids who are ‘suitable cases for treatment.’ Those are the ones who can pass with a little extra help. Color them yellow. Then, color the kids who have no chance of passing this year and the kids that don’t count – the ‘hopeless cases’ – red. You should focus your attention on the yellow kids, the bubble kids. They’ll give you the biggest return on your investment.”

Figure 1 presents a hypothetical example of “educational triage” in the context of SAP. The y-axis is the amount of growth in a student’s test score from the fall to spring test administration. The x-axis identifies the distance a student is from the SAP-defined performance threshold. The vertical line labeled W^* that runs through the apex of the inverted “V” is the performance threshold a student needs to cross to be considered proficient. The farther students reside to the left of W^* in the test score distribution the more likely they are to fail the spring high-stake assessment. The farther students reside to the right of W^* in the test score distribution the more likely they are to pass the spring high-stake assessment. The inverted “V” depicts the basic linear functional form if a school is targeting resources systematically to students particularly important to its accountability rating.

Under this premise, this study predicts a shift of resources away from traditionally high- and low-performing students and toward marginally performing students, under the assumption that devoting resources to students well below or well above the passing threshold holds superfluous marginal effects for a school’s likelihood of meeting AYP. Accordingly, the frequency and magnitude of an achievement tradeoff are functions both of (1) whether a school failed to meet SAP-defined proficiency standard in the prior year; and (2) how much effort might need to be expended by a school on a particular student, or group of students, to reach proficiency, as defined

and measured by whether a student is expected to fail a high-stake test and by how far that student is from reaching the SAP-defined proficiency standard in a particular year. Both of these measures are discussed in greater detail in the next section.

Insert Figure 1 Here

This characterization of SAP-induced resource allocation decision making among failing schools assumes that schools (1) respond strategically to high-stake accountability programs; (2) are both well-informed and well-intentioned in their resource allocation decision making; and (3) face resource constraints. Research built around psychology's classical behaviorist models of motivation suggests external stimuli such as incentives are highly effective in motivating behavior, even when employees are strongly intrinsic (Locke and Latham, 1990; Mohrman and Lawler, 1996). Moreover, a growing body of educational accountability research indicates that schools act strategically in response to high-stake accountability programs (e.g., see Figlio and Getzler, 2002; Figlio, 2005; Jacob and Levitt, 2003; Chakrabarti, 2006; Cullen and Reback, 2006).

NWEA furnishes classroom teachers and building principals with student-specific proficiency reports within three days of the fall test administration, or thereabouts. These reports identify (1) students' performance against proficiency growth targets in prior years; (2) the quartile distribution of classrooms against norm groups; and (3) students' projected individual performance on spring administration of the test (NWEA, 2006). As such, this western state's teachers and principals are well-positioned both to make knowledge-based resource allocation decisions in the short run as a means to avoid SAP sanctions and to track progress of current and incoming students over time.

Reliance on distributional inequities to infer strategic resource allocation decision making in response to SAP assumes implicitly that schools are resource constrained. In effect, resource constraint implies that elevating the performance of marginally performing students necessitates that

schools give up something, somewhere else. If the state's public schools do not operate consistently within a zero-sum view of resources, then examining the distribution of student test score gains to infer strategic resource allocation decision making may be problematic. Theoretically, schools unconstrained by resources are capable of elevating outcomes of marginally performing students without diverting attention and resources away from other students. On the other hand, schools simply may become more efficient; that is, they respond to SAP's systemic incentives by doing more with the same level and distribution of resources available in previous years.

Because of data limitations, this study is unable to precisely predict the impact of SAP on the distribution of student test score gains in the absence of a counterfactual condition. Because the state under study implemented SAP in response to NCLB, and the policy applies to all traditional public and public charter schools in the state, there is no readily available comparison group with which to examine how SAP affects the relative performance of low- and high- achieving students. Consequently, relative performance identified in this study may be the result of customary school behavior irrespective of SAP's threat of sanctions should schools fail to meet AYP.

4. Measuring a School's Short-Run Incentive to Target Resources

A school's short-run incentive to target instruction is a central issue in testing for the impact of failing AYP on the distribution of student test score gains. Although Deere and Strayer (2001) conducted the first work on high-stake accountability programs and achievement tradeoffs, Holmes (2003) was the first researcher to construct an indicator for measuring a teacher's and school's incentive to target instruction in response to the threat of sanctions. Holmes' approach, however, imposed a fixed bandwidth value that restricted the estimated degree of resource targeting at the school-level, thereby likely biasing estimates of whether schools target resources to marginal students.

Reback (2006) developed a more complex technique for estimating a school's incentive to improve the expected performance of certain students. He estimated the marginal effect of a hypothetical improvement in the expected performance of a particular student on the probability that a school would obtain a certain accountability rating. Reback's indicator permitted him to test whether students earned higher-than-expected test score gains when schools were subject to stronger short-run incentives to focus effort and resources on marginally performing students.

This study takes a different approach by constructing an indicator to measure whether a particular student is likely to be targeted. The indicator uses the distance of a student's score from the SAP-defined passing threshold after accounting for a student's expected fall-to-spring gain score in a particular grade and year. The variables, $GAP.PASS_{ij(t-1)}$ and $GAP.FAIL_{ij(t-1)}$, are expressed as the distance of a student's test score from the SAP-defined performance threshold. Use of two variables to express whether a particular student is likely to be targeted is preferred over a single indicator that would impose a strong linear assumption on the model specification; that being, increased gains of a student who starts five points below the cut-off equals the amount by which the gain will diminish when a student starts five points above the cut-off.

