



ILLUSTRATION: GREG CLARKE

Accountability Incentives

Do schools
practice

educational
triage?

“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.”

—Jennifer Booher-Jennings,
“Rationing Education in an Era of Accountability”
Phi Delta Kappan International (June 2006)

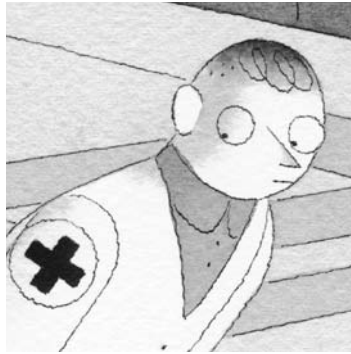
Increasingly frequent journalistic accounts report that schools are responding to No Child Left Behind (NCLB) by engaging in what has come to be known as “educational triage.” Although these accounts rely almost entirely on anecdotal evidence, the prospect is of real concern. The NCLB accountability system divides schools into those in which a sufficient number of students score at the proficient level or above on state tests to meet Adequate Yearly Progress (AYP) benchmarks

BY MATTHEW G. SPRINGER

(“make AYP”) and those that fail to make AYP. The system gives no credit to schools for moving students closer to proficiency or for advancing already-proficient students. If schools intent on meeting minimum competency benchmarks practice educational triage, they dedicate a disproportionate amount of their limited resources to “bubble kids,” students who might otherwise perform just below the proficiency threshold. While these marginally performing students are likely to benefit from increased attention, reallocation of instructional attention leads to a tradeoff whereby the achievement gains of the marginally performing students come at the expense of both the lowest- and highest-performing students.

With congressional proceedings on NCLB’s reauthorization under way, the time is opportune to take a hard look at educational triage claims. If the current law’s minimum competency standard produces gains among students near the proficiency threshold but disadvantages others, the rules of the accountability system need to be modified, perhaps to reward improvements across the entire achievement distribution.

To search for evidence of educational triage, I analyzed three years of test-score and other data on 300,000 students in public schools in a western state. I found none. I concluded that these schools were not responding to NCLB by trading off achievement among students with different baseline levels. Rather, they were successful at raising the performance of students who were otherwise at risk of failing the state test without sacrificing the performance of lower- and higher-performing students. Even in failing schools, students above the proficiency threshold made gains that were greater than one would expect if schools were concentrating resources on students near the threshold. When academic achievement is measured with test-score performance



When academic achievement is measured with test-score performance, the much-politicized argument that NCLB compromises the educational needs and opportunities of high-performing, academically accelerated students holds no water.

in this state, the much-politicized argument that NCLB compromises the educational needs and opportunities of high-performing, academically accelerated students holds no water.

The State’s Accountability Program

The U.S. Department of Education in 2003 approved the state’s accountability plan, which was designed to meet federal guidelines and regulations associated with NCLB. The plan requires all public schools in the state to meet proficiency standards in math and reading for all students and for each of 10 student subgroups, and to test a minimum of 95 percent of students in each subgroup to avoid sanctions. The accountability program measures students’ content knowledge and skills using an Internet-enabled testing system developed by the Northwest Evaluation Association (NWEA), a national nonprofit organization that provides assessment products and related services to school districts. NWEA compares spring assessment results to grade-specific benchmarks to gauge

whether individual students, subgroups of students, and schools meet the state’s proficiency standards.

The particular demographic characteristics of this state limit the generalizability of my findings to other, more demographically diverse states. The state is disproportionately white and rural, with much smaller than typical schools and districts. Approximately 83 percent of students are white, 12 percent are Hispanic, and the remaining 5 percent are black, Asian, Pacific Islander, American Indian, or Native Alaskan. Roughly 40 percent of students were identified as economically disadvantaged based on their eligibility for free and reduced-price lunch. Schools that did not make AYP have a higher percentage of Hispanic students than

schools that did (18 percent vs. 10 percent) and a higher percentage of students eligible for free and reduced-price lunch (48 percent vs. 39 percent). Despite the disadvantage of atypical demographics, this state offered the unique advantage of being able to measure achievement gains within the same school year. The state tests each student twice per year, permitting for measurement of individual students' fall-to-spring test-score gains.

