



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 Massachusetts'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 Massachusetts'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 Massachusetts's system as well as under the systems of 27 other states. We used school data and proficiency cut score¹ estimates from academic year 2005–2006, but applied them against Massachusetts's AYP rules for academic year 2007–2008 (shortened to “2008” in this report).

Here are some key findings:

- We estimate that **17 of 18 elementary schools** and **all 18 middle schools** in our sample **failed to make AYP** in 2008 under Massachusetts's accountability system. (This very high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)

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 Massachusetts Comprehensive Assessment System (MCAS).

² At the same time, it's important to note that Massachusetts has improved more than almost every state on the National Assessment of Educational Progress (NAEP) test. In 2007, for instance, it scored first in the nation in fourth- and eighth-grade math and reading.

- Looking across the 28 state accountability systems examined in the study, we find that **virtually all the states (with the exception of Nevada, which ties Massachusetts) exceed Massachusetts in terms of the number of elementary schools making AYP. In addition, Massachusetts is one of only five states (along with Idaho, Montana, South Carolina, and North Dakota) that had no passing middle schools in the sample (see Figure 1).**²
- Middle schools had even greater difficulty reaching AYP in Massachusetts than did elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is any lower than in the elementary schools.
- The only school in Massachusetts that made AYP had only one subgroup (white).

There are several factors in **Massachusetts** which contribute to only one school making AYP in the study. First, the math proficiency standard ranges from a high of the 77th percentile in grade 4 to the 68th percentile in grades 6 and 8. This means that to be considered proficient, grade 4 students must perform better than 77% of all other students in the nation (calculated from the NWEA norms). The reading standard is somewhat lower, ranging from the 65th percentile in grade 4 to the 30th percentile in grade 8. Second, despite the fact that it's lower, Massachusetts still expects a high percentage (roughly 85%) of its grade 3–8 students to reach the reading standard in 2008. These two dynamics, combined with the fact that Massachusetts does not apply a confidence interval (margin of error) to proficiency rate calculations, contribute to only one school making AYP in the study.