The mean values of $GAP.PASS_{ij(t-1)}$ and $GAP.FAIL_{ij(t-1)}$ are 12.84 (sd = 8.31) and 9.46 (sd = 8.66), respectively. As noted in Table 3, the mean value of $GAP.FAIL_{ij(t-1)}$ is slightly larger in schools that failed to meet AYP. The mean values of $GAP.PASS_{ij(t-1)}$ are nearly equivalent between failing and nonfailing schools. $GAP.PASS_{ij(t-1)}$ and $GAP.FAIL_{ij(t-1)}$ were also modeled using a quadratic and cubic function. The quadratic and cubic regression equations were not statistically significant.

Binary indicator variable, $FAIL.AYP_{j(t-2)}$, denotes whether a school met SAP's minimum proficiency standard and captures a school's incentive to target instruction. $FAIL.AYP_{j(t-2)}$

replicates SAP's proficiency standard calculations by determining whether each public school, and all defined subgroups therein, met the SAP-defined performance threshold in each of the three years represented in the panel following administration of the math test in the spring. This approach helps account for the potential presence of resource targeting both in "Safe Harbor" provision schools that must act strategically in the short run to increase the percentage of students scoring at or above proficiency by 10% year-over-year and in schools where the size of certain failing subgroups fluctuates around minimum "n-size" of 34, thereby making that subgroup's continued exemption from AYP designation an uncertainty and preserving the school's incentive to elevate the performance of students at or near SAP's performance threshold. Four different versions of $FAIL.AYP_{j(t-2)}$ are modeled using minimum "n" thresholds of 34, 32, 28, and zero students; each construction yielded similar results in cross-wise comparisons.

The author also explored if the severity of sanctions a school faced impacted a school's strategic resource allocation decision making. No evidence was found of the severity of sanctions impacting strategic resource allocation decision making. The least squares mean difference using the Tukey-Kramer adjustment for multiple comparisons between schools failing to meet SAP proficiency standard for no years, one year, and two consecutive years indicated that the mean difference between schools failing to meet SAP proficiency standard for one and two years, respectively, was not statistically different. However, this particular investigation was limited given, first, that the data panel encompassed only two years of potential responses by schools to SAP and, second, that SAP's sanctions truly only make an effect after a school fails proficiency for three consecutive years. Additionally, the state under study did not have prior experience with high-stake accountability policy.

5. Data Source, Data Development, and Identification Strategy

Data Source

Primary data for this study come from NWEA's GRD, a data system that collects longitudinal student-level achievement results from approximately 2,200 school districts in 45 states. Starting with the 2002–2003 school year, NWEA's GRD contains test score information for over 90% of this state's students in mathematics, reading, and language arts. GRD assigns each student a unique identifier as long as the student is enrolled in the state's public school system. This tracking mechanism offers access to multiple observations on each individual student in the sample and opportunity to construct a panel data set. GRD also contains demographic and other relevant information on students and schools, including race/ethnicity, gender, free and reduced price lunch status, grade, school type, and school size.

Data Development

Development of the data included selecting eligible schools and students. GRD contains demographic information and test scores for students in traditional public schools, nontraditional public schools (e.g., charter or virtual/on-line schools), and private schools (e.g., Catholic, other religious, and nonreligious schools). Eligible cases were restricted to students enrolled in traditional public schools or public charter schools; private schools are not held accountable under SAP.

There were 51 schools in the state for which the total tested student population was less than SAP's minimum prescribed "n" of 34. Under SAP, because of their small sample size, AYP for these schools was calculated using three years of achievement data to obtain a more consistent and reliable determination. These unusually small schools were removed from the sample, resulting in exclusion of a total of 727, 730, and 708 student observations in the 2002–2003, 2003–2004 and 2004–2005 school years, respectively.

Reconfigured schools were also deleted from the sample. SAP does not hold reconfigured schools accountable until three full years of achievement data are collected. Across all three years,

there were a total of 10 reconfigurations, or eight school consolidations and two school deconsolidations. Eliminating these schools reduced the total number of student observations by 1,921, or approximately 1% of all traditional public school and public charter school students from grades three through eight.

Eligible student observations were restricted to students enrolled in grades three through eight. The dependent variable relies on a fall-to-spring student gain score; therefore, students without both fall and spring RIT scores in a given school year were excluded from analysis. This restriction was also placed on data because the state’s administrative code requires that AYP calculations only include students continuously enrolled in the same public school from the end of the first eight weeks of the school year through the spring test administration period. Although coding procedures employed are not based on between-school student attendance patterns, as defined by law, this restriction is believed to produce a closer approximation to actual AYP calculations. In total, these restrictions eliminated less than 17,500 student observations, or an average of 5.5% of student observations per year. Results were not sensitive to the inclusion of nonfull year students in school-level proficiency calculations.

Identification Strategy

A general linear model is specified as follows. Let $\Delta Y_{ijt} = Y_{ijt} - Y_{ij(t-1)}$ be the math gain score for student i in school j from fall to spring administration of the high-stake test, where t denotes spring administration of the test. Then

$$\begin{aligned} \Delta Y_{ijt} = Y_{ijt} - Y_{ij(t-1)} = & \alpha_0 + \alpha_1 \text{FAIL.AYP}_{j(t-2)} + \alpha_2 \text{HISPANIC}_{ijt} + \alpha_3 \text{WHITE}_{ijt} \\ & + \alpha_4 \text{GAP.FAIL}_{ij(t-1)} + \alpha_5 \text{GAP.PASS}_{ij(t-1)} + \alpha_6 \text{FAIL.AYP}_{j(t-2)} \times \text{GAP.FAIL}_{ij(t-1)} \\ & + \alpha_7 \text{FAIL.AYP}_{j(t-2)} \times \text{GAP.PASS}_{ij(t-1)} + \mathbf{X}_{jt} + e_{ijt}. \end{aligned} \quad (1)$$

A general linear model is the preferred specification for estimating schools’ responses to SAP. In a single stage, this model can isolate whether: (1) a school failed to meet SAP-defined

proficiency standard based on their previous year's performance; (2) a school targeted resources toward particular students depending upon how far above or below that student's expected score was from the SAP performance threshold; and (3) practical differences exist between students enrolled in schools that failed AYP or met AYP.