Data

Data in this study are from the NWEA Growth Research Database. Starting with the 2002–03 school year, NWEA administered tests in mathematics, reading, and language arts to more than 90 percent of the state's students. NWEA furnished fall and spring test scores for the first three years after enactment of the state's accountability program (2002–03 through 2004–05 school years) for students in grades 3 through 8. My analysis focuses on math scores. The statewide percentage of students scoring in the proficient and advanced categories in math has ranged from a low of 53 percent for 8th graders in 2003 to a high of 90 percent for 4th graders in 2005.

The NWEA data set also provides demographic information about students, including the student's grade in school, gender, race, ethnicity, and eligibility for free or reduced-price lunch. School-level characteristics include school type and school size. Although NWEA gathers data for students in traditional public schools, charter schools, and private schools, I limited the study to students enrolled in traditional public schools or public charter schools because private schools are not included in the state's accountability program. I removed from the study sample very small schools, those with fewer than 34 students being tested, given their systematically different treatment under the state's accountability system.

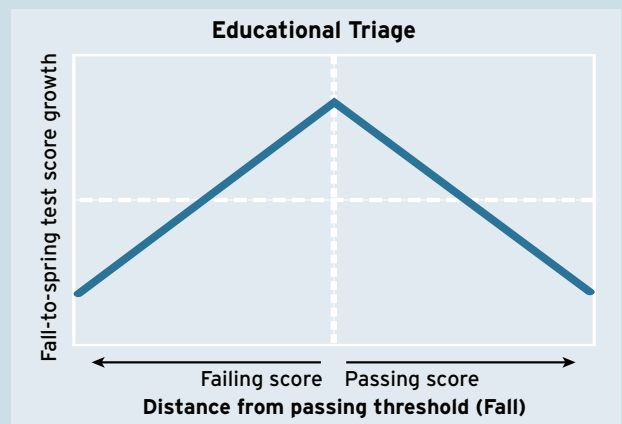
Identifying Educational Triage

My objective was to detect shifts in how schools committed resources to different students, resources such as textbooks and teachers, but also such inputs as teacher attention and choice of curriculum and instructional strategies. Obviously, no formal accounting system tracks the distribution of resources directed at individual students. So I turned to an indirect measure of resource allocation. I infer the priorities of administrators and teachers from educational outcomes, as measured by student performance on the state's math test. If there is a greater-than-expected increase in the achievement of students just below the state's proficiency standard, and this occurs in tandem with a less-than-expected increase in the achievement of high- and

Theory and Reality (Figure 1)

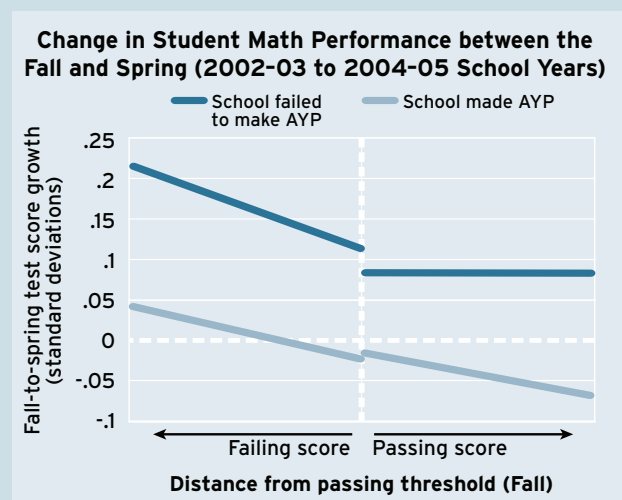
Media reports contend that *No Child Left Behind* has prompted schools to concentrate their resources on students whose test scores lie close to the passing threshold at the expense of the lowest- and highest-achieving students, which would yield test-score gains in the shape of the hypothetical plot below.

Figure 1a



My research shows that schools have not responded to NCLB's incentives with an educational triage strategy. Rather, Figure 1b shows that in both groups of schools the students who experienced the highest gains were those who had the lowest initial test scores.

Figure 1b



SOURCE: Authors' calculations

low-performing students, then I can conclude that educational triage has transpired.

Is it reasonable to expect administrators and teachers to be able to identify students who are likely to be on the cusp of the proficiency threshold at the spring test administration? When speaking of states that use NWEA assessments, the answer is yes. NWEA furnishes classroom teachers and building principals with proficiency reports for each student within days of the fall test administration. The reports include a projection of each student's performance on the spring test.