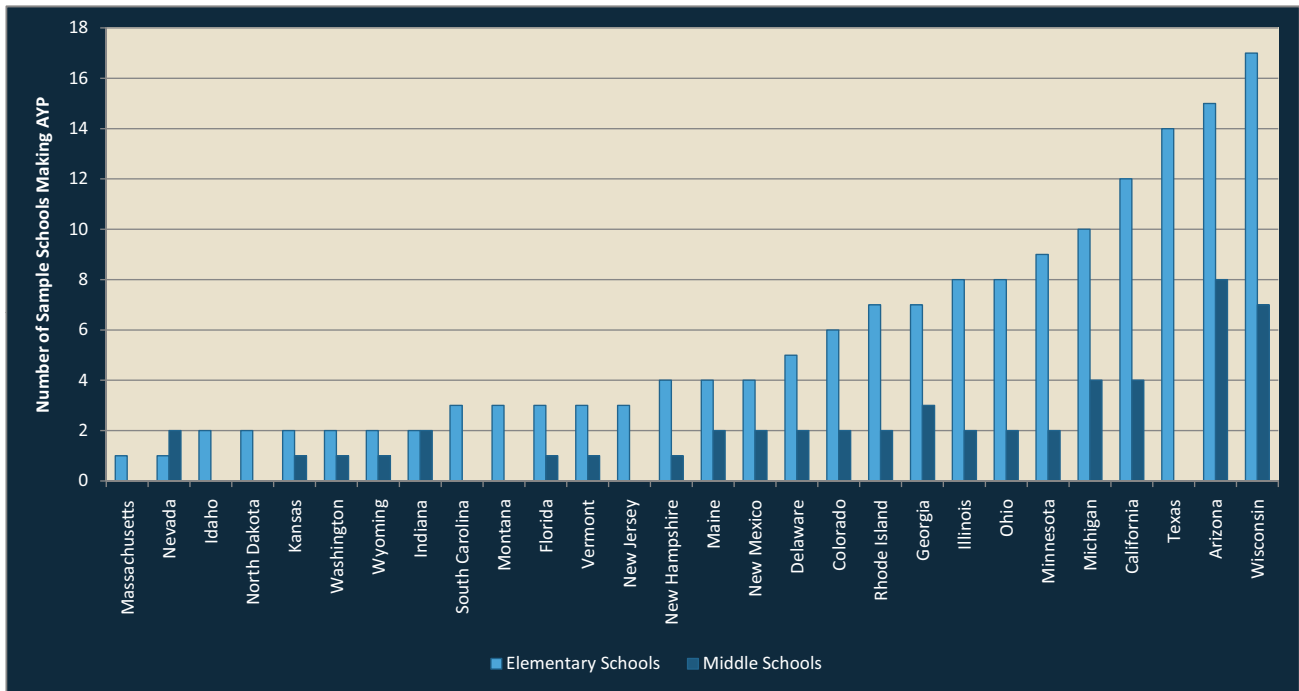


Figure 1. Number of sample schools making AYP by state

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.

- Massachusetts’s high proficiency standards mean that schools will have increasing difficulty in meeting the 100% proficiency requirements of NCLB by 2014.

Introduction

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

dards 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 income³ 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, but not Massachusetts, have also adopted confidence intervals—basically margins of statistical error—to try 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 Ohio, 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 or as English language learners,⁴ 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.⁵

Proficiency cut score estimates for the Massachusetts Comprehensive Assessment System (MCAS) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that **Massachusetts’s definitions of proficiency generally ranked far above the average set by the other 25 states in that study.** These cut score were used to estimate whether students would have scored as proficient or better on the Massachusetts test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Massachusetts 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 Massachusetts’s 2008 AYP rules.

Table 1 shows the pertinent Massachusetts AYP rules that were applied to elementary and middle schools in the current study. Massachusetts’s minimum subgroup size is 40, as long as that constitutes at least 5% of the student population; subgroups can’t be larger than 200 students. The sliding minimum subgroup number used by Massachusetts is not used by most other states, but it means that for many schools, the actual minimum number will be larger than 40.⁶

Massachusetts, unlike most other states examined, does not apply a confidence interval (or margin of statistical

³ Low-income students are those who receive a free or reduced-price lunch.

⁴ Note that we use “students with limited English proficiency (LEP)” or “LEP students” and “English language learners” interchangeably to refer to students in the same subgroup.

⁵ We gave all schools in our sample pseudonyms in this report.

⁶ This means that a school with a total population of 1000 would have a minimum subgroup size of 50 (i.e., 5%), but a school with only 200 students would have a minimum subgroup size of 40, since 5% of 200 (i.e., 10) is below the subgroup minimum of 40. Similarly, a hypothetical school of 5,000 would have a minimum subgroup size of 200, since 5% of 5,000 (i.e., 250) is greater than the subgroup maximum of 200.

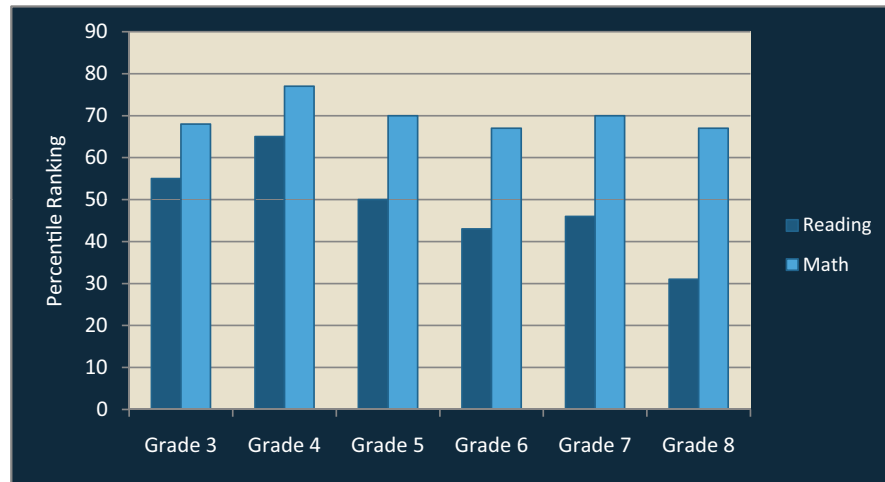


Figure 2. Massachusetts reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Massachusetts'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. Though Massachusetts's cut scores vary by grade and subject, all of the math cut scores and half of the reading cut scores are at or above the 50th percentile.

