INTRODUCTION

NCLB's accountability and intervention provisions were intended to identify and correct underperforming schools. The ultimate goal—for all students to reach high standards—will not be met if schools are graded inconsistently, yet it's well known that NCLB does not establish a uniform benchmark for determining whether schools make Adequate Yearly Progress (AYP), but, instead, allows for quite a bit of state discretion.

First, states can define proficiency in reading/English language arts (hereafter called reading) and math; as a result, proficiency standards vary widely in their rigor and consistency (National Center for Education Statistics 2007; Cronin, Dahlin, Adkins, & Kingsbury 2007a; Kingsbury, Olson, Cronin, Hauser, & Houser 2003). Second, NCLB allows states to establish their own timetables, or annual measurable objectives (AMOs) for moving all students to the proficient level by 2014. Some states require schools to follow a linear trajectory to the 100% proficiency goal, while others use "stair steps" or a back-loaded trajectory (i.e., more of the required improvement must be made in the final few years). Third, in an effort to recognize the potential for error in any assessment, NCLB permits states to use confidence intervals (a.k.a. margins of statistical error) in determining proficiency rates, and also allows states to define both the methodology for estimating the confidence interval and its size. Fourth, NCLB allows states to establish their own rules governing the size that a subgroup—such as Hispanic/Latino or low-income students—must attain within a school for the group's performance to be included in the school's AYP determination. States are allowed to determine the minimum size of these subgroups and, if the number of students in the group falls below this number, they are not counted separately as a subgroup for accountability purposes (though they are, of course, counted in the overall student population).

Given the various state interpretations of NCLB, it is reasonable to ask whether differences in standards, timelines, and rules lead to differences in the schools identified as ineffective. For example, if a school that made AYP in Washington were suddenly dropped into North Dakota, or Ohio, or Florida, or Texas, would it also make AYP there? And if not, what factors within NCLB explain this? Based on this analysis, what can we learn about the variation of the AYP systems used throughout the country? To explore these questions, this study looked closely at a group of 36 schools (18 elementary and 18 middle schools). The performance of these schools on a common assessment was used to estimate whether each school would have made AYP in each of the 28 states whose accountability systems were studied. In other words, this study examines how each school would fare if the 28 different standards and rules used to govern AYP decisions under the No Child Left Behind act (NCLB) in these 28 states were applied to them.

Literature Review

Whether a school makes AYP or not depends on many factors. In this particular study we focused on four of them. They are:

1. The difficulty of the proficiency cut score on the state test.
2. The proportion of students required to reach the proficiency cut score in a given year, also known as the annual measurable objective (AMO).
3. Whether a confidence interval is applied to proficiency results and its size.
4. The minimum count required for a subgroup to be included in AYP determinations.

Proficiency cut scores and AMOs

A relatively large body of research catalogs differences in state implementations of NCLB and their possible impacts. A number of studies document wide disparities in the state proficiency cut scores (McGlaughlin, Bandiera De Mello, et al. 2008; Peterson and Hess 2008; National Center for Educational Statistics [NCES] 2007; Cronin, et al. 2007; Qian and Braun 2005; Kingsbury et al. 2003;

McGlaughlin and Bandeira de Mello 2002). Others have found differences in the various states' improvement trajectories (Chudowsky and Chudowsky 2008; Porter, Linn, and Trimble 2005; Kim and Sunderman 2004). There is, however, little research available that speaks to the interaction between state proficiency cut scores and these trajectories. For example, some states offset some of the effect of a high proficiency cut score with a back-loaded trajectory of improvement. Other states have lower proficiency cut scores but stricter trajectories for improvement. Whether a particular school makes AYP, then, may be as much a function of the improvement trajectory as the standard's difficulty. Little is known about how these interact in any given state.

Confidence intervals

States have the option to apply a confidence interval to their proficiency scores and the majority of states choose to take advantage of this provision (Fulton 2006). Confidence intervals are ostensibly used to account for sampling error. For example, assume opinion pollsters survey voters in the state of Michigan to estimate their support for a highway bond measure. Obviously the pollsters can't call every voter in Michigan, so they take a sample of 1,000 voters that they hope are representative. They find that 47% of the polled voters support the measure. But they also know that if they repeated the survey with a different sample of voters, the estimate could change. A confidence interval is calculated (based on the number of voters polled) to show how greatly results might vary if the population were resampled. If the poll reports a 95% confidence interval of +/- 3 percentage points, that means that, were the population resampled, the poll would be expected to find between 44% and 50% of voters supporting the bond.

