

**How strong is the incentive for educators to game the Adequate Yearly
Progress Requirements of the No Child Left Behind Act?**

September, 2003

John Cronin

For questions call or e-mail:

John Cronin
(503) 624-1951
johnc@nwea.org

How strong is the incentive for educators to game the Adequate Yearly Progress Requirements of the No Child Left Behind Act?

John Cronin, Northwest Evaluation Association

September 2003

Passage and implementation of the *No Child Left Behind Act (NCLB)* has cemented two landmark changes in the criteria federal regulators apply to hold schools accountable for the performance of their students. The first, a change that actually began with adjustments to the accountability requirements of Title I, requires schools to demonstrate improvement by increasing the proportion of students who demonstrate proficiency on their state tests. The second is the NCLB requirement that schools set objectives for adequate yearly progress (AYP) with the goal that all students, in every meaningful demographic category, demonstrate proficiency on their state test by the end of the 2013-2014 school year. The intent of these changes, just as the title of the legislation suggests, is to ensure equity, meaning that all children, regardless of their circumstances, receive an education that assures they leave school with reasonable prospects for success.

The penalties for schools that fail to meet their AYP goals are potentially quite stiff. These penalties escalate in the following fashion:

- Schools that fail to meet their state defined adequate yearly progress targets for two consecutive school years will be identified as needing school improvement for the following year.
- Schools in need of improvement will need to develop a two-year plan to correct their performance. Every student in the school will be given the opportunity to transfer to a better public school within their district.
- Schools that fail to make adequate yearly progress for three consecutive years will be required to offer supplemental educational services to disadvantaged children in addition to the other consequences.
- Schools that fail to meet adequate yearly progress for four consecutive years in addition to the prior consequences, may be required to replace school staff relevant to the failure to make adequate yearly progress, implement new curriculum, extend the school year, and/or reduce the autonomy of school level managers.
- Schools that fail to meet adequate yearly progress for five consecutive years in addition to the prior consequences, may be required to reopen the school as a public charter school, replace all or most of the school staff, contract with a private school management company, and/or turn the operation of the school over to the state (U.S. Department of Education, n.d.).

Educators are likely to take these consequences seriously and I believe they will be motivated to try and meet the expectations of the law. Indeed the consequences are serious enough to raise concern that educators may adopt practices, in their efforts to comply with the law, that are counterproductive to the law's intent. The purpose of this paper is to raise the issue and provide evidence supporting this concern.

NCLB and the high jumping theory of accountability. Imagine a track coach is assigned a group of 100 athletes who will compete in the high jump. The high jumping ability of each of these athletes is widely varied but known. 45 of the athletes can currently jump proficiently, which is defined as high jumping five feet. The coach is told that she will receive \$1,000 for every additional student that jumps five feet after a four-week training period (if coaches only lived in such a world). The coach will also be docked \$1,000 for every student currently jumping five feet who fails to jump five feet on the test.

If coaches wanted to manage or “game” this system to maximize their income, what would they do? First they would find all the athletes who jump **near** five feet, either just above five feet or just below. Then they would coach those athletes feverishly to assure as many of them as possible jump above five feet at the end of the training period.

In the meantime, rational coaches would give the seven-foot high jumpers in this group four-weeks off because they are more likely to get injured practicing than they are to lose two feet on their jump. They would also spend little time with the three-foot high jumpers because their prospects for jumping five-feet within four weeks are poor.

In short, the best strategies for our coaches is to game the system by focusing their energies on the small number of high jumpers who have the best immediate prospects for improving to five feet. We can only hope is that coaches are more altruistic than rational, because only an altruist would put her full energies into ensuring all these athletes receive the coaching they deserve. What causes the potential problem is the degradation of a finely defined scale (feet and inches) capable of measuring small incremental improvements into a two-category scale capable of measuring improvement only when it crosses a single point of measurement.

The NCLB accountability system shares much in common with the scenario outlined above. Proficiency standards measure success in much the same way that success is measured in the high jump at a track meet. States establish a standard by setting a cut score or “bar” that students must exceed in order to get credit for being proficient. When applying the NCLB criteria, schools are rewarded for increasing the proportion of students over the bar. There is no partial credit for near misses, that is no distinction is made between students who barely nick the bar and those who crash through it. In other words, NCLB does not give schools credit for turning 3 foot academic jumpers into 4 foot 11 inch jumpers. AYP goals are only met when the school increases the number of non-proficient students who reach a particular academic threshold, proficiency, while maintaining those students already above the bar.