A general linear model can also handle more than one fixed effect estimator in a single model. A school fixed effects estimator (μ_j) can be added to control for inter-campus differences that are time invariant and likely correlated with student test score gains. Suppose, for instance, that student test score gains in failing schools were, on average, smaller than nonfailing schools. If this is true, omitting school effects would yield biased estimates of the parameters of interest. Additionally, a school effects model is a within-group estimator that attributes only within-school movement to coefficients on parameters of interest. As such, parameter estimates measure how test score gains change within failing schools as the school environment changes.⁸

Other confounding influences have also been taken into account for the proposed general linear model specifications through use of a grade by year interaction ($\eta_g \times \gamma_t$). Whereas NWEA reports attest to the stability, reliability, and validity of the state's assessment (Kingsbury, 2003; Northwest Evaluation Association, 2004), research on non-NWEA designed assessments indicate that test difficulty may change from one year to the next (Yen, 1986; Braun, 1988, Koretz and Barron, 1998; Ballou, 2002). If test difficulty varies from year to year, and/or varies for different student population from year to year, estimates of tradeoffs will be biased toward zero. Because each grade and year represents a new test, the grade-by-year interaction is a suitable control for testing effects.

⁸ Inclusion of school fixed effects also takes into consideration any systematic variation across districts and communities (Hanushek, Kain, Rivkin, and Branch, 2004).

There is also a chance that interdependence of the errors within the same grade at the same school may lead to a Type 1 error. Two strategies are employed to control for clustering: school by grade fixed effects ($\mu_s \times \eta_g$) and school by grade by year fixed effects ($\mu_s \times \eta_g \times \gamma_t$). School by grade fixed effects essentially treats each grade within a school as a separate school, thus taking into consideration any observables common to all students in the same grade within a school. School by grade by year fixed effects control for similarity of outcomes among students in the same grade within a school by removing all the variation in cases over time.

Finally, as intraschool peer composition is likely to change from year to year, thereby rendering a cohort fixed effect an inadequate strategy to control for peer composition, a vector of variables is utilized to control for peer effects (X_{jt}). These variables include poverty status, as measured by eligibility for free or reduced price lunch, and school-level student background characteristics, including percent of students by race/ethnicity. The preferred model specification reported throughout the results section is specified as follows:

$$\begin{aligned} \Delta Y_{ijt} = Y_{ijt} - Y_{ij(t-1)} = & \alpha_0 + \alpha_1 \text{FAIL} .\text{AYP}_{j(t-2)} + \alpha_2 \text{HISPANIC}_{ijt} + \alpha_3 \text{WHITE}_{ijt} \\ & + \alpha_4 \text{GAP} .\text{FAIL}_{ij(t-1)} + \alpha_5 \text{GAP} .\text{PASS}_{ij(t-1)} + \alpha_6 \text{FAIL} .\text{AYP}_{j(t-2)} \times \text{GAP} .\text{FAIL}_{ij(t-1)} \\ & + \alpha_7 \text{FAIL} .\text{AYP}_{j(t-2)} \times \text{GAP} .\text{PASS}_{ij(t-1)} + X_{jt} + \eta_g \times \gamma_t + \mu_s \times \eta_g + e_{ijt} . \end{aligned} \quad (2)$$

6. Results

6.1 Do schools target students expected to fail?

Table 4 displays results from three model specifications. They estimate differences in test score gains for students across the distribution of test score gains at the same time allowing for different outcomes when a student is expected to pass or fail the spring high-stake test. All model specifications include school-level controls for race/ethnicity and free and reduced price lunch status as well as some combination of grade, school, and/or year fixed effects.

Insert Table 4 Here

The estimate on $GAP.PASS_{ij(t-1)}$ indicates the farther a nonfailing student is from the SAP-defined performance threshold the smaller the predicted fall-to-spring gain score, whereas the estimate on $GAP.FAIL_{ij(t-1)}$ indicates that the farther a failing student is from the performance threshold the greater the predicted gain. To put the magnitude of the estimated growth differentials between failing and nonfailing students in perspective, estimates generated by Model 4.2 approximate that the average failing student gains are more than one-third of a standard deviation greater than the average nonfailing student in a similarly situated school. Setting aside for the moment the fact that baseline estimates reported in Table 4 are likely sensitive to specification bias, it is worth noting that the slope coefficient on $GAP.FAIL_{ij(t-1)}$ runs counter to the expected slope coefficient under the “educational triage” hypothesis.

6.2 Have schools responded to failing AYP by raising the test score gains of marginally performing students at the expense of other students?

Table 5 displays estimates from an education production function in which fall-to-spring gain scores are a function of the incentives schools have to focus instruction on marginally performing students. Models 5.1 – 5.3 predict whether: (1) a school that failed AYP in the previous year has responded to SAP’s threat of sanctions by raising the test score gains of marginally performing students; (2) a student that is expected to pass or fail the spring test taking into consideration how far that student is from reaching the SAP-defined performance threshold; and (3) practical differences exist between students enrolled in schools that failed AYP and met AYP.

Insert Table 5 Here

The preferred model specification, labeled as model 5.2, includes school-by-grade and grade-by-year fixed effects and controls for intraschool peer composition. The sign on the α_1 coefficient is positive, and the reported value of .23 (se = .06) in model 5.2 is statistically significant at the $\alpha < .01$ level. This estimate suggests that the average student enrolled in a failing school gained, on average, .23 points more than the average student enrolled in a nonfailing school.