A confidence interval can also be applied to a school's proficiency rate. For example, assume that McKinley Elementary School is required to reach a proficiency rate of 75% in order to reach its AMO and make AYP, but in fact it achieves a proficiency rate of 71%. Assume further, however, that a 95% confidence interval of +/- 6 is calculated by the state and applied to the results. Since McKinley's actual proficiency rate of 71% is within 6 points of the target of 75%, the school would meet this AMO.

Rogosa (2003) argues that the very concept of a confidence interval violates the integrity of a proficiency requirement. In McKinley's case, the school's "real" proficiency rate is as likely to be 65% as it is to be 77%, meaning that the school is far more likely to have failed to reach the proficiency target of 75% than it is to have reached the target. Thus, it would be more reasonable to say that McKinley's status is, at best, *undetermined*. When states use confidence intervals for purposes of NCLB, however, the assumption is that McKinley reached the target.

Other researchers question whether the very concept of the confidence interval is misapplied. Confidence intervals are normally used to compensate for sampling error, but state tests are not administered to a sample of students within a school—they are administered to 95% or more of the eligible students. Thus, the most common justification for the use of confidence intervals wouldn't be appropriate when applied in these circumstances. (M. West, personal communication 2008). This generally leads to an alternate justification for use of the confidence interval, namely, that the state test represents a sample of student performance at a single time, with results possibly varying if students were resampled on a different date. To extend the analogy to opinion polls and voting, this is akin to arguing that election results should be subject to a confidence interval; if the difference in votes between two candidates is within some confidence interval, we should ignore the outcome and revote because the results might be different if we voted the following Tuesday.

The states that employ confidence intervals typically use ranges between 95% and 99% probability, where higher probability means a larger margin around the target value. The differences in the size and application of confidence intervals by the various states can lead to vastly different AYP findings (Erpenbach and Forte 2005; Simpson, Gong and Marion 2005; Porter, Linn, and Trimble 2005). Porter and colleagues found, for example, that the application of a 99% confidence interval increased the proportion of schools that would make AYP in Kentucky schools from 61% to 90% in 2003. The effect of the confidence interval is especially great for small schools or subgroups. In these circumstances, a school

with a proficiency rate far below the actual goal may meet the standard if a large confidence interval is employed.

Minimum subgroup sizes

For purposes of NCLB, schools are accountable for the performance of every subgroup of students that exceeds a minimum size established by each state. These requirements vary widely from as few as five students to as many as one hundred or even more. The number of subgroups contained within a school is influenced by three factors: the size of the school itself (a school of 1,000 students with a 10% Native American population is likely to be required to count this subgroup although a school of 100 students with the same proportion of Native Americans will not); the ethnic diversity within the school; and the state's minimum *n* (number of students in sample) requirement. The requirement that proficiency targets be met by all accountable subgroups has led to considerable debate on whether this results in a "diversity penalty" in which racially integrated schools face more difficulties in reaching AYP than more homogenous schools.

Several previous studies (U.S. Department of Education 2006; Kim and Sunderman 2004; Novak and Fuller 2003; Kane and Staiger 2002) have found that schools serving diverse students were at higher risk for failing to make AYP. In a critique of these studies, Rogosa (2005) claimed that the diversity penalty has been overstated, in part because in many low-income schools, different subgroups may have the same membership. In an inner Los Angeles suburb, for example, the Hispanic/Latino, low-income, and limited English proficient (LEP)[1] subgroups may essentially be composed of the same students, meaning that the proficiency outcome for the Hispanic/Latino students is unlikely to differ from that of the other groups.

Moreover, the term "diversity penalty" is itself problematic, because it can imply that holding educators accountable for failing to educate traditionally disadvantaged children is somehow unfair. It is perhaps fairer to question whether accountability and sanctions should be targeted toward poorly performing subgroups as opposed to the entire school (e.g., offering choice to the students in a failing subgroup rather than the entire school).

Still, there are many schools in which the general student population meets its AMO, yet the school fails to make AYP because of the performance of a single subgroup. In 2004, for example, a report from the U.S. Department of Education (2006) found that in 23% of cases schools failed to make AYP because a single subgroup missed an AMO.

