

APPENDIX A:

COMPLETE METHODOLOGY

The purpose of this study was to explore how key elements of NCLB, in this case proficiency cut scores, proficiency rate targets, subgroup sizes, and confidence intervals, interact to affect the AYP status of schools. We hoped to shed light on such questions as “Would a school with a population and performance mix that makes AYP in California also be likely to make AYP in New Hampshire, Washington, or South Carolina?” We pursued this by applying each state’s proficiency cut scores and several key rules related to AYP to achievement data from a multistate sample of schools that were chosen to reflect a broad range of student performance, income, and growth in student achievement.

Sample

We started by creating two samples. The first was a sample of states for which we could compare cut scores and AYP rules. The second was a sample of schools for which we could use achievement data to evaluate the impact of the various cut scores and rules on their possible AYP status.

States Sample

In all, we included 28 states in our study (see Table A-1). States were included in the study if sufficient student records from state and 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.² Twenty-six of these cut score estimates were originally reported in *The Proficiency Illusion* (Cronin et al. 2007a). To estimate the majority of cut scores used in this study, we used achievement data from the 2005–2006 school year. Since *The Proficiency Illusion* was published, cut scores for 3 additional states were estimated using achievement data from the 2006–2007 school year. Cut scores were estimated for grades three through eight, and these were used to determine the proficiency rates of the sample schools. There were some exceptions, as follows:

Table A-1. States and grades included in the study sample and terms used for alignment estimate*

State	Term	Grades †
Arizona	Spring 2005	3,4,5,6,7,8
California	Spring 2006	3,4,5,6,7,8
Colorado	Spring 2005	3,4,5,6,7,8
Delaware	Spring 2006	3,4,5,6,7,8
Florida	Spring 2007	3,4,5,6,7,8
Georgia	Spring 2007	3,4,5,6,7,8
Idaho	Spring 2006	3,4,5,6,7,8
Illinois	Spring 2006	3,4,5,6,7,8
Indiana	Fall 2006	3,4,5,6,7,8
Kansas	Fall 2006	3,4,5,6,7,8
Maine	Spring 2006	3,4,5,6,7,8
Massachusetts	Spring 2006	3,4,5,6,7,8
Michigan	Fall 2005	3,4,5,6,7,8
Minnesota	Spring 2006	3,4,5,6,7,8
Montana	Spring 2006	3,4,5,6,7,8
Nevada	Spring 2006	3,4,5,6,7,8
New Hampshire	Fall 2005	3,4,5,6,7,8
New Jersey††	Spring 2006	3,4,5,6,7
New Mexico	Spring 2006	3,4,5,6,7,8
North Dakota	Fall 2006	3,4,5,6,7,8
Ohio	Spring 2007	3,4,5,6,7,8
Rhode Island**	Fall 2005	
South Carolina	Spring 2006	3,4,5,6,7,8
Texas††	Spring 2006	3,4,5,6,7
Vermont	Fall 2005	
Washington	Spring 2006	3,4,5,6,7,8
Wisconsin	Fall 2005	3,4,5,6,7,8
Wyoming	Spring 2007	3,4,5,6,7,8

*The table shows that a number of states administer their state assessment in the fall. For these states we estimate the cut score using fall data and convert that estimate to the equivalent spring score, using percentile ranks. This permits us to evaluate each state’s results using NWEA data from a single term.

** Rhode Island, Vermont, and New Hampshire use the New England Common Assessment Program Tests. Cut score estimates for these states are based on the estimates for New Hampshire.

†The same grades were included for both math and reading.

††Because eighth-grade cut scores for New Jersey and Texas couldn’t be estimated, we didn’t include these states in the middle school portion of the study.

² We require a sample of 700 or more students at each grade to generate a cut score estimate.

