DISCUSSION

NCLB was intended to ensure that all schools set high standards for reading and math, and to hold all students accountable to these standards, regardless of their ethnicity, income, or other differences. Unfortunately, the strategy chosen to implement these goals creates an illusion of accountability that will not get us to these results, in part because it was too lax in establishing guidelines around standards and rules and too inflexible in its requirements for outcomes.

NCLB has given states the discretion to establish proficiency cut scores, the required trajectory for improvement, minimum subgroup sizes, and confidence intervals. Our results show that the product of these differences bears no resemblance to a coherent system. Not only do the proficiency cut scores themselves vary greatly, but the variance in improvement trajectories, subgroup sizes, and policies for application of confidence intervals result in wildly different Adequate Yearly Progress results for the schools in our sample. It appears, then, that the federal government has implemented a system in which geography had as much to do with our schools' AYP status as their students' academic performance. In addition, it was sometimes impossible to distinguish between the high-performing and underperforming schools in our sample. We could argue that NCLB has been too lax in allowing this degree of discretion.

Conversely, the law requires 100% of students, including 100% of students in every subgroup, to achieve the states' proficiency standards by 2014. In the meantime, each and every subgroup is required to meet the Annual Measured Objectives that are set for schools each year. These subgroups include low-income students and ethnic minorities, but they also include subgroups whose members have documented academic challenges, such as Limited English Proficient students and Students with Disabilities students and SWDs. Although the sample schools in the study met proficiency goals for their overall student populations in the majority of cases, the performance of subgroups within the sample schools was far worse. All eligible minority subgroups within a school met their proficiency objectives in only 20% to 30% of cases. But eligible LEP and SWD populations fared even worse. Within the sample schools, these two groups met their proficiency objectives in just 3% to 6% of cases. This means that the relative difficulty of the cut scores and the AMOs are essentially irrelevant, because LEP and SWD subgroups failed even in states with low cut scores and AMOs. In this regard, we could argue that NCLB has been too strict.[1]

Of course the bottom line for schools is whether they ultimately make AYP. Applying these rules to the elementary sample, we found that AYP results differed dramatically across the states studied. The number of schools in the sample that made AYP varied from 1 in Massachusetts and Nevada to 17 in Wisconsin. Ultimately there was no consistency in the way elementary schools were judged, meaning that there is likely to be no consistency in the way sanctions are applied.

The results for the middle school sample were consistent but grim. In 5 states none of the schools in the sample met AYP; in 6 other states, only 1 school made AYP. In general, the higher rates of failure can be attributed to the fact that middle schools were accountable for more subgroups. In many cases, the failing subgroups were low-income students and ethnic minorities. But in almost all cases in which the school was accountable for a LEP or SWD subgroup, the school failed.

We could take this to mean that the AYP fate of many schools is tied to the performance of their lowest per-

[1] It's important to note that federal reports regarding SWD and LEP subgroup performance differ from our findings here. The National Assessment of Title I: Interim Report (2006) concluded that 23% of schools (they were not broken down by elementary and middle) failed to make AYP in 2003-2004 due to the performance of a single subgroup. Of this 23%, the breakdown was as follows: 13% of schools missed AYP due to the performance of students with disabilities, 4% because of LEP performance, 3% because of low- income student performance, and 3% because of the performance of a single ethnic group. The differences between the federal report and this one may be due to several factors, including: (1) the relatively new NCLB guidelines that exclude small percentages of LEP students and SWDs from taking the state test or that allow them to take alternative assessments; (2) the fact that this report does not calculate the impact of safe harbor on subgroup performance; and (3) the study sample is not nationally representative.

forming subgroup, frequently a subgroup with documented learning challenges. From our results, we could also extrapolate that a school's best strategy for making AYP would be to rid itself of the LEP and SWD subgroups because the presence of one essentially guarantees failure, even in circumstances where these two subgroups outperform similarly identified students in other schools. If that's truly the case, it's unlikely that the current handling of subgroups within NCLB is likely to improve the results achieved.

Some might conclude that we're arguing for different or lower proficiency standards—or both—for LEP students and SWDs. Let's be clear: That's not our argument at all. Instead, we believe the evidence shows that evaluating schools primarily on whether their students meet a fixed, arbitrary, and often low proficiency bar serves all students poorly, including LEP students and SWDs. After all, these students are not members of a homogenous subgroup. LEP students may include some who enter the United States in their teenage years with no formal schooling alongside others who may have attended elite private schools abroad and have exposure to multiple languages. SWDs can range from learners who are academically gifted but challenged by dyslexia, those who perform below their ability because they have behavioral issues, and those with significant cognitive barriers that make learning slower and more difficult. How well is a gifted, dyslexic learner served by meeting a standard that's set to the least-common denominator of performance? And what about a student in Massachusetts (a state with high standards and difficult targets) who has shown promising growth despite huge learning difficulties, but has not yet achieved proficiency? Is that student served well if her school is sanctioned because she and some of her peers did not all achieve a standard that's set to college readiness?

We strongly believe that parents should know how their child is progressing relative to their family's aspirations (which are almost always college readiness). But checking off the number of students who cross a fixed—and low—proficiency bar is a poor way to judge school effectiveness. We believe students would be better served by a model that focuses on how effective schools are in promoting student growth. Such a model would require schools to focus their energy on all students—high-, average-, and low-performing—as well as members of subgroups, which could only be beneficial to both school and student. And a model like this would keep schools from focusing all of their energy on the relatively few students who have the best prospects for crossing a proficiency bar during the current year.

On a technical note, the use of confidence intervals seems to have emerged as a coping mechanism for some of NCLB's design problems. Ostensibly the confidence interval exists to account for the possibility of some form of measurement error in the performance of the student population. In 8% to 11% of cases, a school that wouldn't have met the AMOs for overall proficiency ended up meeting its target with the assistance of a confidence interval. We included (but did not report) the confidence interval in the calculation of subgroup performance as well. There is no doubt that the confidence interval helps many subgroups meet their AMOs, subgroups that wouldn't have otherwise met these targets. But the fact that the vast majority of schools (particularly among our middle school sample) still ultimately failed to make AYP suggests that the confidence interval was not the "difference maker" with many schools. That said, we think the logic for including confidence intervals in NCLB's accountability system is weak, and we doubt confidence intervals would be required in a more consistent, rational accountability system.

Taken as a whole, the evidence from the sample suggests that NCLB, as currently implemented, is not a discriminating system. A tremendous amount of money and energy has been spent to create the illusion of accountability. 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 AYP far too often, and some pretty mediocre schools that slid by in some states. So in reality, what passes for accountability feels more like a high-dollar crapshoot. Some schools may really be failing—no doubt that's so—but they get off easy. For others, the dice aren't as kind—they get labeled as failing but are truly competent.

Either way this is not the type of accountability that will, in the long run, really improve schools, states, or nations.