Colorado

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, with a particular focus on groups that have traditionally been left behind. Under NCLB, states submit accountability plans to the U.S. Department of Education detailing the rules and policies to be used in tracking the adequate yearly progress (AYP) of schools toward these goals.

This report examines Colorado's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Colorado's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Colorado's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Colorado's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **12 of 18 elementary schools** and **16 of 18 middle schools** in our sample **failed to make adequate yearly progress** in 2008 under Colorado's accountability system. (This rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students.)
- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools making AYP in Colorado was exceeded in 10 other sample states. In addition, Colorado was one of 10 states with two passing middle schools in the sample (see Figure 1).**
- Most of the schools in our sample that failed to make AYP in Colorado are meeting expected targets for their overall populations but failing because of the performance of individual subgroups, particularly students with disabilities (SWD)[2] and English language learners.[3]

Colorado is a state with an interesting set of rules, which, when working in tandem, put the state in the middle of the sample distribution in terms of how many schools make AYP. First, Colorado's proficiency standards (or cut scores) are relatively easy to achieve. All of them are at or below the 25th percentile in both reading and math. Still, while Colorado's cut scores are low, its annual targets for proficiency–which vary depending on subject and grade–are fairly ambitious (ranging from 79 to 88 percent in 2008); thus, some schools do not make AYP in Colorado *despite* its undemanding proficiency standards. Another wrinkle is that Colorado's minimum subgroup size is 30, smaller than most other states we examined. This means that schools in Colorado will have more subgroups to account for than schools in most other states. In Colorado, then, schools large enough to have many accountable subgroups fail to make AYP while very small, homogenous schools tend to make AYP, even if their overall student achievement is lower.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Colorado Student Assessment Program (CSAP).

[2] SWDs are defined as those students following individualized education plans.