- New Hampshire, Rhode Island, and Vermont report on AYP using a common, jointly developed state test called the New England Common Assessment Program (or NECAP), and all three states use the same proficiency cut scores on that test to evaluate student performance. The rules used to evaluate school AYP, though, including annual targets, differ across the three states. Our estimated cut scores on NECAP were derived from a sample of New Hampshire students, but our AYP analyses consider each state's rules separately.
- No school districts within Maryland use NWEA tests for math, so cut score estimates were available only for reading. Consequently, although Maryland reading cut scores were reported in *The Proficiency Illusion*, Maryland is not included in the current study.
- Sample sizes were inadequate to produce eighth grade cut score estimates in Texas and New Jersey. In these cases, we analyzed only elementary schools under the AYP rules in these two states.
- **Student performance (net student achievement in reading and math):** The average raw scale score difference between the students' performance and the median performance (based on NWEA [2005]) for their grade in this subject. As a rule of thumb, a difference of six scale score points is roughly equivalent to a difference of one school year in median achievement.
- **Income level (proportion of school population eligible for free or reduced-price lunch):** This was the only available variable that is a surrogate for family income.
- **Student grade (elementary and middle school groupings):** One finding from *The Proficiency Illusion* (Cronin et al. 2007a) was that middle schools tended to have more difficult standards than elementary schools relative to the NWEA norms. In addition, some states set different AMOs (percentages of students required to meet standards) for elementary and middle school grades. Finally, middle schools, on average, enroll more students than elementary schools. As a result, we created two study groups one composed entirely of middle schools, the other comprising only elementary schools.
- **Student growth (net student growth in reading and math):** This is the average scale score difference on NWEA's assessment, the Measures of Academic Progress (MAP) between student scores in fall to spring terms relative to the NWEA RIT Point Norms (NWEA 2005). This metric compares the average growth of students to the growth of students who started with the same scale score in that grade.

Example Schools

We chose 36 schools to serve as example schools in the study, treating the data from students in these schools as if the school existed in each of the 28 sample states (26 for middle schools). We designed the school selection process to produce a group of schools that reflected breadth in student achievement, school size, diversity, and student growth. The selected schools do not necessarily reflect the demographics of the nation as a whole, nor was that our intention. To create the sample, we contacted 20 school systems to request their participation in the study. Eight school systems that included 153 district and charter schools in the states of Arizona, California, Illinois, Kansas, South Carolina, Washington, and Wisconsin agreed to participate. These school systems supplied student demographic data and state test results to supplement NWEA achievement data that were already stored. Of these schools, 103 were elementary schools and 50 were middle schools. Before we selected the schools, we compiled data on each, relative to the following variables:

Within the elementary and middle school groups, we ranked and classified schools relative to their peers on the achievement, income, and growth variables. Three categories (high, middle, low) were created for the student achievement and student growth variables, with the upper third of schools assigned a classification of high, the middle third assigned average, and the bottom third assigned low. For the income classification, we created two categories. The fifty percent of schools with the highest free or reduced-price lunch population were classified as low income, the other half as high income.

Next, we compiled these classifications into a code that described the achievement, income, and growth status of each school. Thus, a school classified as *high, high, low* (HHL) would be classified as high-achieving, high-income, and low-growth. Eighteen codes were possible (3 achievement × 3 growth × 2 income).

In addition to selecting schools that reflected diversity on these three criteria, we also attempted to select schools in which student performance on their respective state tests was closely predicted by the NWEA assessment. Accordingly, we tried to find schools in which the estimated proficiency rate of students in both reading and math on the NWEA test was within 5% of their actual proficiency rate on their particular state's test.

Here are details of the process we used to select schools:

1. For each cell (e.g., high-achievement, high-income, high-growth), we attempted to find one or more schools with that cell assignment. If there was no school with that cell assignment, we attempted to assign a school with an adjacent assignment, proceeding in the following order (growth → achievement → income). Tables A-2 and A-3 present the results of the sample schools relative to these criteria.
2. Once one or more schools were identified, we selected schools whose predicted proficiency rate on both the NWEA reading and math assessment was within 5% of the actual proficiency rate attained by the school on their own state test. If more than one school met this criterion, we randomly selected a school. If no school met this criterion, we attempted to find a school that met the criterion from an adjacent cell. Tables A-4 and A-5 report the performance of the sample schools on these criteria.
3. In circumstances in which no school met the requirement for predicted proficiency, we selected the school

whose actual state test performance was most closely predicted by the NWEA assessment.