NCLB also places no demands on schools for improving the academic performance of students who have already reached the proficiency bar. Schools that turn six-foot academic jumpers into eight-foot leapers receive no credit in the accountability formulae for doing so (other than the self-satisfaction that comes from a job well done).

The accountability criteria, therefore, create a risk that the intention of No Child Left Behind may be subverted if well-meaning educators, fearing for their reputations and livelihoods, adopt practices that game the system by focusing their energies on the students who have the best immediate prospects for jumping the proficiency bar. While in the long run these gaming strategies are likely to be counterproductive because of the law’s requirement that all students reach proficiency by the 2013-14 school year, gaming strategies may nevertheless stall consequences and create a temporary safe-haven from the accountability penalties. There are many educators with long-term vision, and most educators want to be altruistic, but it might be

difficult for them to sustain good intentions in the face of immediate consequences that threaten their professional reputations and the survival of their workplace.

Linn (2003) argues that, while there are cases in which defining categories of performance are essential one should not consolidate test scales into two to five point categorical measures unless there is a compelling reason. All of us would cede, for example, that having pass/fail categories on a state driving test is useful.

But when one is attempting to determine whether improvement has occurred, categorical measures may disguise more than they reveal. That's why one never reads medical studies of weight loss treatments in which criterion used to measure the effectiveness of the treatment is a improvement in a categorical measure like "thin/not thin". Instead, medical researchers take advantage of the precise gradations available on a scale so they can assess the incremental effect of a treatment. After all, for those of us that are overweight, there is a health benefit in being 10 pounds lighter, even if we don't cross that magical threshold that is "thin".

He goes on to argue, that when one is assessing the capacity of school systems to improve student performance, many methods superior to the use of cut scores are available.

Methodology. Our was interested in trying to determine whether the Adequate Yearly Progress (AYP) expectations associated with the NCLB standards are structured to provide an incentive for educators to narrow their instruction to the population of students near the proficiency bar. In an effort to assess the potential risk associated with this issue, we analyzed reading results on the Colorado State Assessment Program (CSAP) for a group of 4,012 students from 8 Colorado school systems who completed this assessment as 4th graders in spring 2001 and again as 5th graders in spring 2002. Using a range of scores that varied from 10 to 90 points around the cut score, we determined the number and proportion of the student population that

changed their status around the CSAP's *proficient* cut score. We used the ranges to determine how closely the student starting scores were associated with most proficiency status changes. We then determined the proportion of status changers relative to the entire student population and the proportion of actual test content contained in that range of the CSAP scale. The scale difficulty of CSAP test content is public information that is available by downloading the appropriate item map from the Colorado Department of Education's Website (Colorado Department of Education, n.d.)

In addition, we designed a simple simulation that would allow us to estimate the costs and benefits of programs designed to produce status changes with this group of students. The simulation involved designing a hypothetical program that would improve the numbers of students achieving proficiency by a fixed proportion for a specified cost. We used the simulation to determine how much money it took to produce status changes for students near the proficiency bar as opposed to students who performed at greater distances from the proficiency bar.

Results. As we expected, a relatively small proportion of students (13.63%) actually changed status against the proficiency bar between 4th and 5th grade during the study period (see table 1). Although this data imply a likelihood that many students are not close enough to the proficiency standard to have a meaningful chance of changing status within a year, it is not sufficient evidence by itself to assess the size of that risk.

Table 1 – Students changing status against the proficiency cut score on the CSAP between 4th and 5th grade

Students who improved		Students who maintained status		Students who slipped	
N	%	N	%	N	%
256	6.38%	3465	86.36%	291	7.25%

The next step was to determine starting position of status changers relative to the proficiency bar. We found, once again as you might expect, that the vast majority of the status changes in 5th grade reading occurred among students whose 4th reading scores already put them near the proficiency bar. 52% of the students who improved and 50% of those who slipped were clustered in a 15 point range on either side of the proficiency standard. But only 18% of the total student sample scored within 15 points of the proficiency bar and just 11% of the test content is in that difficulty range. We further found that over 90% of the status changes occurred in a range within 45 points of the bar, but that range covered only about half of the sample population (52%) and 35% of the test content.

