PREFACE

By John Cronin and Michael Dahlin

Set standards. Test students. Sanction schools that don't measure up.

This is the NCLB formula for accountability, and it seems simple and compelling. Thanks to the passage of NCLB, we have proficiency standards and testing for all students in grades through 3 through 8, plus one high school grade. We have a no-excuses requirement that 100% of students achieve these proficiency standards, and a firm deadline for achieving them by 2014. There are also strict sanctions imposed on schools that do not meet the Annual Measured Objectives (AMOs), the proficiency rates required to stay on track for the 2014 deadline.

This is NCLB's sixth year of implementation. Large numbers of schools have been identified as underperforming and many of those schools have been sanctioned. As far back as 2005, over 10,000 schools across the United States had failed to make adequate yearly progress (or AYP) for two years in a row, thus putting them in "program improvement" (National Education Association 2006). And this year, California alone has 2,241 schools, about 22%, in program improvement (*San Francisco Chronicle* 2008). These numbers have increased dramatically in the past three years and the pace will likely accelerate as the Act's 2014 deadline draws closer.

We have standards, we have deadlines, and now we have a large round-up of K-12 suspects. Were we as cynical as Captain Renault from the film *Casablanca*, a round-up of the usual suspects would be all we needed to maintain an illusion of accountability, and it would little matter whether our suspect schools were really culprits in some crime against learning. To their credit, Former President Bush, Senator Ted Kennedy, Margaret Spellings and others who have driven support for this reform are not Captain Renault. The 2007 blueprint for reauthorizing NCLB stated the sentiments of those who support NCLB in plain, ambitious terms; its goal being to deliver "…steady academic gains until all students can read and do math at above grade level, closing for good the nation's achievement gap between disadvantaged and minority students and their peers (pg. 1)" (U.S. Department of Education 2007). The statement is quite sweeping; it does not suggest that the law's intent is merely limited to eliminating achievement gaps within a state. Rather, her statement refers to these as *national objectives*, which can be achieved only by wiping out differences in the performance of groups of students across states.

The strategy for achieving these objectives under NCLB might be best described as a "strict-loose" approach. NCLB's requirements for setting standards, testing students, and specifying deadlines are clearly strict. However, NCLB is loose in giving states wide latitude to determine both the difficulty of the proficiency standards (or cut scores) and the annual benchmarks that schools must achieve in order to make "Adequate Yearly Progress" (AYP) between now and 2014. Furthermore, NCLB allows states to set their own accounting rules for how students are categorized for evaluation. These rules include, among others, determining the minimum number of students in various groups that are separately accountable under NCLB, whether to apply a confidence interval (or margin of error) to proficiency results and, if a confidence interval is applied, its size.

If educational equity is the goal, then the strict-loose approach must achieve some degree of consistency in results for it to be reached. After all, if we accept that a school ruled "in need of improvement" in Florida, would not get that same label if it happened to be in New Jersey, California, or Illinois, then we are not truly eliminating achievement gaps – we are merely replacing gaps based on race or poverty with gaps based on geography.

If the goal of ensuring that all students achieve high standards is a **national objective**, then it is reasonable to ask whether this "strict-loose" approach is producing some modicum of consistency. Thus we, alongside our colleagues from Fordham, undertook a study to investigate two research questions.

1. Is there enough consistency among the various state proficiency standards and objectives to conclude that expectations across the states are similar? Does making AYP reflect equivalent achievement across the various states?

2. Do states apply the standards, timelines, and the various state rules in a manner that results in consistent judgments about schools across states? Would a school that meets Florida's expectations, in reality, also meet the expectations of New Jersey, or California, or Illinois?