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

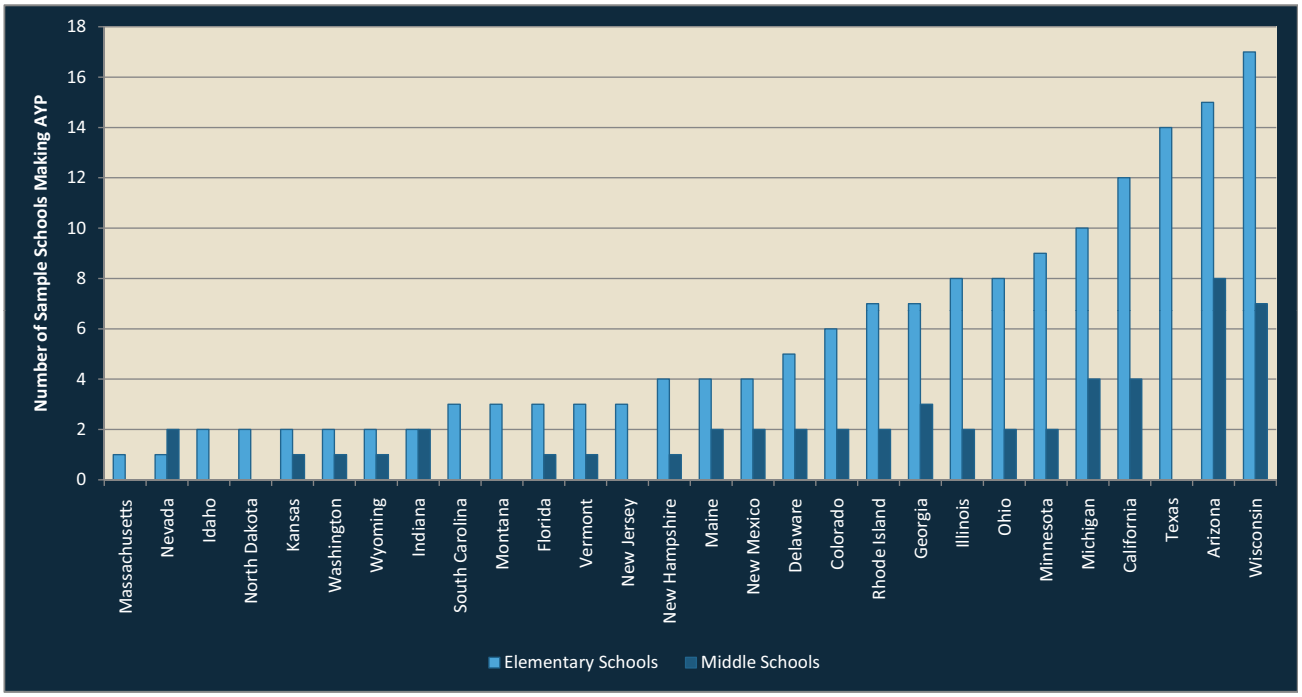

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- One sample school that failed to make AYP in most other states made AYP in Colorado. This is probably because Colorado's proficiency standards (or cut-off scores) are relatively easy compared to other states; this school also had fewer accountable subgroups.

- **Still, while Colorado's proficiency standards are low, its annual targets for proficiency are fairly ambitious (ranging from 79 to 88 percent in 2008); thus, large numbers of schools do not make AYP in Colorado *despite* its undemanding proficiency standards.**

- In Colorado, as in most states, schools with fewer subgroups attain AYP more easily than schools with more subgroups, even when their average student performance is lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- In Colorado, as in most states, middle schools have greater difficulty reaching AYP than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower.

- A strong predictor of a school making AYP under Colorado's system is whether it has enough limited English proficient (LEP) students[4] to qualify as a separate subgroup. Almost every single school with even one such subgroup failed to make AYP.[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Colorado's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a

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

[5] We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, mostly because of the likely differences between how LEP students and SWDs are treated in MAP, the assessment we used in this study, and in the Colorado Student Assessment Program, the standardized state test. Specifically, the U.S. Department of Education has issued new NCLB guidelines in recent years that exclude small percentages of LEP students and SWDs from taking the state test or that allow them to take alternative assessments. In this study, however, no valid MAP scores were omitted from consideration.

computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013-2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Colorado, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

Proficiency cut score estimates for the Colorado Student Assessment Program (CSAP) are taken from *The Profi-*

[6] Low-income students are those who receive a free or reduced-price lunch.

[7] We gave all schools in our sample pseudonyms in this report.

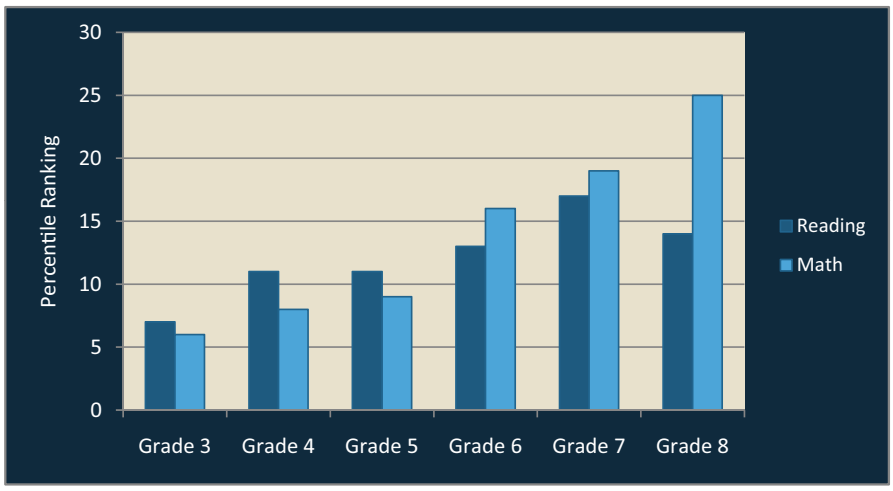

Figure 2. Colorado reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Colorado's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Colorado's cut scores are at or below the 25th percentile.

ciency Illusion (as shown in Figure 2), which found that Colorado's definitions of proficiency ranked well below the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Colorado test, given their performance on MAP.[8] Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Colorado AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Colorado's 2008 AYP rules.