The α_6 coefficient, identified by $FAIL.AYP_{ij(t-2)} \times GAP.FAIL_{ij(t-1)}$, is the failing student conditioning effect. The sign on α_6 is positive and the value of .08 (se = .0047) is statistically significant at the $\alpha < .01$ level, demonstrating that, on average, this state's schools are responding to failing AYP by elevating the performance of failing students relative both to nonfailing students in failing schools and to failing students in nonfailing schools. In substantive terms, the estimates on the failing student conditioning effect indicate that an average failing student enrolled in a failing school gains 3.93 points more than an average passing student also enrolled in a failing school and 1.22 points greater than a failing student in a nonfailing public school. These results contradict what one would expect to find if schools were practicing “educational triage”.

The α_7 coefficient, identified by $FAIL.AYP_{ij(t-2)} \times GAP.PASS_{ij(t-1)}$, is the passing student conditioning effect. The sign associated with the α_7 coefficient is positive, and the reported value of .0116 is statistically significant at the $\alpha < .01$ level. This estimate indicates that the average student expected to pass the spring test who attended a school that failed to meet AYP the previous year will have a larger fall-to-spring gain score than a student that was located at the same point on the $GAP.PASS_{ij(t-1)}$ distribution but was enrolled in a nonfailing school.

However, as displayed in Fig. 2, the expected test score gains of passing students in failing schools are below average and decrease with higher $GAP.PASS_{ij(t-1)}$ values. Although Fig. 2 shows that the appearance of elevated learning opportunities for low-performing students comes at the

expense of traditionally high-performing students; it is possible that mean reversion is contaminating schools' true response to failing to meet AYP (Chay, McEwan, and Urquiola, 2005). The next section addresses the potential threat of mean-reverting bias using a technique for standardizing individual gain scores defined in Hanushek et al. (2005). This transformation of the dependent variable also permits checking the sensitivity of other estimates reported in Table 5.

Insert Figure 2 Here

6.3 Do estimated effects persist after standardization of students' fall-to-spring gain scores?

Mean-reverting bias is addressed by calculating a standardized fall-to-spring gain score for each student based on a comparison of a student's nominal gain and the average gain in test scores for all students by year and grade and then re-estimating all model specifications using the standardized gain score as the dependent variable. To implement, the state's initial distribution of students' fall test score is divided into 20 equal intervals for each year and grade combination, and the mean and standard deviation score gain is computed for all students starting in a particular interval for each of those combinations. A student's gain score is standardized by taking the difference between that student's nominal gain and the mean gain of all students in the interval over the standard deviation of all student gains in the interval. Gains in each interval are distributed with a mean of zero and standard deviation of one.

Table 6 displays estimates from a series of model specifications using the standardized gain score as the dependent variable. Most results are qualitatively similar to those parameters reported in the previous section. Both the signs on the coefficients and the attendant significance levels have remained stable. Once again, for both failing and nonfailing schools, the linear trends associated with $GAP.PASS_{ij(t-1)}$ and $GAP.FAIL_{ij(t-1)}$ stand in sharp contrast to the expected shape if "educational triage" in response to SAP was taking place in failing schools.

There are some noteworthy changes in the predicted fall-to-spring gain scores as illustrated in Fig. 3. Most striking is the fact that the gains of failing students in failing schools are above average, whereas schools are responding to failing AYP by raising the test score gains of failing students in failing schools. No longer do gains of failing students in failing schools appear to be occurring at the expense of nonfailing students.

Insert Figure 3 Here

7. Conclusion

This study analyzed longitudinal, student-level test score data from one western state to answer the following research questions:

1. Do schools target students who are expected to fail the spring high-stake assessment?
2. Have schools responded to failing AYP by raising the test score gains of marginally performing students relative to the test score gains of other students?
3. Have schools responded to failing AYP by raising the test score gains of marginally performing students at the expense of other students?

Results contradict what one would expect to find if failing schools were practicing “educational triage” in response to SAP. The analyses presented provide empirical evidence that the state’s public schools responded to failing AYP by raising the performance of students expected to fail the spring high-stake assessment. The further a failing student is from the SAP performance threshold, the greater is their predicted fall-to-spring test score gains. The results also provide evidence of the state’s public schools responding to failing AYP by raising the test score gains of failing students relative to that of nonfailing students.

Most estimates reported were robust across model specifications when the dependent variable was standardized to account for the potentially confounding influence of mean-reverting measurement error. The most striking change was the fact that the gains of failing students in failing schools are above average, although schools are responding to failing AYP by raising the test score gains of failing students in failing schools. This suggests that the state's public schools responded to failing AYP by raising the test score gains of failing students in failing schools without sacrificing the test score gains of other students in failing schools. It is also interesting to note that the average failing student in a failing school gained more than the average nonfailing student enrolled in a similarly-situated failing school and that the average fall-to-spring gain score for a failing student in a failing school was significantly larger than the average failing student in a school that met AYP. In summary, this pattern of results tends to suggest that failing schools in the state may be deriving positive utility from operational slack, and that the threat of sanctions of increasing severity stimulated greater operational efficiency within these schools.

Means by which schools become more efficient remain a salient question and important direction for future inquiry. On the one hand, schools unconstrained by resources are capable of elevating outcomes of low-performing students without diverting attention and resources away from other students. On the other hand, schools simply may become more efficient, in that they respond to systemic incentives by doing more with the same level and distribution of resources available in previous years. Schools may become more efficient by substituting resources across outcomes or engaging in system-gaming or other opportunistic behavior. Recognizing that SAP does not hold schools accountable across the spectrum of activities found in traditional public schooling, schools may focus additional resources on high-stake tests and subjects (e.g., math, reading, and language arts) to the detriment of low-stake activities (e.g., science, social studies, art, and physical education).