Table 1. Massachusetts AYP rules for 2008

Subgroup minimum <i>n</i>	Race/ethnicity: 5% of the student population but with a minimum of 40 and maximum of 200	
	SWDs: 5% of the student population but with a minimum of 40 and maximum of 200	
	Low-income students: 5% of the student population but with a minimum of 40 and maximum of 200	
	LEP students: 5% of the student population but with a minimum of 40 and maximum of 200	
CI	Applied to proficiency rate calculations?	
	Not used	
AMOs	Baseline proficiency levels as of 2002 (index)	2008 targets (index)
READING/LANGUAGE ARTS		
Grade 3	70.7	85.4
Grade 4	70.7	85.4
Grade 5	n/a	85.4
Grade 6	n/a	85.4
Grade 7	70.7	85.4
Grade 8	n/a	85.4
MATH		
Grade 3	n/a	76.5
Grade 4	53.0	76.5
Grade 5	n/a	76.5
Grade 6	53.0	76.5
Grade 7	n/a	76.5
Grade 8	53.0	76.5

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; n/a = not available

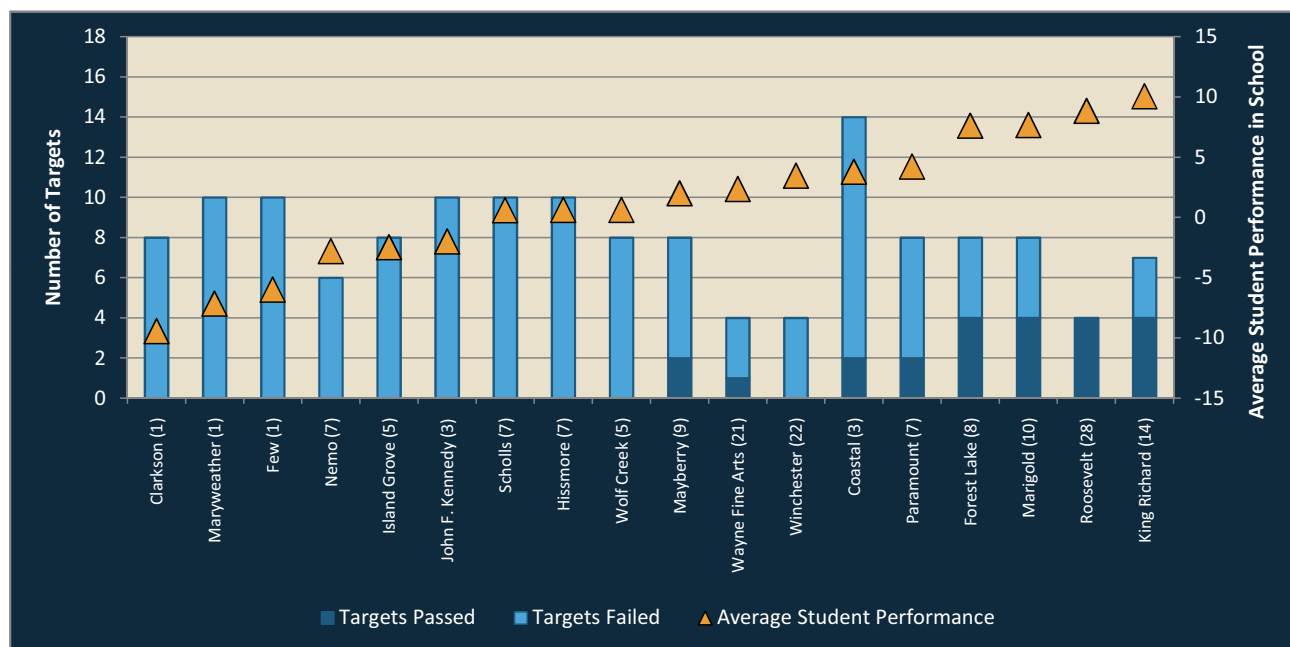


Figure 3. AYP performance of the elementary school sample under Massachusetts's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Massachusetts'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 that the school failed. Marigold Elementary, for example, met four of its eight 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, and 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.