Table 1 shows the pertinent Colorado AYP rules that were applied to elementary and middle schools in this study. Colorado's minimum subgroup size is 30, smaller than most other states we examined.[9] This means that schools in Colorado will have more subgroups to account for than schools in most other states.

Furthermore, most states also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates. Colorado, like most other states in the study, uses a 95% confidence interval. This means even though the AMO might require a school to attain, for instance, 88.4% reading proficiency among its grade 3 students, and 88.4% reading proficiency among its grade 3 students in each subgroup, the real target can be lower, particularly with smaller groups. Note, too, that for different grades and subjects, Colorado applies different AMOs, although all are relatively demanding for 2008.

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states in-

[8] NCLB requires three levels of proficiency: basic, proficient, and advanced. Colorado uses four levels of proficiency on its state test (unsatisfactory, partially proficient, proficient, and advanced). In order to comply with NCLB guidelines, Colorado merged the "partially proficient" and "proficient" categories for AYP purposes. Thus, "partially proficient" students in Colorado are considered "proficient" in terms of AYP accounting. Colorado, however, continues to report four categories of proficiency in its state reporting of CSAP results.

[9] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

Table 1. Colorado AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 30	
	SWDs: 30	
	Low-income students: 30	
	LEP students: 30	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	77.5	88.4
Grade 4	77.5	88.4
Grade 5	77.5	88.4
Grade 6	74.6	86.8
Grade 7	74.6	86.8
Grade 8	74.6	86.8
MATH		
Grade 3	79.5	89.0
Grade 4	79.5	89.0
Grade 5	79.5	89.0
Grade 6	60.7	79.7
Grade 7	60.7	79.7
Grade 8	60.7	79.7

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

clude attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each school's subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Colorado's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Colorado's 2008 AYP rules. **Six elementary schools made AYP while 12 failed to make it.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. Most schools making