The Need for This Study

Ultimately the interactions among the state standards, proficiency trajectory, confidence interval, school enrollment, and minimum subgroup size determine whether a school makes AYP. But, even though it's evident that the standards and rules differ greatly across states, it's extremely difficult to judge or compare the effect that these differences have on the results for individual schools. If a state's application of these rules leads to an overly permissive environment in which nearly all schools, no matter how deficient, make AYP, then we might say that NCLB produces an *illusion* of educational equity. If the application of these rules leads to great inconsistency in the way similar schools are judged across states, it might be more persuasive to argue that these differences lead to unreliable decisions and a subsequent waste of resources. Then again, if AYP findings are fair and consistent *in spite* of differences in applying the rules, we could argue that these complex processes, although messy, still produce the desired result.

Alas, we have found no research to date that examines the interactions between the difficulty of the proficiency standards and the various rules across states. We intend for this study to fill a critical gap in the research by helping policy makers evaluate the consistency of proficiency expectations across states, and determine whether NCLB is consistent in its effect.

[1] Note that we use "LEP students" and "English language learners" interchangeably to refer to students in the same subgroup.

METHODOLOGY

In this section, we give a brief overview of the methods we used to conduct this study. Appendix 1 contains a complete description of our methodology.

Research Question

The purpose of the study was to explore how differences in the various state implementations of NCLB—in this case differences among the states in proficiency cut scores, AMOs, subgroup sizes, and confidence intervals—might interact to affect the AYP status of 36 schools. To address this question, we applied the proficiency cut scores of 28 states and their key AYP rules to a multistate sample of schools.

Sampling and Overall Approach

To begin we created two samples. The first was a sample of states for which we compared cut scores and AYP rules. The second was a sample of schools for which we used achievement data to evaluate the impact of the various state cut scores and rules on their possible AYP status.

In all, we evaluated 28 states in the study. We included a state in the study if sufficient student records from state testing and Northwest Evaluation Association (NWEA) testing were available to permit a robust estimate of the state's proficiency cut scores in both reading and math for grades three through eight.

Our sample of 36 schools was drawn from seven school systems serving 153 schools and located in six states. It was created to reflect the diversity within the American educational system. The sample included schools large and small from both high- and low-income communities. Some of the sample schools served many ethnic groups, others only one or two. Some educated large numbers of students from special populations and some did not. Our sample included traditional public schools, magnet schools, and charter schools. Across the sample, both student achievement and growth varied greatly. We should emphasize that our goal in creating this sample was diversity and not "representativeness." We tried to create a sample that would allow applying proficiency standards and rules to a wide variety of circumstances. Thus we wanted to know if a high performing, non-diverse school, a low performing, diverse school, or a low-performing homogeneous school would make AYP in more states. Creating a "representative" sample of 36 schools, were that even possible, would not have permitted us to engage in this kind of experimentation.

All 36 of these schools participated in both the appropriate state test and NWEA testing during the 2005–2006 academic year. Because NWEA tests are calibrated to the proficiency cut scores of the 28 states included in the study, we had a means to estimate how students in each school would perform relative to the proficiency cut scores in these states. Thus, we could take a school that may have achieved a 70% proficiency rate in Illinois and estimate what its proficiency rate might have been in Wisconsin, Minnesota, New Jersey, or other states. In addition, we could estimate the proficiency rates for various subgroups within each school. Armed with that information, we could assess whether the proficiency rates achieved by the school and its subgroups would have been sufficient to meet the annual proficiency targets required by all 28 states.

We validated that NWEA estimates of a school's proficiency rate within its own state (based on NWEA tests) closely matched the actual results achieved by the school on their own state assessment. If NWEA's estimates of results for a school are a fair reflection of their actual performance on their own state test, they are also likely to produce reasonable estimates of the school's performance on the tests of other states.

Estimating State Test Results

For *The Proficiency Illusion* (Cronin et al. 2007a), researchers aligned the results on NWEA's Measures of Academic Progress (MAPs) with the proficiency cut scores of 26 states. The alignment procedure that was used is outlined in detail in that report, but briefly, alignment was estimated by comparing the performance of a single

group of students who participated in both NWEA testing and their respective state's test. The process used, known as "equipercentile equating," is quite straightforward. Assume that 50% of a group of students achieved proficiency on their state's test. If we find the point on the NWEA scale that represents the performance of 50% of the group, that point would represent the score on the NWEA test that is equivalent in difficulty to the proficiency cut score on the state assessment. The accuracy of this process was validated in a pilot study (Cronin et al. 2007b) which found that the equipercentile equating method generally produced projected results that were within three percentage points of the actual state test proficiency rate for the five-state study group.