error) to measurements of student proficiency rates. This means that **schools in Massachusetts will have a more difficult time meeting their proficiency targets than similar schools in other states that do use confidence intervals.** Unlike most states examined, however, Massachusetts targets are measured against an index rather than a proficiency percentage, meaning that partially proficient students receive partial credit.⁷

Note that we were unable to examine the impact 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's 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 include 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 subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are

⁷ Six of the states (Minnesota, Rhode Island, Vermont, Wisconsin, New Hampshire, as well as Massachusetts) in our 28-state sample use an index that gives full credit to students who achieve proficient (or better) and partial credit to students performing at lower levels. Consequently, the resultant score in states using this "hybrid" model is always higher than the actual proficiency percentage (giving students partial credit for achieving lower proficiency levels is obviously better than no credit, at least for the schools' ratings). The index provides a fair amount of help when annual targets are below 50%; however, once targets rise above 75%, the index has far less impact.

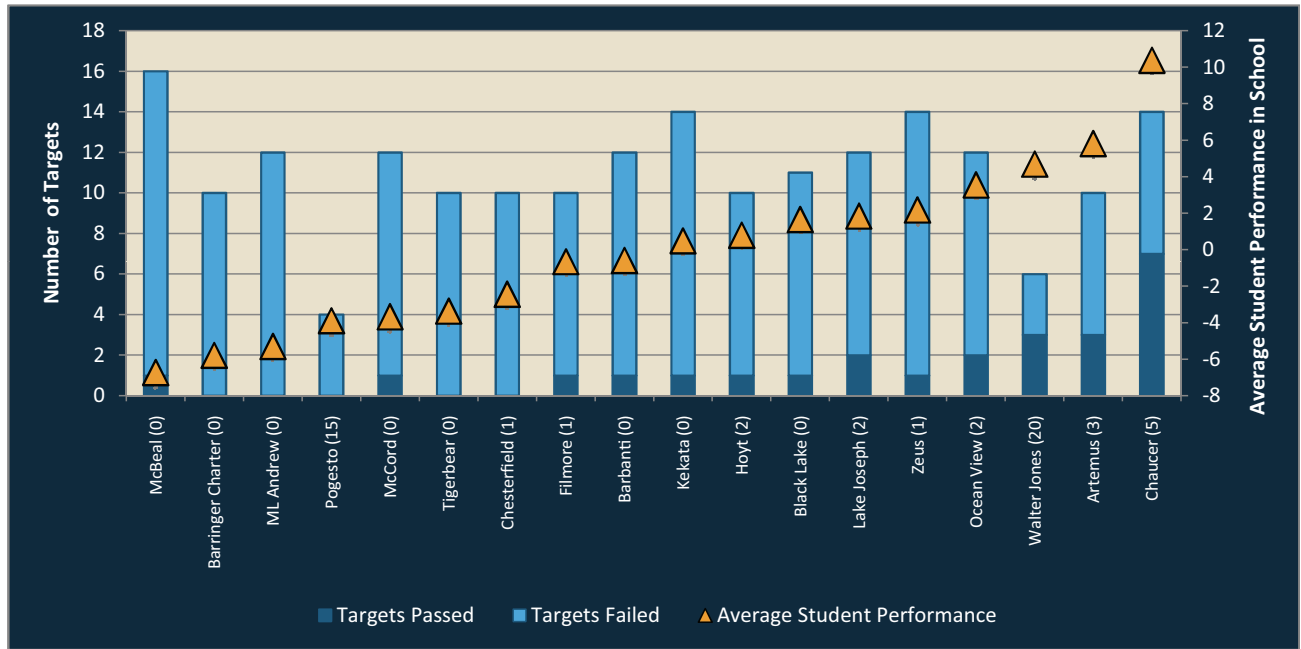


Figure 4. AYP performance of the middle school sample under Massachusetts's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Massachusetts AYP rules (as described in Table 1). The bars show the number of targets that each school had 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 did not make AYP, so any light blue means that the school failed. Chaucer, for example, met half its 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, and 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.