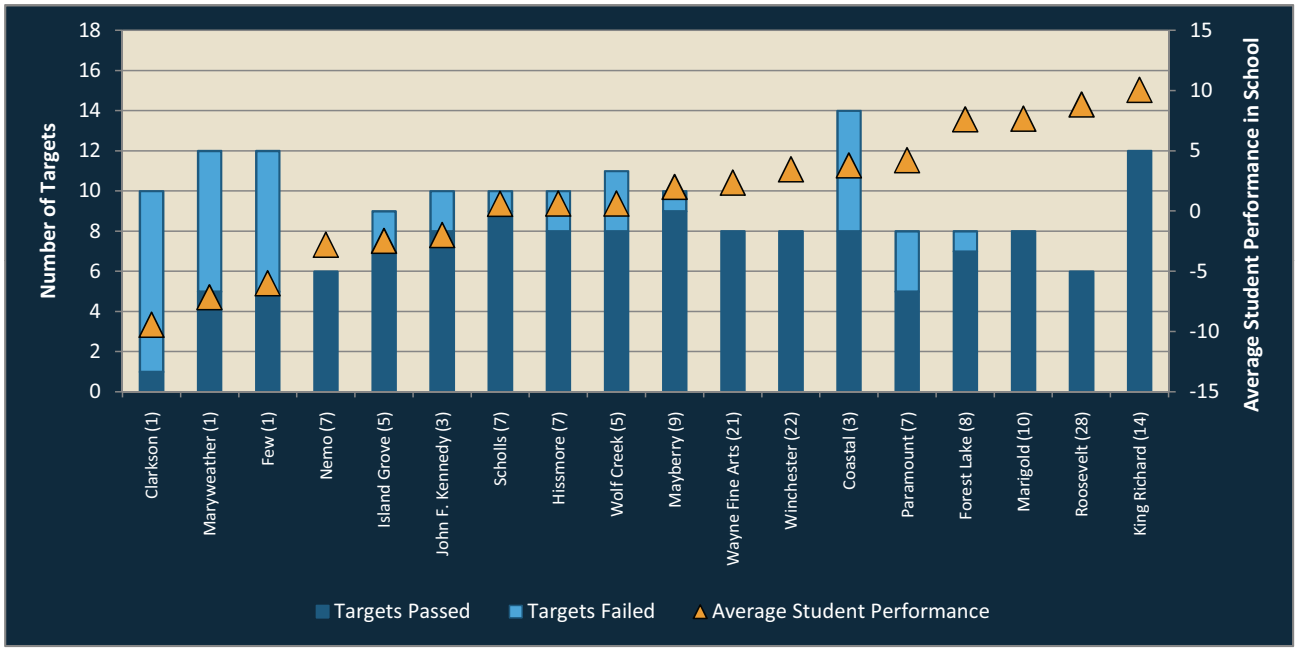

Figure 3. AYP performance of the elementary school sample under Colorado's 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under Colorado's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means the school fails. Mayberry Elementary, for example, met nine of its ten targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

AYP are in the right half of the figure, meaning that the higher performing students were found at these schools.

Yet almost without exception, the only schools actually to make AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet (since each subgroup has its own separate targets to meet). For example, Nemo and Roosevelt made AYP, but have only six targets each.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Colorado AYP rules. **Out of 18 middle schools in our sample, only 2 made AYP** – one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which have relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by Colorado's confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates are increased by the application of the confidence interval. The orange lines show the annual measurable objective needed to meet the targets. These figures show that only two elementary schools (Clarkson and Maryweather) and one middle school (Pogesto) were assisted by the confidence intervals. However, we know from Figure 3 that Clarkson and Maryweather still failed to make AYP because of low subgroup performance.

The effect of confidence intervals on reading proficiency rates for elementary and middle schools is much the same (not shown). In reading, no elementary school is assisted by the confidence interval, but one middle school (Kekata) is helped. However, like Maryweather, Kekata failed to make AYP because of poor subgroup perform-