Table 2 – Analysis of 4th to 5th grade reading status changes on the Colorado State

Assessment Program reading assessment

Distance around proficient cut score	Student population in this range		Students dropping below proficient		Students improving above proficient		Proportion of CSAP content in this difficulty range
	N	% of total	N	% of total	N	% of total	
+/- 5 points	222	6	67	26	23	8	4%
+/- 15 points	708	18	134	52	146	50	11%
+/- 25 points	1181	29	189	72	209	72	20%
+/- 35 points	1650	41	220	86	247	85	27%
+/- 45 points	2072	52	237	93	261	90	35%
Totals	4012		256		291		

The results show very clearly that the vast majority of the status changes occur in range of scores that comprise a very narrow range of the test content and a relatively small proportion

of the tested student population. If one's motive is to game the system, there is good reason to believe that focusing on students within 15 points of the proficiency bar offers the highest potential payoff. As one moves beyond that range, the number of students who change status declines rapidly.

Even when we use a much broader range of performance, a range 45 points above and below the proficiency cut score that encompass over 90% of all status changes, we still cover just over half of the total student sample and just over one-third of the test content. Delivering instruction that is geared toward this group would greatly narrow the range of the curriculum and would not provide appropriate focus for a near majority of the students.

Results of the simulation. To provide some cost-benefit estimates, we designed a hypothetical reading intervention. We assumed that the intervention would cost \$100 for each student included in the program and would be required for all students in any defined score range around the proficiency bar. We also assumed the intervention would increase the number of students who improved in status in that defined score range by 60% and decrease the number of students who slipped below the proficient cut score by 35%. The assumptions we used are entirely arbitrary and intentionally generous, that is, we purposefully chose assumptions that would tend to overestimate the effectiveness of these interventions. The reasons for doing so will become more apparent when we review the data, but our motive was to employ a model that would understate the actual costs of an effective intervention. We did this to assure that the marginal cost of moving students who were longer distances from the proficiency bar would be estimated as conservatively as possible. So let's be clear that the model's purpose is simply to be illustrative, it is not intended to reflect the impact of any real educational intervention in use.

The simulation was intended to model the kind of investments schools might make if they were focused on improving the number of student achieving proficiency on their state test.

We were interested in seeing whether the payoff for the investment changed as more students, further from the proficiency cut score, were included.

Table 4 shows the results of the simulation. They show that our hypothetical program is quite efficient in improving results for students who are close to the proficiency bar. When applied to students within 15 points of a proficient score, we improved the number of proficient students by 135 (this was about 4% of the entire student population). The cost for each success was \$524 per success. In other words, investing in five students produced one additional favorable result.

Unfortunately as we moved farther from the proficiency bar, the productivity of the program declined dramatically. When we move 45 points from the proficiency bar the cost per success rises to \$866, meaning that we must invest in eight students to get one additional favorable result. More discouraging, however, is the marginal cost per success when we move from 35 points around the proficiency cut score to 45. By marginal cost per success, we mean the cost uniquely associated each success produced by the additional children served by the program in the range of 35 to 45 points from the proficiency cut score. When we make that decision, we include an additional 422 students in the program (2072 are included at 45 points, 1650 at 35), at an increased cost of over \$42,000. This large investment produced only 12 new proficient students. The benefit from increasing the range was very small relative to the cost. Each of those 12 students had a marginal cost of \$2,813, meaning that we paid \$2,813 to buy every additional success when we moved from 35 to 45 points above and below the proficiency bar. While we can buy success in this simulation, buying them becomes very expensive as the distance from the cut score increases.

Table 3 – Analysis of hypothetical program to increase reading proficiency rates on the CSAP

Distance around proficiency bar	Students in this range	Cost of program	Number of additional proficient students created by program	Cost per success	Marginal cost of program	Marginal cost per success
+/- 5	222	\$22,200	37	\$600		
+/- 15	708	\$70,800	135	\$524	\$48,600	\$496
+/- 25	1181	\$118,100	191	\$618	\$47,300	\$844
+/- 35	1650	\$165,000	225	\$733	\$46,900	\$1,379
+/- 45	2072	\$207,000	237	\$866	\$42,200	\$2,813

Let me emphasize that the problem is not with the concept of the program, the problem is with the mechanism used to evaluate its effectiveness. Every student who received the program was likely to receive some benefit; the problem is that the benefit only became visible when that student’s score improved enough to carry him or her over the proficiency bar. For students near the bar, an improvement of 5 points might be enough. But for students 40 points from the bar far more dramatic improvements, improvements as great as 39 points, had no effect on the proficiency rating.