The names of the schools selected were changed to protect their anonymity. We also altered the state and school type for Barringer School, whose identity might be discerned from the school's size and unusual configuration if its state and school type were known.

The data indicate that the elementary schools as a group showed slightly higher than average student performance and slightly higher than average growth when compared with students in NWEA's norming group as a whole (NWEA 2005). The average performance of the middle school group was also higher than the norming group, although the growth of these students was slightly below average. Because the study group had slightly higher than average performance, this group might achieve higher rates of proficiency than a group of schools randomly selected from NWEA's 2005 norming population.

In constructing our sample, we didn't aggregate any information that would communicate the projected proficiency rate of students (on the NWEA test) or the actual size of any subgroup within a school, with the exception of the free and reduced-price lunch rate. We did this intentionally to ensure that the selection process was as free as possible from bias that might derive from having direct knowledge of how the school might fare under the AYP rules of any given state. For example, if we had known that one of the selected schools had 41 Hispanic/Latino students, we would also know that this particular subgroup would be large enough to require AYP consideration in some states but not others. Not compiling this kind of information in advance helped to ensure that the schools—although selected purposefully—were not chosen with knowledge that a school's selection would produce a predetermined result in the various states.

Table A-2. Status of elementary school study group on the selection variables*

Pseudonym	State	Type	State tested in Math	NWEA performance†	Income (percentage in parentheses)	Performance (percentage of average growth in parentheses)‡	Assigned category§	Actual category
King Richard	Illinois	District elementary	415	High (+10.1)	High (13)	High (140)	HHH	HHH
Roosevelt	Wisconsin	District magnet (gifted)	284	High (+8.9)	High (13)	Middle (103)	HHL	HHM**
Marigold	Illinois	District elementary	372	High (+7.7)	High (17)	Middle (122)	HHM	HHM
Forest Lake	South Carolina	District elementary	378	High (+7.6)	High (34)	High (152)	HLH	HHH**
Paramount	Arizona	District elementary	270	Middle (+4.2)	High (37)	Middle (142)	MHM	MHM
Coastal Intermediate	South Carolina	District intermediate	550	Middle (+3.8)	Low (58)	High (131)	MLH	MLH
Winchester	California	District elementary	262	Middle (+3.5)	High (13)	High (139)	LHH	MHH**
Wayne Fine Arts	Wisconsin	District alternative	168	Middle (+2.4)	High (22)	Low (97)	MHL	MHL
Alice Mayberry	South Carolina	District elementary	295	Middle (+2.0)	Low (60)	Low (88)	HLL	MLL**
Wolf Creek	California	District elementary	281	Middle (+0.6)	High (25)	High (133)	MHH	MHH
Scholls	South Carolina	District elementary	279	Middle (+.06)	Low (61)	High (111)	MLM	MLH
Hissmore	South Carolina	District elementary	274	Middle (+0.6)	Low (75)	Middle (103)	MLM	MLM
Island Grove	Washington	District elementary	280	Middle (-2.5)	High (40)	Middle (117)	LHM	MHM**
John F. Kennedy	South Carolina	District elementary	268	Middle (-2.0)	Low (75)	Low (94)	MLL	MLL
Nemo	Wisconsin	District elementary	188	Middle (-2.8)	High (33)	Low (93)	LHL	MHL**
Few	Arizona	District elementary	263	Low (-6.0)	Low (90)	High (135)	LLH	LLH
Maryweather	Arizona	District elementary	224	Low (-7.1)	Low (80)	Middle (113)	LLM	LLM
Clarkson	California	District elementary	434	Low (9.4)	Low (87)	Low (55)	LLL	LLL

*Group is sorted by math and reading performance. Within the table, H stands for high, M for middle, and L for low.

†The number in parentheses reflects the average scale score difference in performance and growth (in math and reading) between students in the school and those in the norming group.

‡The number in parentheses represents the average scale score improvement shown by this school relative to a matched group of students from the NWEA norming group. One hundred percent means that a school is on target in terms of expected growth. 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.