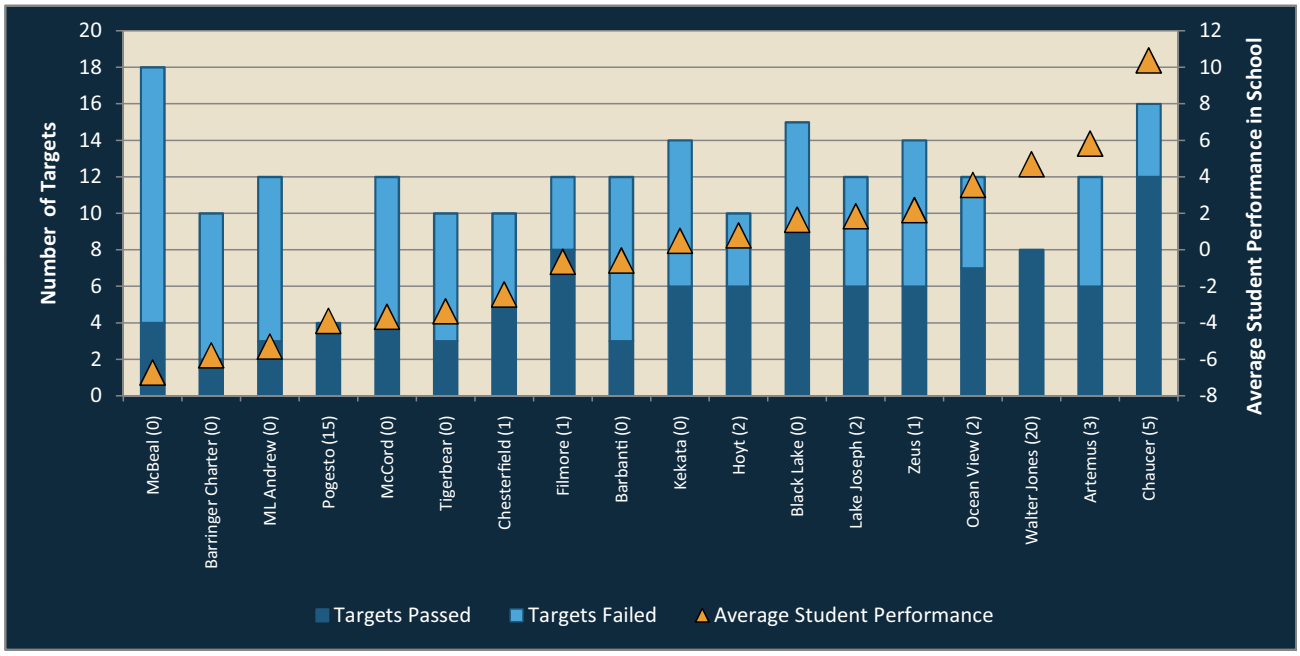

Figure 4. AYP performance of the middle school sample under Colorado's 2008 AYP rules

Note: This figure shows how each middle school within the sample would have faired under Colorado's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup did not make AYP, so any light blue means that the school failed. Hoyt, for example, met 6 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

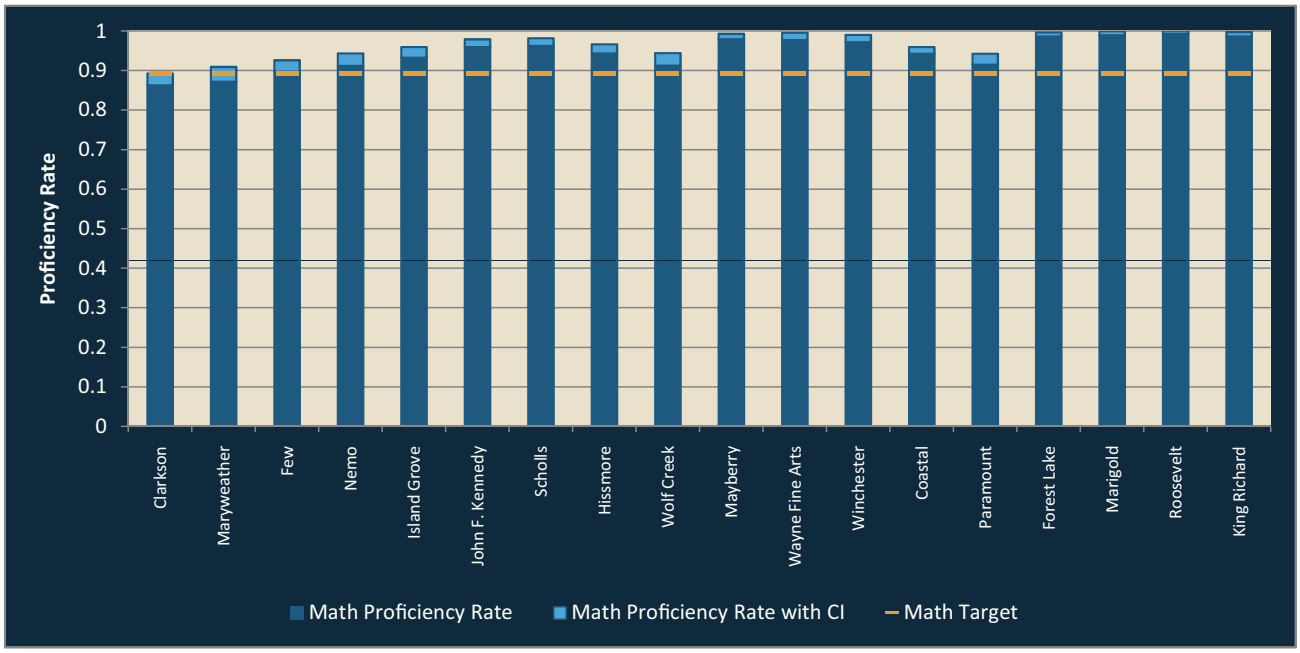

Figure 5. Impact of the confidence interval on elementary school math proficiency rates under the Colorado 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample elementary schools, Clarkson and Maryweather, were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