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 Massachusetts's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Massachusetts's 2008 AYP rules. **Only one elementary school made AYP while seventeen 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. The only ele-

mentary school (Roosevelt) that made AYP had just one subgroup, which resulted in only four targets for the school to meet (two targets for the overall population in reading and math, and two more targets for the white subgroup in reading and math).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Massachusetts AYP rules. None of the 18 middle schools made AYP.

Where Do Schools Fail?

Figure 3 shows that having few targets is crucial to making AYP, but neither Figures 3 or 4 indicates which subgroups failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 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

Table 2. Elementary school subgroup performance of sample schools under the 2008 Massachusetts 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	AYP	Targets				
Clarkson	46.8%	52.0%	N	N			N	N	N	N					N	N					8	0	0%	N	1	
Maryweather	50.8%	57.9%	N	N			N	N	N	N					N	N			N	N	10	0	0%	N	1	
Few	56.3%	60.5%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1	
Nemo	57.4%	70.6%	N	N					N	N									N	N	6	0	0%	N	7	
Island Grove	58.5%	71.4%	N	N					N	N					N	N			N	N	8	0	0%	N	5	
JFK	64.7%	69.5%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	3	
Scholls	70.2%	73.3%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	7	
Hissmore	69.7%	74.4%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	7	
Wolf Creek	65.6%	73.9%	N	N					N	N					N	N			N	N	8	0	0%	N	5	
Alice Mayberry	70.0%	76.9%	N	N					N	N	N	N							Y	Y	8	2	25%	N	9	
Wayne Fine Arts	68.2%	85.1%	N	N															N	Y	4	1	25%	N	21	
Winchester	70.9%	81.0%	N	N															N	N	4	0	0%	N	22	
Coastal	75.3%	77.5%	N	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	2	14%	N	3	
Paramount	76.4%	79.4%	N	N					N	N					N	N			Y	Y	8	2	25%	N	7	
Forest Lake	83.8%	86.2%	Y	Y	N	N			N	N									Y	Y	8	4	50%	N	8	
Marigold	83.0%	85.6%	Y	Y	N	N			N	N									Y	Y	8	4	50%	N	10	
Roosevelt	84.5%	92.1%	Y	Y															Y	Y	4	4	100%	Y	28	
King Richard	83.9%	90.1%	Y	Y	N	N			N										Y	Y	7	4	57%	N	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.

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 Massachusetts 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:

- Four elementary schools (Forest Lake, Marigold, Roosevelt, King Richard) met the reading and the math targets for their overall school population.
- Five middle schools met reading targets for their overall population and only one middle school (Chaucer) met its math target for its overall school population.
- Most of the subgroups in both elementary and middle schools failed to meet their targets.

Table 3. Middle school subgroup performance of sample schools under the 2008 Massachusetts 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	45.2%	71.5%	N	N	N	N	N	N	N	N	N	N			N	N	N	N	N	N	Y	16	1	6%	N	0	
Barringer Charter	48.5%	71.4%	N	N	N	N			N	N	N	N			N	N						10	0	0%	N	0	
ML Andrew	44.9%	78.1%	N	N	N	N			N	N	N	N			N	N					N	N	12	0	0%	N	0
Pogesto	44.0%	82.4%	N	N																	N	N	4	0	0%	N	15
McCord Charter	48.2%	80.3%	N	N	N	N			N	N	N	N			N	N					N	Y	12	1	8%	N	0
Tigerbear	55.1%	76.0%	N	N	N	N			N	N	N	N									N	N	10	0	0%	N	0
Chesterfield	57.9%	79.1%	N	N	N	N			N	N	N	N									N	N	10	0	0%	N	1
Filmore	57.9%	83.3%	N	N	N	N			N	N					N	N					N	Y	10	1	10%	N	1
Barbanti	56.1%	79.3%	N	N	N	N	N	N	N	N					N	N					N	Y	12	1	8%	N	0
Kekata	64.5%	82.3%	N	N	N	N	N	N	N	N	N	N			N	N					N	Y	14	1	7%	N	0
Hoyt	63.9%	83.6%	N	N	N	N			N	N	N	N									N	Y	10	1	10%	N	2
Black Lake	68.4%	84.3%	N	N	N	N			N	N	N	N			N						N	Y	11	1	9%	N	0
Lake Joseph	64.6%	86.6%	N	Y	N	N	N	N	N	N					N	N					N	Y	12	2	17%	N	2
Zeus	68.6%	85.1%	N	N	N	N	N	N	N	N	N	N			N	N					N	Y	14	1	7%	N	1
Ocean View	68.4%	90.7%	N	Y	N	N	N	N	N	N					N	N					N	Y	12	2	17%	N	2
Walter Jones	75.1%	86.1%	N	Y					N	N											Y	Y	6	3	50%	N	20
Artemus	75.0%	88.1%	N	Y	N	N			N	N					N	N					Y	Y	10	3	30%	N	3
Chaucer	79.1%	93.4%	Y	Y	N	N	N	N	N	N			Y	Y	N	Y					Y	Y	14	7	50%	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.