§Performance/income/growth

**Indicates that the selection was from an adjacent cell

Table A-3. Status of middle school study group on the selection variables*

Pseudonym	State	Type	State tested in math	NWEA performance†	Income (percentage in parentheses)	Performance of average growth in parentheses) ‡	Assigned category§	Actual category
Chaucer	California	District middle	1083	High (+10.4)	High (10%)	High (175%)	HHH	HHH
Walter Jones	Arizona	District magnet	165	High (+6.5)	High (38%)	Middle (111%)	HLH	HHM**
Artemus	Illinois	District middle	749	High (+5.8)	High (17%)	Middle (91%)	HHM	HHM
Ocean View	California	District middle	599	High (+3.6)	High (22%)	Middle (85%)	HHL	HHM**
Zeus	South Carolina	District middle	947	Middle (+2.2)	High (42%)	Middle (85%)	MHL	MHM**
Lake Joseph	Washington	District middle	801	Middle (+1.8)	High (34%)	High (111%)	LHH	MHH**
Black Lake	South Carolina	District middle	1380	Middle (+1.7)	Low (46%)	Middle (87%)	HLM	MLM**
Hoyt	South Carolina	District middle	1012	Middle (+0.8)	Low (55%)	Low (79%)	HLL	MLL**
Kekata	South Carolina	District middle	885	Middle (+0.5)	Low (57%)	Middle (103%)	MLM	MLM
Barbanti	California	District middle	1459	Middle (-0.6)	High (45%)	High (130%)	MHH	MHH
Filmore	Washington	District middle	674	Middle (-0.7)	High (40%)	Middle (96%)	MHM	MHM
Chesterfield	South Carolina	District middle	539	Middle (-2.4)	Low (63%)	Low (75%)	MLL	MLL
Tigerbear	South Carolina	District middle	702	Middle (-3.4)	Low (78%)	Middle (87%)	MLH	MLM**
McCord	Wisconsin	Charter	730	Low (-3.7)	High (41%)	Middle (95%)	LHL	LHM**
Pogesto	Washington	District intervention	83	Low (-3.9)	Low (46%)	Middle (107%)	LHH	LLM**
Barringer (K-8)	***	***	2198	Low (-5.0)	Low (81%)	Low (77%)	LLL	LLL
ML Andrew	Wisconsin	District middle	651	Low (-5.3)	High (37%)	Middle (85%)	LHM	LHM
McBeal	Arizona	District middle	808	Low (-6.7)	Low (58%)	Middle (87%)	LLM	LLM

*Group is sorted by math and reading performance. Within the table, H stands for high, M for middle, and L for low.

†The number in parentheses reflects the average scale score difference in performance and growth (in math and reading) between students in the school and those in the norming group.

‡The number in parentheses represents the average scale score improvement shown by this school relative to a matched group of students from the NWEA norming group. One hundred percent means that a school is on target in terms of expected growth. 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.

§Performance/income/growth

**Indicates that the selection was from an adjacent cell

*** Because of the school's very large student population, the state and type was removed to preserve its anonymity.

Table A-4. Comparison of sampled elementary schools' actual state test performance to estimated performance on NWEA test

Pseudonym	State	State math		NWEA math		Count	State proficiency rate, %		NWEA proficiency rate, %		Difference, %	
		Count	Count	Count	Count		Math	Reading	Math	Reading	Math	Reading
King Richard	Illinois	415	296	29	95.3	89.8	95.6	89.1	-0.3	0.7		
Roosevelt	Wisconsin	284	297	-5	93.3	96.3	94.9	98.3	-1.6	-2.0		
Marigold	Illinois	372	278	25	94.4	91.0	96.0	86.3	-1.6	4.7		
Forest Lake	South Carolina	378	373	1	69.3	69.5	68.1	69.1	1.2	0.4		
Nemo	Wisconsin	188	215	-14	65.9	81.3	69.3	85.1	-3.4	-3.8		
Few	Arizona	263	291	-11	75.0	54.3	70.8	59.1	4.2	-4.8		
Maryweather	Arizona	224	219	2	58.6	51.1	63.5	54.8	-4.9	-3.7		
Clarkson	California	435	356	18	19.8	32.4	18.6	32.3	1.2	0.1		
Wolf Creek	California	281	218	22	60.9	54.8	57.8	54.8	3.1	0.0		
Winchester	California	262	212	19	59.9	58.2	64.2	63.0	-4.3	-4.8		
Wayne Fine Arts	Wisconsin	168	174	-4	79.2	92.3	84.0	95.0	-4.8	-2.7		
Paramount	Arizona	270	269	0	84.3	79.7	83.0	80.7	1.3	-1.0		
Scholls	South Carolina	279	268	4	44.9	48.5	48.0	44.8	-3.1	3.7		
Coastal Intermediate	South Carolina	550	523	5	60.7	49.9	57.2	52.1	3.5	-2.2		
Island Grove	Washington	280	238	15	58.9	71.6	58.8	71.2	0.1	0.4		
Alice Mayberry	South Carolina	295	290	2	46.4	48.6	43.4	47.8	3.0	0.8		
John F. Kennedy	South Carolina	268	269	0	33.3	40.8	32.7	38.1	0.6	2.7		
Clarkson	California	274	263	4	37.6	46.6	41.4	47.3	-3.8	-0.7		