To investigate these questions, we found a sample of 36 schools that reflect the diversity within the American educational system. Students in these schools took achievement tests that predict their proficiency status on 28 state tests with a high degree of accuracy. From this achievement information, estimates of the school's proficiency rates could be produced for each of the states studied. Thus, if a school achieved a proficiency rate of 70% in Illinois, it was possible to estimate what that proficiency rate would be if the school were located in Wisconsin, Minnesota, Indiana or other states. Once the proficiency rate is known, we can determine whether that proficiency rate would have been sufficient to reach the state's annual proficiency targets (AMOs) and whether the school would likely make AYP. Finally, it's possible to estimate whether a school that made AYP in Illinois would also it in other states.

With respect to the first question, the results of this study demonstrate that proficiency standards across states are vastly different. Case in point: one elementary school in our sample that achieved a predicted 80% proficiency rate under Wisconsin standards, achieved a 52% proficiency rate under Massachusetts standards, and only a 19% proficiency rate in California.

But standards are only one part of the equation. Each state also has AMOs, which are timetables of targets that require increasing proportions of students to achieve proficiency between now and 2014 (the NCLB deadline for achieving 100% proficiency). This study and others (e.g., Chudowsky and Chudowsky 2008) show that these timetables vary as much as the standards. But what is the result?

Consider Wolf Creek Elementary, a California school in our sample. Its students achieved a 54% reading proficiency rate and met their AMO. If Wolf Creek were relocated to South Carolina, we estimate their students would achieve about the same proficiency rate, 53%, since South Carolina's reading cut scores are roughly comparable to California's. But this rate of proficiency would fail to meet South Carolina's AMO (hence Wolf Creek fails to make AYP). In other words, we could have the same students produce the same proficiency rate in two states, and get two very different AYP outcomes. To make matters worse, consider what happens if Wolf Creek is relocated to New Jersey (whose state test is easier to pass). The school's estimated proficiency rate now rises to 80%, but in New Jersey, 80% is not high enough to meet the AMO. But had we dropped Wolf Creek into Michigan, whose state test is roughly equal in difficulty to New Jersey's, 80% proficiency would have been high enough to meet the AMO. So in Michigan, Wolf Creek Elementary would make AYP! Does this seem confusing? Take heart, because it is!

Is Wolf Creek on the path to "all students achieving high standards"? Who knows? How could one possibly tell? Performances that were a hit in Fresno bombed in Trenton. A school we might call a rose in Ann Arbor would not smell as sweet in Spartanburg…

Of course we recognize that the background and achievement of students vary from state to state. But there's no reason to believe that there's less need for math and reading competence in California than there is in South Carolina. And even if NCLB is successful in getting 100% of students to proficiency by 2014, all it will mean is that we have created an Orwellian system in which all students are proficient, but some are more proficient than others.

The second question we asked in this study was whether the state accountability systems created under NCLB make consistent judgments about schools across the various states. Whether sanctions achieve their desired end

depends on how effectively they are deployed. For the system to work, sanctions must target schools that are actually underperforming. Unfortunately, this study found little consistency across states in how NCLB is implemented, and rarely were adequately performing schools differentiated from underperforming ones.

Many years ago, one of the study authors taught high school. At this school, it was typical for nearly all the students enrolled in choir classes to receive "A" grades. One wouldn't know from the grading system that some of the students were highly-motivated, vocally gifted stars; that others were recreational singers of average talent; and that yet others took the class to get an easy grade. In this same school was another teacher who dedicated her efforts to finding *failure* somewhere inside every student. This teacher was legendary for giving pop quizzes, counting them triple if the students performed poorly, or discounting them by half if students performed too well.

In this study, state accountability systems fit both of these archetypes. Despite their large differences in achievement and growth, nearly all of the sample elementary schools made AYP under some accountability systems. Roughly one-third of the states have a combination of proficiency standards, AMOs, and rules that were met by the overall school populations in every single school within our sample. In such states, one could reasonably argue that students would be better served by higher proficiency standards, more aggressive targets, stricter rules, or perhaps all three.