Furthermore, schools may become more efficient by tracking students by ability. A principle assumption in the “educational triage” hypothesis is that curriculum and instruction will focus on the least rigorous portion of content tested and required to a student to meet performance thresholds as the marginal gain in the passing rate from teaching more advanced material will be small. If schools are not restricted to offering the same instruction to all students in the same classroom, then tradeoffs may not be evident when measured by the distribution of student test score gains. This solution assumes that schools can easily track students, differentiate their curriculum, and/or create small group tutoring sessions, and do so without incurring additional cost. R

Research needs to continue to monitor potential achievement tradeoffs within the context of NCLB-induced state accountability programs to better understand if the law’s minimum competency standard produces perverse incentives and requires modification, perhaps to reward improvements across the entire achievement distribution. For example, responses by schools with historically low-performance relative to the average performance level in that state, and/or schools governed by a high-stake accountability program with particularly rigorous standards, might differentially influence school behavior when compared to a school with above average performance in a state that has relatively weak standards. Similarly, distributional effects may be different across time.

In conclusion, analyses from this state’s early experience with high-stake accountability policy provides clear evidence that the response by failing schools to SAP’s threat of sanctions has not increased incentives for schools to target marginally performing students, nor does it appear as though increased performance by traditionally low-performing students is occurring at the expense of traditionally high-performing students after standardizing the fall-to-spring gain score. This pattern of results tends to suggest that failing schools in the state were able to benefit low-performing students in ways that were consistent with having operational slack, and that the

accountability system's threat of sanctions of increasing severity may stimulate greater operational efficiency within failing schools. Nonetheless, one must remember that these findings have a particular meaning in this context because this study cannot know how these students would have fared in the complete absence of NCLB.

References

- Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, 23(2), 171–191.
- Ballou, D. (2002). Sizing up test scores. *Education Next*. Retrieved September 8, 2006, from <http://www.educationnext.org/20022/10.html>
- Ballou, D., Liu, K., & Rolle, E. (2005). *Response to No Child Left Behind among Tennessee schools*. Unpublished working paper, Peabody College of Vanderbilt University.
- Booher-Jennings, J. (2005). Below the bubble: “Educational triage” and the Texas accountability system. *American Educational Research Journal*, 42(2), 231–268.
- Booher-Jennings, J. (2006). Rationing education in an era of accountability. *Phi Delta Kappa International*. Accessed from http://pdkintl.org/kappan/k_v87/k0606boo.htm
- Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13(1), 1–18.
- Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305–331.
- Chakrabarti, R. (2006). *Do public schools facing vouchers behave strategically? Evidence from Florida* (Working Paper Series). Nashville, TN: National Research and Development Center on School Choice.
- Chay, K., McEwan, P.J., & Urquiola, M. (2005). The central role of noise in evaluations that use test scores to rank schools. *American Economic Review*.
- Cullen, J., & Reback, R. (2006). *Tinkering towards accolades: School gaming under a performance accountability system*. Unpublished manuscript, University of Michigan.
- Deere, D., & Strayer, W. (2001). *Putting schools to the test: School accountability, incentives, and behavior*. Unpublished manuscript, Texas A&M University.
- Diamond, J. B., & Spillane, J. (2004). High-stake accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record*, 106(6), 1145–1176.
- Figlio, D. N. (2005). *Testing, crime and punishment* (NBER Working Papers No. 11194). Cambridge, MA: National Bureau of Economic Research. Retrieved September 8, 2006, from <http://www.nber.org/papers/w11194.pdf>
- Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability and disability: Gaming the system?* (NBER Working Papers No. 9307). Cambridge, MA: National Bureau of Economic Research. Retrieved September 8, 2006, <http://www.nber.org/papers/w9307.pdf>

- Grissmer, D. W., & Flanigan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas*. Washington, DC: National Education Goals Panel.
- Hanushek, E. A., Kain, J. F., Rivkin, S. G., & Branch, G. F. (2005). *Charter school quality and parental decision making with school choice* (NBER Working Papers No. 11252). Cambridge, MA: National Bureau of Economic Research. Retrieved September 8, 2006, from <http://www.nber.org/papers/w11252>
- Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297–327.
- 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-stake testing in the Chicago public schools. *Journal of Public Economics*, 89(5-6), 761–796.
- Jacob, B., & Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3), 843–877.
- Kingsbury, G. G. (2003, April 24). *A long-term study of the stability of item parameter estimates*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Koretz, D., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. Washington, DC: Educational Resources Information System.
- Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly*, 38(4), 494–529.
- Locke, E.A., & Latham, G.P. (1990). *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice Hall.
- Mohrman, S. A., & Lawler, E. E. (1996). Motivation for school reform. In S.A. Fuhrman & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 115–143). San Francisco: Jossey-Bass.
- Neal, D. and Schanzenbach, D.W. (2007). *Left Behind by Design: Proficiency Counts and Test-Based Accountability*. NBER Working Paper No. 13293. Cambridge: National Bureau for Economic Research.
- Nicotera, A., Teasley, B., and Berends, M. (2006). *Examination of Student Movement in the Context of the Federal Transfer Provision*. National Center on School Choice Working Paper.

- Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA achievement level test and measures of academic progress*. Portland, OR: Author.
- Northwest Evaluation Association (2006). *Assessment system classroom reports*. Portland, OR: Author.
- Peterson, P.E. and West, M.R. (2006). The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments. The Program on Education Policy and Governance at Harvard University Working Paper #05-01.
- Reback, R. (2006). *Teaching to the rating: School accountability and the distribution of student achievement*. Unpublished manuscript, Barnard College, Columbia University.
- Stecher, B. M. (2002). Consequences of large-scale, high-stake testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79-100). Santa Monica CA: RAND.
- Swanson, C., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis*, 24(1), 1–27.
- Yen, W. M. (1986). Normative growth expectations must be realistic: A response to Phillips and Clarizio. *Education Measurement: Issues and Practice*, 7, 16–30.