Tables 4 and 5 summarize subgroup performance for sample elementary and middle schools, respectively. In examining these, a few points become clear. First, none of the subgroups did very well with the reading and math tests, most likely because Massachusetts's proficiency standards are among the highest in the nation, and because unlike most other states, it does not use confidence intervals as a tool to boost its reported proficiency rates. The only subgroups within the sample elementary and middle schools that ever reached their targets are the white and Asian subgroups (with the exception of one Hispanic subgroup at Chaucer)—

neither of which is traditionally academically disadvantaged. It is likely that as NCLB's 100% proficiency deadline approaches, schools in Massachusetts will face increasing sanctions because of their current high standards.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that schools that failed in the majority of other states failed in Massachusetts too.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Massachusetts 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	8	8	8
Students with limited English proficiency	4	4	4
Low-income students	15	15	14
African-American students	5	5	5
Asian/Pacific Islander students	0	0	0
Hispanic students	7	7	7
American Indian/Alaska Native students	0	0	0
White students	16	9	8

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Massachusetts 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	16
Students with limited English proficiency	7	7	7
Low-income students	17	17	17
African-American students	10	10	10
Asian/Pacific Islander students	1	0	0
Hispanic students	13	13	11
American Indian/Alaska Native students	1	1	1
White students	17	14	4

Nevertheless, Massachusetts does produce some anomalies. Winchester and Wayne Fine Arts Elementary Schools both made AYP in the majority of the other states examined, but failed in Massachusetts. The same pattern holds true for Walter Jones Middle School. These failures are almost certainly the consequence of Massachusetts's higher proficiency standards and lack of confidence intervals, compared to the other states examined. In fact, the only school within our sample

that made AYP under the Massachusetts rules was Roosevelt Elementary, which had a much smaller proportion of traditionally academically disadvantaged students (e.g., low income) and far fewer subgroups (and hence, fewer targets to meet) (see Table 6).

Concluding Observations

This study examined the test performance data of stu-

Table 6. Comparisons between schools that did and didn't make AYP in Massachusetts

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	1	17	0	18
Average student body size	262	307	n/a	859
Average % low income	13	48	n/a	45
Average % nonwhite	19	42	n/a	44
Average performance [†]	8.85	0.78	n/a	-0.05
Average % growth [‡]	103	116	n/a	98
Average number of targets to meet	4	8	n/a	11

[†] 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.

dents from 18 elementary and 18 middle schools across the country to see how these schools would fare under Massachusetts's AYP rules (and AMOs) for 2008. Among this sample, only one elementary school and no middle schools—one from a sample of 36—would have made AYP in Massachusetts. Looking across the 28 state accountability systems examined in the study, this puts Massachusetts at the very low end of the sample distribution in terms of the number of schools making AYP (see Figure 1). Massachusetts' high proficiency standards (and lack of confidence intervals to boost proficiency rates) will mean that schools will have increasing difficulty in meeting the 100% proficiency requirements of NCLB by 2014.

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, Massachusetts'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. In the pre-NCLB era, such schools might have been considered 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. In the case of Massachusetts, is it “fair” that a state is penalized for having rigorous proficiency standards and annual targets? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, 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 re-authorization 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.