Discussion. NCLB was clearly intended to drive change in schools and the Adequate Yearly Progress requirement is the operational expression of the change expected. In literal terms, the change expected is to make more non-proficient students proficient. The categorical scale is binary. Unless you turn zeroes into ones, improvement doesn’t count.

We were surprised at just how strong the incentive to game the Adequate Yearly Progress requirements of NCLB might be. In our study of Colorado data, more than half of the fourth grade student population had no realistic probability of crossing the proficiency threshold in a single year. The accountability system provides no rewards for working with these students or punishments for ignoring them. This may encourage a culture within schools that focuses

educators' short-term resources and energies on only the *nearly proficient*, those who have the most potential to provide an immediate payoff relative to the proficiency bar. Were this to happen, its effects would be counterproductive to the intent of the law. At the low end of the achievement scale, students who are not close enough to the proficiency bar to have immediate prospects for success would likely become children left behind, because there is no immediate reward defined in the accountability system for improving their performance unless they cross the *proficient* cut score. Of equal concern is that the program could also encourage no child left behind behavior, with students performing well above the proficient level suffering benign neglect because there is no incentive for educators to focus energy here. These incentives are likely to be strongest in those schools that find themselves on the NCLB accountability ladder rungs. The stronger the consequences, the greater the incentive desperate educators have to try and game the system.

I try to be optimistic that the commitment of educators to equity might be enough to prevent this kind of culture from emerging. But the results of the cost simulation have raised my concern that the forces driving administrators to produce immediately measurable proficiency rate improvements will overwhelm their commitment to attend to the needs of all students. The simulation demonstrates very dramatically that even when interventions are very effective, they are produce visibly efficient results on the binary proficient/not proficient scale only when applied to students already performing near the cut score. Resources are scarce and when one attempts to offer those interventions to students at larger distances from the cut score, it becomes extremely expensive to produce results that seem to justify the program (the word *seem* is used purposefully because while program probably produces good results for many children, just a few improve enough to trigger some change in their proficiency status). Once again the problem is not the intervention, the problem is with the binary scale that is used to judge improvement in

results. Given the rules of the game, it will take considerable intestinal fortitude for administrators and boards of education to resist the impulse to target their investments solutions to narrow ranges of students who have promise to show large, immediate payoffs against the proficiency bar, even though that behavior is clearly against the long term interests of most children and families, including large segments of the at-risk population that is NCLB's primary concern.

Despite the difficulty, we hope educators resist the urge to use the Adequate Yearly Progress criterion as their *only* criteria for program evaluation. We concede not only that the expectation that schools must meet AYP targets is legislatively mandated, but that it's also a good idea to help more students become proficient. It is simply unfortunate that we have chosen a scale with just a single mark to assess improvement. The result of this is that most students are not "in play" relative to the criteria for improvement. When exploring alternatives, districts should consider criteria that use more gradated measures to evaluate the improvement of all children, so that improvement in the near proficient is not rewarded if it is offset by slippage among other children. We believe methods that use the measurement and improvement of student growth to evaluate effectiveness meet these criteria and are good alternatives.

Let there be no misunderstanding. The *No Child Left Behind Act* embraces the right principle, all students are entitled to a quality education to assure their opportunity for future success. The Act is an ambitious undertaking and we have no doubt that there are elements of the Act that may evolve over time. As the implementation of the Act is refined, we hope regulators will consider alternatives to the current Adequate Yearly Progress model that may better support the intentions embodied in the law.

References

Colorado Department of Education. 2002 CSAP Grade 5 Reading Item Map. Retrieved March 26, 2003 from http://www.cde.state.co.us/cdeassess/asitemmap_02g5rd.htm

Linn, R. (2003, September 1). Performance standards: Utility for different uses of assessments. Education Policy Analysis Archives, 11 (31). Retrieved September 2, 2003 from <http://epaa.asu.edu/epaa/v11n31/>.

U.S. Department of Education. No Child Left Behind: Frequently Asked Questions for Children and Families. Retrieved March 26, 2003 from <http://www.nclb.gov/next/faqs/accountability.html>