Note: Light peach shading indicates a greater than 10% difference in the percentage of students tested.

Appendix A

Table A-5. Comparison of sampled middle schools' actual state test performance to estimated performance on NWEA test

Pseudonym	State	State math		NWEA math		Count		State proficiency rate, %			NWEA proficiency rate, %			Difference, %	
		Count		Count		Difference, %	Math	Reading	Math	Reading	Math	Reading	Math	Reading	
Chaucer	California	1083		1118		-3%	67.8%	68.8%	69.5%	73.5%	-1.7%	-4.7%			
Ocean View	California	599		626		-5%	58.7%	63.8%	52.1%	63.6%	6.6%	0.2%			
Artemus	Illinois	749		426		43%	89.5%	86.7%	92.0%	82.4%	-2.5%	4.3%			
Walter Jones	Arizona	165		172		-4%	87.0%	89.3%	85.5%	85.7%	1.5%	3.6%			
Zeus	South Carolina	1018		947		7%	42.6%	41.3%	46.7%	39.9%	4.1%	1.4%			
ML Andrew	Wisconsin	651		746		-15%	67.6%	75.6%	71.0%	82.2%	-3.4%	-6.6%			
Barringer Charter (K-8)	Illinois	2198		2463		-12%	73.5%	64.1%	76.2%	63.2%	-2.7%	0.9%			
Pogosto	Washington	83		54		35%	27.7%	52.3%	31.5%	53.7%	-3.8%	-1.4%			
McCain	Arizona	808		888		-10%	53.0%	58.7%	56.0%	59.2%	-3.0%	-0.5%			
Barbanti	California	1459		1430		2%	43.8%	45.5%	42.9%	45.3%	0.9%	0.2%			
Filmore	Washington	674		584		13%	42.2%	63.9%	46.2%	60.2%	-4.0%	3.7%			
McCord Charter	Wisconsin	730		790		-8%	65.8%	78.0%	71.4%	83.2%	-5.6%	-5.2%			
Chesterfield	South Carolina	539		523		3%	35.1%	28.7%	30.2%	25.8%	4.9%	2.9%			
Hoyt	South Carolina	1012		975		4%	35.1%	31.4%	36.9%	36.0%	-1.8%	-4.6%			
Kekata	South Carolina	885		855		3%	39.6%	35.7%	42.6%	35.3%	-3.0%	0.4%			
Black Lake	South Carolina	1380		1310		5%	45.0%	35.0%	45.6%	32.8%	-0.6%	2.2%			
Tigerbear	South Carolina	702		676		4%	30.6%	25.9%	32.1%	27.3%	-1.5%	-1.4%			
Lake Joseph	Washington	801		695		13%	48.4%	68.1%	54.8%	67.3%	-6.4%	0.8%			

"Light pink shading" indicates a greater than 10% difference in the percentage of students tested.

"Light peach shading" indicates differences in actual and estimated percent proficient that exceed 5 percent.