Since *The Proficiency Illusion* was published in 2007, NWEA has completed estimates for three additional states (and lost one of the original states), now giving us cut score estimates for 28 states. These estimates allowed us to take a student score on the NWEA assessment in one state, and use that score to project whether the student is likely to be proficient in each of the 28 states studied. From there, we were able to project the number of students in each sample school who were likely to be proficient. We could also calculate estimated proficiency rates for each school and its various subgroups.

Note that we were unable to estimate cut scores for eighth grade students in two states, New Jersey and Texas, because of insufficient data. As a result of this limitation, we compared results for the elementary school sample across all 28 states studied, but limited comparisons for the middle school sample to the 26 states in which we had cut score estimates through grade eight.

Estimating a School's AYP Status

Although NCLB requires each state to achieve a target of 100% proficiency for its schools by 2014, each state establishes annual benchmarks for proficiency that increase as schools draw nearer to this deadline. These benchmarks are the AMOs we mentioned earlier. To avoid sanctions, schools must meet the proficiency rate required by the AMO each year.

In addition to setting the AMOs, states also determine minimum subgroup size, and whether and how to apply a confidence interval to a school's proficiency results. For purposes of this analysis, we used the state accountability plans that were in place as of February 2008 (U.S. Department of Education 2008) to document the rules in place at that time. By applying a state's rules to our example schools' data, we were able to project whether a school within the sample would likely achieve several key elements used to determine AYP within that state.

The entire set of rules governing AYP is very complex and it was not possible, based on the data available to us, to estimate the actual status of schools in the sample against all of the AYP rules for the states. As a result, we focused our evaluation on several key AYP rules:

- We evaluated whether the overall performance of students, which we estimated based on spring 2006 results on the NWEA assessment, met the AMOs that the state had set for the 2007–2008 academic year.[2]
- For all ethnic subgroups with counts that exceeded the minimum subgroup size for evaluation, we determined whether their performance, as estimated on the spring 2006 NWEA assessment, was sufficient to meet the proficiency target the state set for the 2007–2008 academic year.
- All students with disabilities (SWDs) were included in the school's sample if they also took some form of their state's assessment. If the count for this subgroup exceeded the minimum subgroup size for evaluation, we determined whether the performance of this group met their AMOs.

[2] As indicated, this report builds on *The Proficiency Illusion* (2007), which used 2005–2006 NWEA data to estimate proficiency cut scores in 26 states. At the time, those were the most recent NWEA data available, and we were unable to update the estimates based on newer data for this report. However, by comparing the 2005–2006 data to the 2007–2008 AYP rules from each state, we're able to use states' most recent annual proficiency targets, which have increased quite dramatically since 2006.

- All students reported as LEP pupils by their schools were included in the school's sample if they also took their state's assessment. Once again they were evaluated against the AMOs if the size of the group exceeded the minimum size.

- All students who were reported by their schools as eligible for free or reduced lunch were included in the sample if they also took their state's assessment. This subgroup was evaluated against the AMO when its count exceeded the minimum size.

- For states that used confidence intervals as part of their AYP calculation, we applied the calculation in circumstances when a subgroup's performance fell short of meeting the required proficiency rate.

To make AYP, elementary and middle schools must also test 95% of their eligible students and meet a standard related to an alternate indicator (generally daily attendance). Data were not available to allow us to evaluate the performance of the sample schools in relation to these two indicators.

Schools that fail to meet an AMO can still make the AYP requirements through a "safe harbor" provision in NCLB. To do this, a school must reduce the number of nonproficient students within a failing subgroup by at least 10% relative to the previous year. We did not evaluate the safe harbor provision as part of this study. As a result, readers should expect that some schools that failed to make AYP in our study might make it in real life.

This methodology allowed us to estimate the proficiency results and status relative to several key AYP rules for each of the 36 schools in the sample. Metaphorically speaking, we were able to drop a school that made AYP in California into states like New Mexico, Illinois, and New Jersey and estimate whether that school would also make AYP there, based on that state's AYP rules.