On the other hand, many of the state accountability systems seemed designed to ensure school failure. Shockingly, the highest performing elementary school in our sample failed to make AYP in thirteen of the twenty-eight states studied, and the highest performing middle school failed in twenty-three states. Under the accountability systems in Massachusetts and Idaho, to cite two examples, every single middle school within our sample failed to make AYP.

The accounting rules used to define subgroups differ across states, and this one factor largely explains the indiscriminate effect of NCLB in certain states. NCLB requires that proficiency be achieved on the same timetable for all subgroups within a school, a goal meant to eliminate racial or income-based educational disparities. This "no-excuses" aspect is one of NCLB's most attractive features; it does not permit educators to write off the performance of minority or other traditionally disadvantaged groups. To the extent that NCLB has focused attention on improving the performance of these subgroups, it can be called a success.

While disaggregation is laudable, in practice the subgroup requirements cause the most diverse schools—particularly in states with more ambitious proficiency cut scores— to fail AYP. In about 30% (elementary sample) to 50% (middle school sample) of cases, low-income students failed to make their 2008 annual targets. In over one-half of the cases, one or more groups of minority students failed to make their AMO.

The results for limited English proficient (LEP) students and students with disabilities (those with Individualized Education Plans) were more depressing. These groups almost universally failed to meet AMOs regardless of the state they were in. In only 2% to 4% of the cases we evaluated did a group of LEP students actually achieve their AMO, even in states with relatively low proficiency cut scores and in states that "boost" their observed performance rates by reporting confidence intervals (or margins of error). Similarly, in only 2% to 6% of cases did students-with-disabilities (SWDs) achieve their targets. Ultimately even the highest performing schools—schools whose own LEP or SWD subgroups outperformed most or all of the same students in other schools—generally failed their AMO.[1]

Looking at the data, we would conclude that states have two possible strategies to cope with this problem, both of which are untenable. One is to avoid having subgroups. In general, schools within our sample that did not have LEP or SWD subgroups

[1] For reasons explained in the report, however, our estimates of SWD and LEP subgroup performance may be lower than is actually the case.

had a fighting chance of making AYP. So, if states were to set the minimum n size requirement so high that these subgroups escaped separate reporting, schools could up their AYP odds. The other solution would be to create proficiency standards so low that they could be met by 100% of students. Clearly, both of these solutions are at odds with NCLB's intended goals.

Simply put, it's a hard knocks life for states trying to implement NCLB in a manner consistent with its intent. When states adopt high standards, when they set AMOs on a rigorous timetable, when they establish rules about minimum subgroup sizes that are reasonable, then their schools are inevitably seen as failures under NCLB. For the schools in our sample, this was a plain, irrefutable fact. When confronted with these odds, educators in some of our better schools might be forgiven for feeling like new recruits in military basic training: They can make up their bunks immaculately, shine their boots to a high polish, learn all the drills to perfection, but still get 500 push-ups from the drill sergeant because he found a stray bristle on a toothbrush.

As currently implemented, NCLB is not a discriminating system. A tremendous amount of money and energy has been spent to create the impression that there is accountability, and there are large numbers of schools throughout the United States that are in some phase of sanctions. But the accountability is not coherent. We found states where most schools failed to make AYP and others where nearly every school made it. We found demonstrably good schools that failed to make AYP far too often, and some pretty mediocre ones that slide by in some states. Thus what seems like accountability is an illusion. Good schools get sanctioned, bad schools get off, and ultimately students get shafted, since maintaining this illusion has a cost. When good schools get sanctioned, resources are wasted and we risk causing quick-fix, panic driven, counterproductive change in schools that may ultimately hurt students. When bad schools get off, their students are denied opportunities (what we unfortunately now call "sanctions") that might lead to a better education, including the chance to attend a different school, or receive supplemental services, or simply obtain assurance that the workings of a perennially dysfunctional school will be addressed and corrected.

It 's long past time to dispel the accountability illusion.