Consider a hypothetical example of the distribution of changes in test scores under educational triage. In Figure 1, the y-axis is the amount of growth in a student's test score from the fall to the spring test administration. The x-axis identifies a student's distance from the state-defined proficiency threshold. The vertical line in the middle of the graph is the threshold a student needs to cross to be considered proficient. The farther a student lies below the performance threshold in the fall, the more likely that student is to fail the spring assessment. The inverted "V" depicts the simplified pattern of gains one would expect to see if a school disproportionately targets resources, such as instructional time and teacher focus, to students particularly important to its accountability rating, that is, to students hovering around the state-defined proficiency threshold. If this practice were the case, the greatest fall-to-spring achievement gains would occur among students around the threshold, while other students would struggle to match expected test-score gains.

My basic strategy, then, was to compare fall-to-spring test-score changes among students who were expected to be either nearer or farther from the state-defined proficiency threshold following spring testing. A question of particular interest was



My objective was to detect shifts in how schools committed resources to different students, resources such as textbooks and teachers, but also such inputs as teacher attention and choice of curriculum and instructional strategies.

whether schools that failed to make AYP in the previous school year responded strongly to the incentive to target instruction to students at risk of falling just short of the proficiency threshold.

Statistical Controls

The first step in my statistical analysis was to rank all the students within the same grade and year by their performance on the fall exam. I then divided the students into 20 groups of equal size and calculated a standardized test-score change for each student that measured their respective performance relative to students within the same performance group. Use of standardized test scores helps address what statisticians call reversion to the mean. When one takes repeated measures of some event or behavior, such as test-score performance among students, the measurements at the low and high ends of the resultant distribution tend over time to converge toward the average value for the population under study. With respect to students and test scores, reversion to the mean suggests that students with scores in the upper or lower tail of the test-score distribution

are likely to perform closer to the average when tested more than once. This effect may mask a school's actual response to the threat of failing AYP by producing the illusion that schools are helping low-performing students while neglecting high-performing students. By comparing each student's gain to gains among students who performed at a similar level and would have experienced a similar, natural shift toward the average score, I can better separate legitimate test-score gains and losses from change associated with mean reversion.

I also made my best effort in the statistical analysis to isolate the change in test scores that could be attributed reasonably to the resources schools dedicate to teaching students.

Specifically, I separated out the effects on test-score gains of a student's race and ethnicity, as well as accounted for the influence of a student's peers, by evaluating the influence of demographic characteristics of the student body, including average income level and percentage of minority students. Given a large data set, I also was able to account for characteristics of schools that I could not directly measure but that might influence student achievement over the school year. Finally, I took precautions against shaping the results according to changes in test difficulty from year to year for each grade and for students in a given school.

Results

Despite many media claims of educational triage, I found no evidence of failing schools engaging in coordinated targeting of students near the state-defined proficiency threshold. In the state under study, public schools that had failed to make AYP focused instruction on the entire range of low-performing students in the subsequent school year, and did so without negative impact on high-performing students.

Figure 1b shows the changes in standardized test scores, across the full range of student performance, that can be attributed reasonably to teacher and school performance and to decisions about how the school allocates resources among students. In schools that failed to make AYP in the previous year, students who were expected to fall well below proficiency gained more than students nearest the proficiency threshold. The lowest-performing students gained about 0.20 standard deviations, roughly twice the improvement of those students whose expected gains were to leave them just below proficiency. Students expected to be proficient did not lose ground; the most advanced students performed comparably to other already-proficient students.



Despite many media claims of educational triage, I found no evidence of failing schools engaging in coordinated targeting of students near the state-defined proficiency threshold.

In schools that did make AYP, lower-performing students met expectations, with the largest of those gains coming from students expected to perform the weakest. It is interesting to note, in contrast, that higher-performing students in these schools lost ground from one year to the next. Remarkably, proficient students enrolled in failing schools experienced larger test-score gains than proficient students in non-failing schools. Remember that these patterns of gains and losses would look much different if schools did indeed engage in educational triage. Under educational triage, students near the proficiency threshold would attain the largest gains, while students dispersed away from this threshold and toward the tails of the achievement distribution would suffer diminished performance.

Conclusion

Although there is no evidence that schools in the study sample targeted resources to particular students, they may have allocated resources toward outcomes measured by the accountability system. For instance, schools may have taught to the tests in math and reading while neglecting science, social studies, the arts, and physical education. The apparent absence of educational triage in one state does not invalidate documented accounts of the practice in particular schools, nor encroach upon other arguments to modify NCLB's proficiency-based school rating system. Nonetheless, as the reauthorization debate continues, policymakers should take note that educational triage was not evident in the first statewide analysis of the issue.

Matthew G. Springer is 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.