Table 1

Annual Proficiency Goals and Intermediate Incremental Increase Required to Reach 100% Proficiency by the 2012-13 School Year

Goals	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13
Reading											
Annual	--	3	3	3	3	3	3	4	4	4	4
Intermediate	66	66	72	72	78	78	84	84	92	92	100
Math											
Annual	--	4	5	5	5	5	5	5	5	5	5
Intermediate	51	51	60	60	70	70	80	80	90	90	100

Table 2

Approved High-Stake Assessment Proficiency Scores

	Grade								
	2	3	4	5	6	7	8	9	10
Reading									
Basic	174	185	192	198	203	207	210	213	216
Proficient	182	193	200	206	211	215	218	221	224
Advanced	193	204	211	217	222	226	229	232	235
Language Arts									
Basic	176	186	193	200	204	207	211	213	214
Proficient	184	194	201	208	212	215	219	221	222
Advanced	197	207	214	221	225	228	232	234	235
Math									
Basic	174	185	194	202	208	214	222	229	231
Proficient	185	196	205	213	219	225	233	240	242
Advanced	201	212	221	229	235	241	249	256	258

Note. Districts are permitted to impose more stringent proficiency standards. However, all districts are currently following these accountability standards.

Table 3

Select Sample Statistics

	Made AYP	Failed AYP	All
<u>Student Characteristics</u>			
<i>American Indian / Alaska Native</i>	0.0143 (0.1187)	0.0209 (0.1428)	0.0159 (0.1251)
<i>Asian / Pacific Islander</i>	0.0156 (0.1239)	0.0138 (0.1166)	0.0152 (0.1221)
<i>Black</i>	0.0093 (0.0958)	0.0072 (0.0844)	0.0088 (0.0931)
<i>Hispanic / Latino</i>	0.0978 (0.2969)	0.1829 (0.3865)	0.1188 (0.3235)
<i>White, Non-Hispanic</i>	0.8593 (0.3477)	0.7733 (0.4186)	0.8380 (0.3684)
<i>Other</i>	0.0037 (0.0607)	0.0020 (0.0443)	0.0033 (0.0571)
<i>Free and Reduced Price Lunch Status</i>	0.3870 (0.4870)	0.4780 (0.4995)	0.4095 (0.4917)
<u>Student Test Scores</u>			
<i>Growth in Test Score from Fall to Spring</i>	9.25 (7.5746)	8.28 (7.3800)	9.01 (7.5395)
<u>Indicator Variables</u>			
<i>Student Gap (Failing)</i>	8.76 (8.0183)	9.72 (8.1877)	9.46 (8.6648)
<i>Student Gap (Passing)</i>	12.88 (8.2530)	12.75 (8.4918)	12.84 (8.3085)

(Standard deviations in parentheses)

Table 4

Influence of Distance from State Accountability Program Performance Threshold on Student Test Score Gains

Regression Type	General Linear Model		
Dependent Variable	Student Math Gain (Fall-to-Spring)		
Indicator Variables	<i>Distance from SAP Performance Threshold (GAP.FAIL and GAP.PASS)</i>		
Fixed Effects (model)	School Grade*Year (4.1)	School*Grade Grade*Year (4.2)	School*Grade*Year (4.3)
<i>GAP.FAIL</i> (a_1)	0.2067 (0.0022)***	0.2064 (0.0022)***	0.2072 (0.0022)***
<i>GAP.PASS</i> (a_2)	-0.1069 (0.0014)***	-0.1068 (0.0021)***	-0.1052 (0.0014)***
R ²	0.1959	0.2152	0.2420
Observations	307758	307758	307758

* | ** | *** Estimates statistically significant from zero at the 10, 5, and 1 percent levels, respectively.

School-level controls for race/ethnicity and free and reduced price lunch status included.

Estimates are robust when suspicious values removed; suspicious values defined by values 1.5 IQR above Q3 and 1.5 IQR below Q1.

Table 5

Influence of Distance from State Accountability Program Performance Threshold on Student Test Score Gains

Regression Type	General Linear Model		
Dependent Variable	Student Math Gain (Fall-to-spring)		
Indicator Variables	School Failed to Make AYP (FAIL.AYP), Distance from Performance Threshold (GAP.FAIL and GAP.PASS)		
Fixed Effects (model)	School Grade x Year (5.1)	School x Grade Grade x Year (5.2)	School x Grade x Year (5.3)
<i>FAIL.AYP</i> (α_1)	0.2543 (0.0592)***	0.2378 (0.0590)***	...
<i>HISPANIC</i> (α_2)	-0.7405 (0.0679)***	-0.7505 (0.0673)***	-0.7373 (0.0668)***
<i>WHITE</i> (α_3)	0.3804 (0.0587)***	0.3769 (0.0582)***	0.3818 (0.0577)***
<i>GAP.FAIL</i> (α_4)	0.1872 (0.0002)***	0.1865 (0.0026)***	0.1859 (0.0026)***
<i>GAP.PASS</i> (α_5)	-0.1141 (0.0016)***	-0.1144 (0.0016)***	-0.1123 (0.0016)***
<i>FAIL.AYP</i> x <i>GAP.FAIL</i> (α_6)	0.0811 (0.0047)***	0.0825 (0.0047)***	0.0846 (0.0047)***
<i>FAIL.AYP</i> x <i>GAP.PASS</i> (α_7)	0.0107 (0.0047)***	0.0116 (0.0032)***	0.0092 (0.0032)***
R ²	0.1994	0.2187	0.2451
Observations	307758	307758	307758

*, **, *** Estimates statistically significant from zero at the 10%, 5%, and 1% levels, respectively.

School-level controls for race/ethnicity and free and reduced price lunch status included.

Estimates are robust when suspicious values removed; suspicious values defined by values 1.5 IQR above Q3 and 1.5 IQR below Q1.