Estimating Proficiency Rates

Because each state implements its own tests and sets its own cut scores, we can't directly compare a Wisconsin test result to one in North Dakota. Several previous studies, however, have made comparisons among state tests by aligning their cut scores to a common instrument. Most of these aligned proficiency cut scores to the scale used for the National Assessment for Educational Progress (McGlaughlin et al. 2008; NCEES 2007; Qian and Braun 2005; McGlaughlin and Bandeira de Mello 2002, 2003; McGlaughlin 1998a, 1998b). NWEA's MAPs were used to estimate state cut scores for *The Proficiency Illusion* and other studies (Cronin et al. 2007a; Kingsbury et al. 2003). Results on the MAP assessment were combined with the estimated cut scores for this test to estimate proficiency rates for the sample.

MAP tests are computer-adaptive assessments in the basic skills. Starting in grade two and continuing through high school, these tests are taken by students in more than 3,000 school systems in 49 states and several foreign countries. The MAP tests were developed in accordance with the test design and development principles outlined in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999). The *Computer-Based Testing Guidelines* (2000) of the Association of Test Publishers and the *Guidelines for Computerized-Adaptive Test Development and Use in Education* (American Council on Education 1995) are used to guide test development and practices related to NWEA's use of computer-adaptive testing. Content on the MAP assessments is aligned to the curriculum standards for each state in which it is used, so that the test is a reasonable reflection of the content that students are expected to learn in each state. Because evidence related to the general content validity of MAP assessments is available in Appendix 1 of *The Proficiency Illusion*, we refer interested readers to that document for a more complete discussion of the assessment, its measurement characteristics, and the associated scale.

To estimate proficiency cut scores for *The Proficiency Illusion*, we created a sample population of students who

took both MAP tests and their respective state assessment. Next we calculated the proportion of students in this sample population who performed at a proficient or above level on the state test. Once this was known, we found the score on the MAP scale that would produce an equivalent proportion of students. For example, assume that students must achieve a score of 300 on their state test and that 75% of our sample population achieved that score. If 75% of that sample performed at a scale score of 200 on the MAP assessment, a score of 200 on the MAP score would be equivalent to the state passing score of 300. This is a common method for estimating cut scores across tests and is known as the equipercentile or distributional method.

To evaluate the efficacy of this method, a pilot study of five states was conducted in which the distributional method was used to evaluate how accurately cut scores from one sample predicted the proficiency status of individual students in a second sample in each state (Cronin, Kingsbury, Bowe, & Adkins, 2007b). The results indicated that the cut scores estimated from MAP testing with the first sample accurately predicted the proficiency status of 84% of the students in the second sample in reading and 86% of the students in math. In addition, when applied to the entire sample, the predicted proficiency rate for the sample in each state fell within an average of 3 percentage points of the actual results for the group in reading, and within an average of 2.1 percentage points of the actual results in math.

The latter finding is particularly important for purposes of this study, because it demonstrated that when the estimated MAP cut scores are used, a school's projected proficiency results on the MAP assessment consistently came within 3 points of duplicating its actual results on its state assessment. This means that these methods for estimating cut scores can also be applied to make a reasonable prediction of a school's approximate proficiency rate on its state test.

The cut scores reported in *The Proficiency Illusion* were used for 25 of the states in the sample. These cut scores were estimated from data collected during the spring 2005, fall 2005, or spring 2006 testing terms. An addi-

tional 3 states were included in this study and data for these estimates came from spring 2007 testing data. Sampling data associated with the 25 states studied can be found in Appendix 3 of *The Proficiency Illusion*. The projected MAP percentile ranks associated with proficiency in the 28 states in this study are reported in Appendix B and Appendix C of this document.

The estimated cut scores for each of the states were applied to the 36 sample schools' spring 2006 MAP results in reading and math in order to determine the projected proficiency status of each student relative to each state's standard. Accordingly, students whose MAP scores were equal to or greater than the projected cut score for a state were identified as proficient in that state. From this information, we calculated an estimate of the overall proficiency rate that represented the proportion of students who scored proficient at each school, and derived an estimate of the proficiency rate for the subgroup populations within each school.