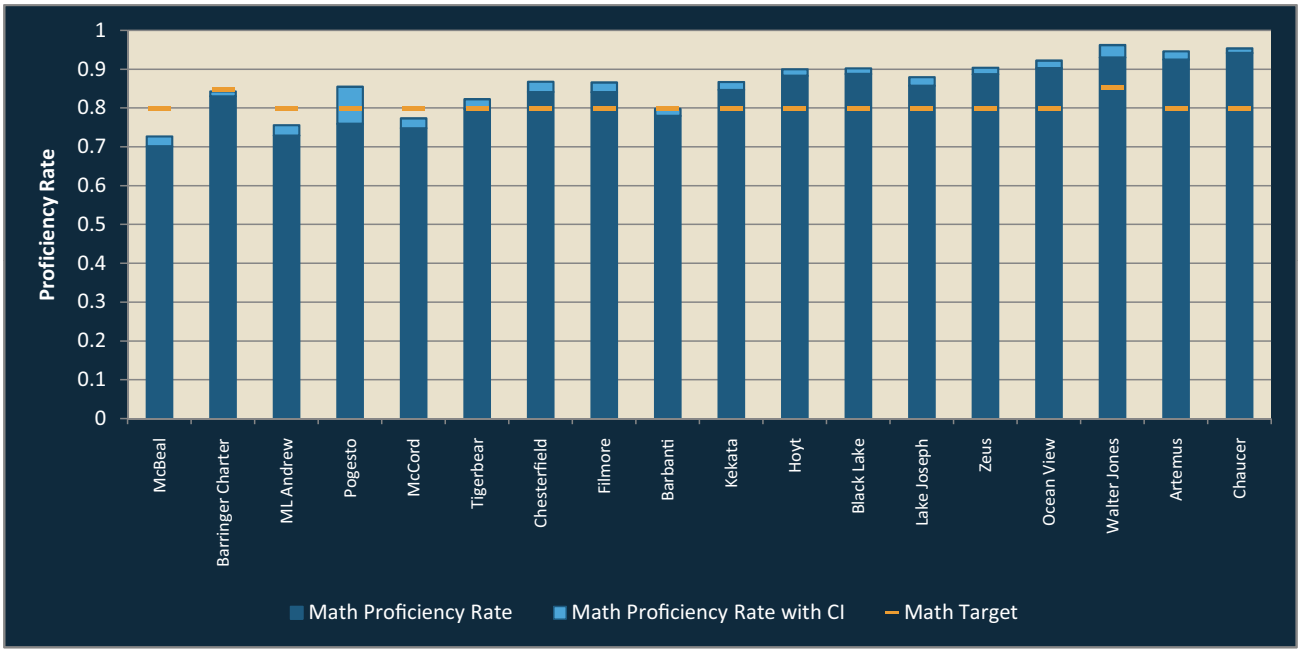

Figure 6. Impact of the confidence interval on middle school math proficiency rates under the Colorado 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample middle schools, Pogesto, was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

ance (Figure 4). **In short, applying the confidence interval has very modest impact on AYP decisions for the sample elementary and middle schools in Colorado.**[10]

Where do schools fail?

Figures 3 and 4 illustrate how schools with low or middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Tables 2 and 3 list information on individual subgroup performance for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 Colorado rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Overall, most elementary schools performed fairly well in terms of meeting AYP targets.
- Three elementary schools failed to meet reading targets for their overall school population. No elementary schools failed in math.
- Four middle schools failed to meet math targets for their overall population and five failed in reading.