Table 6

Influence of Distance from State Accountability Program Performance Threshold on Standardized Student Test Score Gains

Regression Type			
Dependent Variable	Standardized Student Math Gain (Fall-to-spring)		
Indicator Variables	School Failed to Make AYP (FAIL.AYP), Distance from Performance Threshold (GAP.FAIL and GAP.PASS)		
Fixed Effects (model)	School Grade x Year (6.1)	School x Grade Grade x Year (6.2)	School x Grade x Year (6.3)
<i>FAIL.AYP</i> (α_1)	0.0564 (0.0089)***	0.0534 (0.0090)***	...
<i>HISPANIC</i> (α_2)	-0.1087 (0.0103)***	-0.1099 (0.0102)***	-0.1083 (0.0101)***
<i>WHITE</i> (α_3)	0.0558 (0.0089)***	0.0557 (0.0088)***	0.0560 (0.0088)***
<i>GAP.FAIL</i> (α_4)	0.0034 (0.0004)***	0.0035 (0.0004)***	0.0036 (0.0004)***
<i>GAP.PASS</i> (α_5)	-0.0024 (0.0002)***	-0.0024 (0.0002)***	-0.0021 (0.0002)***
<i>FAIL.AYP</i> x <i>GAP.FAIL</i> (α_6)	0.0016 (0.0007)**	0.0018 (0.0007)***	0.0021 (0.0007)***
<i>FAIL.AYP</i> x <i>GAP.PASS</i> (α_7)	0.0014 (0.0005)***	0.0016 (0.0005)***	0.0012 (0.0005)***
R ²	0.0410	0.0644	0.0960
Observations	307758	307758	307758

*, **, *** Estimates statistically significant from zero at the 10%, 5%, and 1% levels, respectively.

School-level controls for race/ethnicity and free and reduced price lunch status included.

Estimates are robust when suspicious values removed; suspicious values defined by values 1.5 IQR above Q3 and 1.5 IQR below Q1.

Figure 1. Illustrative Example of Educational Triage in Response to No Child Left Behind

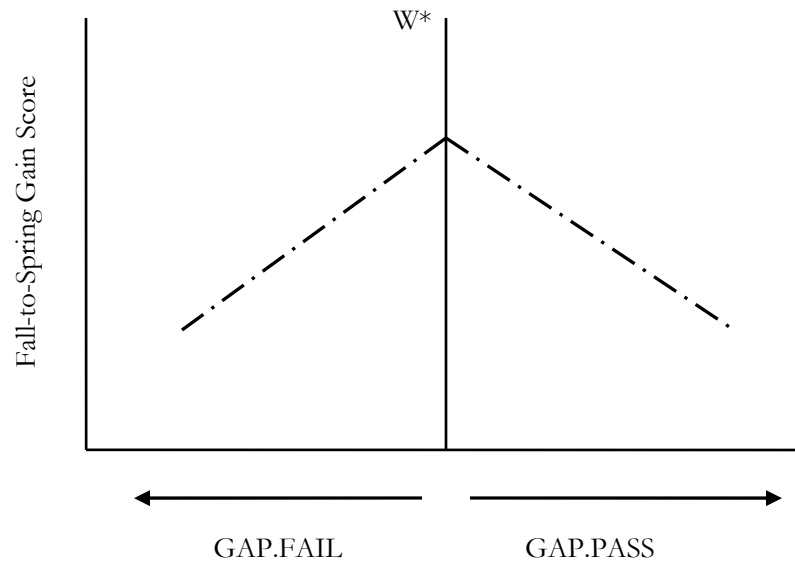


Figure 2

Influence of Distance from State Accountability Program Performance Threshold on Student Test Score Gains by School AYP Status

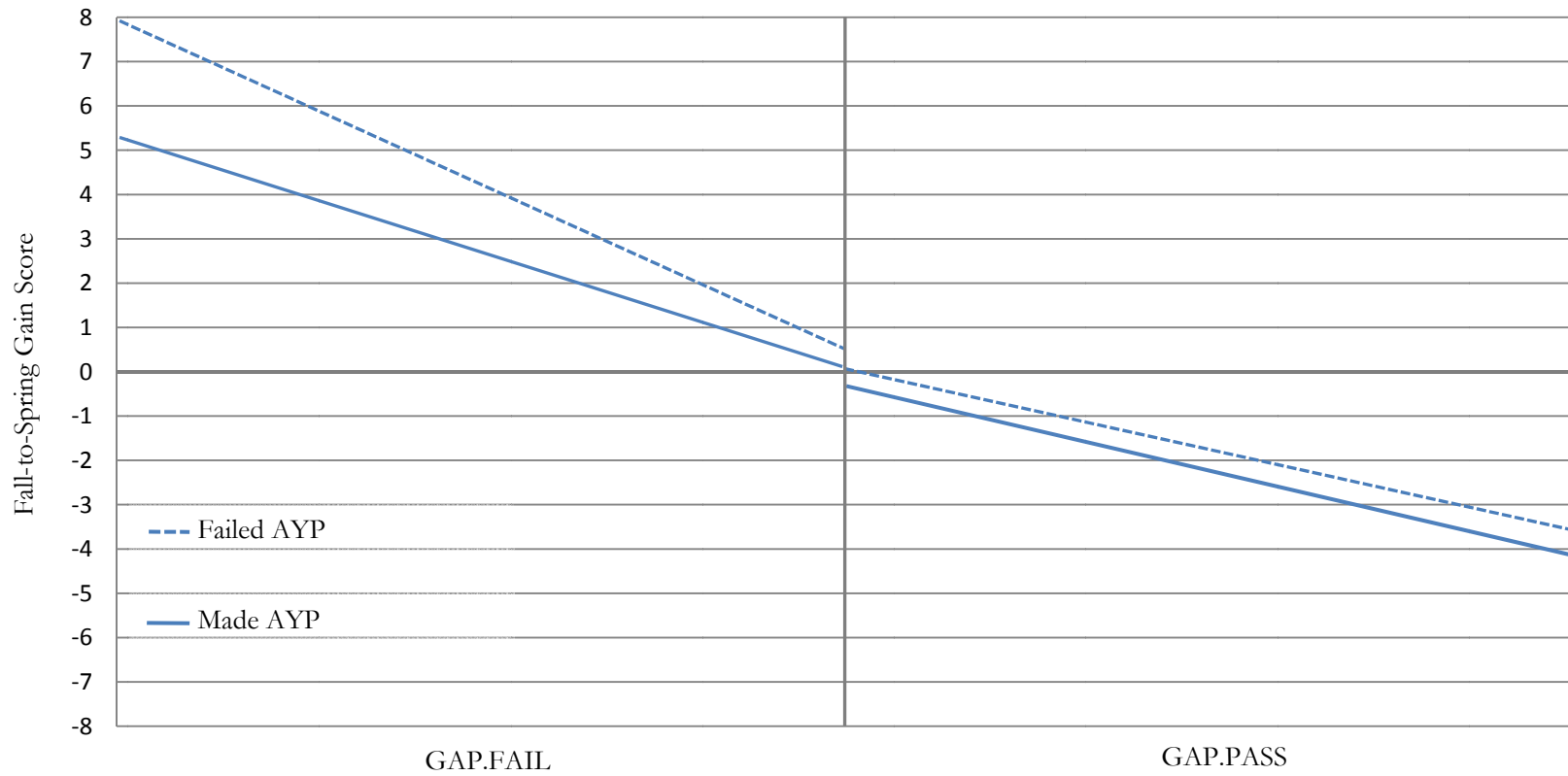
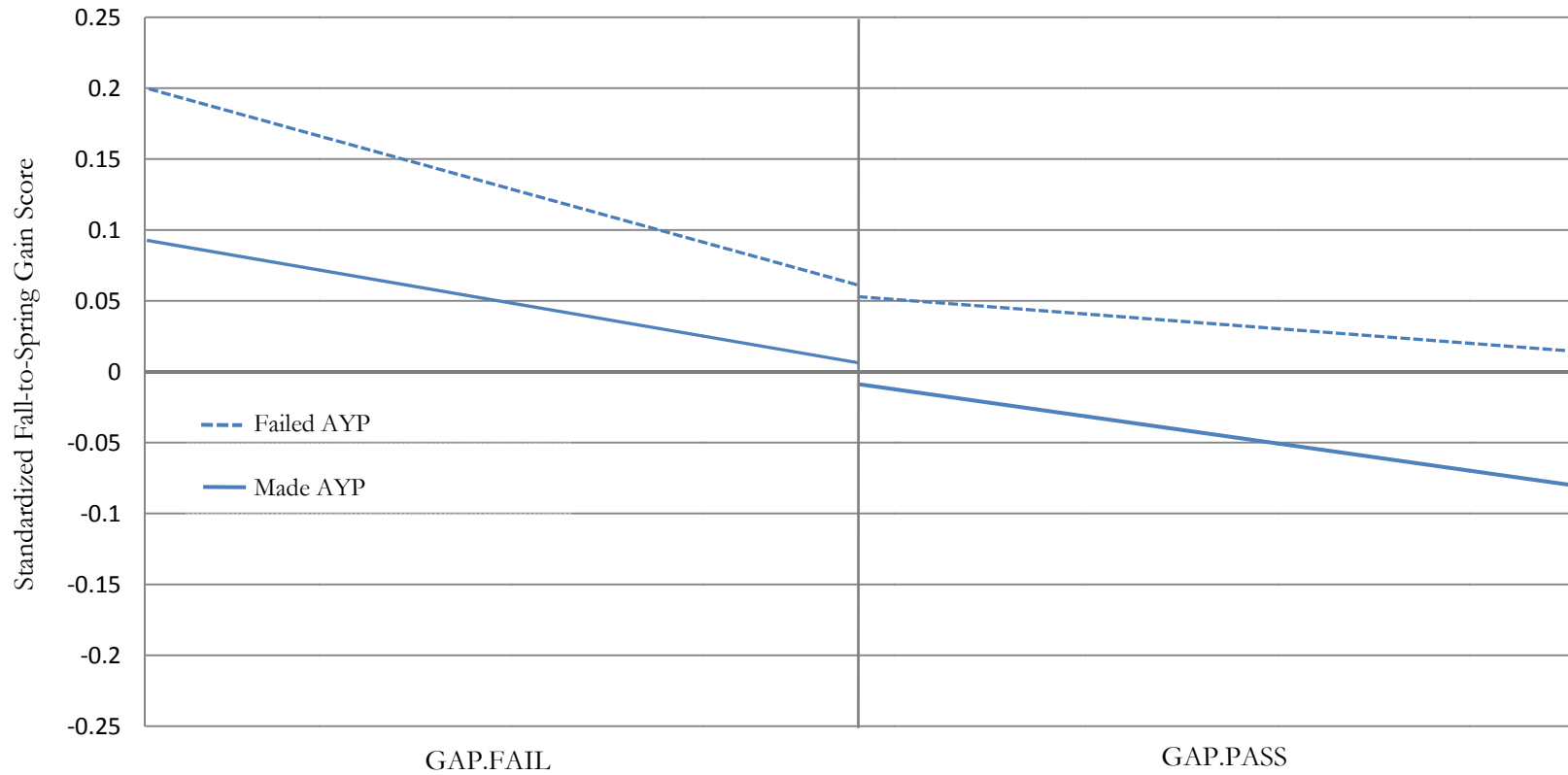


Figure 3*

Influence of Distance from State Accountability Program Performance Threshold on Student Test Score Gains by School AYP Status



*Erratum: Figure 3 differs from the abridged version published in the Winter issue of Education Next due to a reporting error by the author. All substantive conclusions remain the same.