[10] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Colorado AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	86.2%	76.9%	Y	N	N	N	N	N	N	N					N	N					10	1	10%	N	1
Maryweather	87.2%	76.7%	Y	N	N	N	N	N	Y	N					Y	N			Y	Y	12	5	42%	N	1
Few	89.7%	80.8%	Y	N	N	N	N	N	Y	N					Y	N			Y	Y	12	5	42%	N	1
Nemo	91.2%	89.8%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Island Grove	93.3%	88.1%	Y	Y				N	Y	Y					Y	N			Y	Y	9	7	78%	N	5
JFK	95.9%	88.4%	Y	Y	Y	N			Y	Y	Y	N							Y	Y	10	8	80%	N	3
Scholls	96.3%	90.3%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	7
Hissmore	94.3%	91.6%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	91.3%	89.0%	Y	Y	N	N		N	Y	Y					Y	Y			Y	Y	11	8	73%	N	5
Alice Mayberry	97.9%	93.4%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	9
Wayne Fine Arts	97.7%	98.9%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	21
Winchester	97.2%	95.3%	Y	Y	Y	Y									Y	Y			Y	Y	8	8	100%	Y	22
Coastal	94.3%	89.7%	Y	Y	N	N	N	N	Y	N	Y	Y			Y	N			Y	Y	14	8	57%	N	3
Paramount	91.4%	90.3%	Y	Y					N	N					Y	N			Y	Y	8	5	63%	N	7
Forest Lake	98.7%	96.0%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	8
Marigold	98.9%	96.4%	Y	Y	Y	Y			Y	Y									Y	Y	8	8	100%	Y	10
Roosevelt	99.3%	99.0%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	98.6%	97.6%	Y	Y	Y	Y	Y	Y	Y	Y					Y	Y			Y	Y	12	12	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

- Three (Scholls, Alice Mayberry, Forest Lake) of the twelve failing elementary schools didn't make AYP because of one target.
- Every LEP subgroup and almost every SWD subgroup at the middle school level did not meet targets in reading and math.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively.[11] As shown, the performance of students with disabilities is proving most challenging for schools under Colorado's system, particularly for middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* of 30. In fact, every single middle school with a SWD population large enough to qualify as a separate subgroup failed to meet its math and reading targets for these students (except Ocean View). Students with LEP also struggled to meet the state's targets; all middle schools with a LEP population large enough to qualify

[11] Recall that elementary students do better on Colorado's math test than middle school students perhaps because Colorado's proficiency scores are easier in math than in reading at the elementary grades (see Figure 2).

Table 3. Middle school subgroup performance of sample schools under the 2008 Colorado AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	70.2%	75.1%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	18	4	22%	N	0
Barringer Charter	83.0%	85.4%	N	N	N	N			N	N	N	N			Y	Y					10	2	20%	N	0
ML Andrew	72.9%	83.3%	N	N	N	N			N	N	N	N			Y	N			Y	Y	12	3	25%	N	0
Pogesto	75.9%	88.9%	Y	Y															Y	Y	4	4	100%	Y	15
McCord Charter	74.8%	85.2%	N	Y	N	N			N	N	N	N			N	Y			Y	Y	12	4	33%	N	0
Tigerbear	79.7%	81.4%	Y	N	N	N			N	N	N	N							Y	Y	10	3	30%	N	0
Chesterfield	84.1%	84.8%	Y	Y	N	N			Y	N	N	N							Y	Y	10	5	50%	N	1
Filmore	84.1%	89.4%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	1
Barbanti	78.0%	83.5%	Y	N	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	0
Kekata	84.7%	85.3%	Y	Y	N	N	N	N	Y	N	Y	N			N	N			Y	Y	14	6	43%	N	0
Hoyt	88.3%	88.7%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	2
Black Lake	88.8%	88.6%	Y	Y	N	N	N		Y	N	Y	N	Y	Y	Y	N			Y	Y	15	9	60%	N	0
Lake Joseph	85.8%	90.0%	Y	Y	N	N	N	N	Y	Y					N	N			Y	Y	12	6	50%	N	2
Zeus	88.7%	88.6%	Y	Y	N	N	N	N	Y	N	Y	N			N	N			Y	Y	14	6	43%	N	1
Ocean View	90.3%	94.1%	Y	Y	N	Y	N	N	N	Y					N	Y			Y	Y	12	7	58%	N	2
Walter Jones	93.0%	93.7%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	92.5%	90.9%	Y	Y	N	N			N	N			Y	Y	N	N			Y	Y	12	6	50%	N	3
Chaucer	94.2%	96.1%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP

as a separate subgroup failed to meet math and reading targets for these students.

Moreover, Hispanic students in Colorado struggled to meet targets as well. At the elementary level, 6 of the 9 qualifying subgroups failed to meet their reading targets. At the middle school level, 6 of 14 qualifying subgroups failed to meet both reading and math targets.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Colorado's NCLB accountability system is, in most respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Colorado, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Colorado.

One exception is Nemo elementary school (see Figure 3) which failed to make AYP in 21 states, yet succeeded in Colorado. Examining Table 2, we can see that Nemo didn't meet the minimum numbers for the LEP and

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Colorado AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	13	6	10
Students with limited English proficiency	7	4	6
Low-income students	17	2	5
African-American students	6	0	1
Asian/Pacific Islander students	0	0	0
Hispanic students	9	1	6
American Indian/Alaska Native students	0	0	0
White students	17	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Colorado AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	15
Students with limited English proficiency	9	9	8
Low-income students	17	8	12
African-American students	11	6	10
Asian/Pacific Islander students	4	0	0
Hispanic students	14	8	8
American Indian/Alaska Native students	1	1	1
White students	17	0	0

SWD subgroups, which created difficulty for many other schools in the sample. Nemo also enrolled fewer than the minimum numbers of African American or Hispanic students to qualify as accountable subgroups. With fewer subgroups, and in a state with relatively easy proficiency standards (Figure 2), Nemo made AYP in Colorado, even when other schools with higher average performance failed.

This is consistent with the patterns shown in Table 6, which compares the sample schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, elementary schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus fewer targets to meet), and much lower per-

Table 6. Comparisons between schools that did and didn't make AYP in Colorado, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	6	12	2	16
Average student body size	231	342	124	951
Average % low income	19	60	42	45
Average % nonwhite	26	48	27	46
Average performance†	4.93	-0.63	0.40	-0.11
Average % growth‡	116	115	109	97
Average number of targets to meet	8	10	6	13

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. Less than 100% growth means that the average student is increasing *by less* than normative amounts, while percentages over 100 mean that the average student is *exceeding* normative growth expectations.

centages of nonwhite students. Similarly, middle schools that make AYP have slightly higher performing students, on average, than middle schools that don't make it, but have smaller total enrollments, smaller nonwhite populations, and fewer subgroups (and thus targets to meet).

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Colorado's AYP rules (and AMOs) for 2008. We found that only 6 elementary schools and 2 middle schools — 8 in all, from a sample of 36—would have made AYP in Colorado. Looking across the 28 state accountability systems examined in the study, this puts Colorado in the upper middle of the distribution in terms of the number of schools making AYP (see Figure 1). Colorado's cut scores are low but its annual targets for proficiency are fairly high; thus, large numbers of schools did not make AYP in Colorado despite its low proficiency standards.

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Colorado's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income students. Almost all of the sample schools met the Colorado reading and math targets for their overall populations, i.e., without considering subgroup results. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and students with disabilities are more generous under the current state assessment sys-

tem,[12] doesn't the massive failure of these students, especially in middle schools, to meet Colorado's targets indicate that a new approach is needed for holding schools accountable for their performance? Yes, schools should redouble their efforts to boost achievement for LEP students and students with disabilities, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

[12] See Footnote 5.