Estimating the AYP Status of Schools

The intent of NCLB is to ensure that 100% of each school's students achieve proficient performance in reading and math by the year 2014. To hold schools accountable for progress toward this goal, states set gradually escalating benchmark rates for proficiency that must be achieved by schools each year. These benchmarks, called AMOs, must not only be achieved by the student population as a whole, but also by ethnic subgroup members, low-income students, SWDs and LEP students whose group size exceeds the minimum count required by the state. NCLB also requires at least 95% of the school's enrolled students to take the standard version of the state test, and directs states to identify another indicator of school performance beyond test scores. States generally use attendance as the indicator for elementary and middle schools.

In order to make AYP, schools must meet all the criteria with each and every subgroup. Failing to make AYP for two consecutive years leads to the imposition of sanctions that escalate if the school fails to meet AYP in successive years. Sanctions range from requiring that schools offer students an opportunity to transfer after their

school fails to make AYP for two consecutive years, to eventually closing or reconstituting the school after it fails for six years in a row.

For schools that do not meet the proficiency requirement for any subgroup, many states employ a confidence interval as a safety net. The confidence interval is a statistical measure that provides a margin of error, much like that reported as part of public opinion polls. If the observed proficiency rating for a failing subgroup, plus the estimated margin of error, meets the required proficiency rating, that subgroup is still considered to have met that AMO.

For example, assume that Washington Elementary School (a hypothetical school) tests 100 students from Subgroup E in reading, and assume that a 50% proficiency rate is required to meet the AMO for that group. But only 49 students (49%) pass the reading test. If a 95% confidence interval around the observed pass rate were applied, it might yield a margin of error of approximately ± 4 points, depending on the variability within the sample. Consequently, the confidence interval around the observed pass rate would be 49% plus or minus 4 points, or 45% to 53%. Because the upper range of this interval (53%) exceeds the pass rate of 50% required to meet the AMO in this example, that subgroup would have passed.

Schools that fail to meet the proficiency testing requirements required by NCLB in any given year may also meet an AMO if they meet the criteria necessary to qualify for the act's safe harbor provision. To do this, a school must reduce the number of nonproficient students within a failing subgroup(s) by at least 10% relative to the previous year. If that is accomplished, the school will meet the AMO for that subgroup if at least one additional academic criterion is met. The additional academic criterion varies across states and school levels (e.g., elementary versus high school), but may include attendance rates, graduation rates, percentages of students performing above proficient, or other such indicators. In our study, only a single year's performance data were available at the subgroup level, so it wasn't possible for us to evaluate whether a school might have achieved safe harbor status.

The entire set of rules governing AYP is extraordinarily complex. In addition, based on the data available to us, it wasn't possible to estimate the actual status of the schools in our sample against all the rules. For purposes of this study, then, we limited our evaluation of AYP status to the following rules:

- We evaluated whether the overall performance of students, as estimated by spring 2006 results on the NWEA assessment, would have been sufficient to meet the AYP proficiency target that the state had set for the 2007–2008 academic year.
- 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 set by the state for the same school year. We used ethnic identifiers supplied by the school to assign students to a subgroup. Because these identifiers are not always consistent across school systems, each student had to be reclassified into one of five ethnic subgroups: White, African American, Hispanic/Latino, Asian/Pacific Islander, or American Indian/Alaska Native. Students who were identified as mixed-race, such as White and Native American, were classified with the respective nonwhite subgroup. Students of unknown or unspecified race were removed from the analysis.
- All SWDs in a given school 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 its performance met the AMO for this subgroup.
- All LEP students in a given school were included in the school's sample if they also took their state's assessment. Once again, they were evaluated against the AMO if the count exceeded the minimum size.
- All low-income students in a given school were included in the sample if they also took their state's as-

essment. This subgroup was evaluated against the AMO when its count exceeded the minimum size.

- Students were evaluated in each subgroup for which they qualified. Consequently, the test result of an Asian student who had been classified as LEP would be considered three times, once when determining whether the school as a whole met its AMO, once when considering whether the Asian/Pacific Islander subgroup met its AMO, and once when considering whether the LEP group met that AMO. This application is consistent with the current NCLB rules (Sunderman 2006).
- 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. Some states apply confidence intervals to the proficiency rate; others apply confidence intervals to student scores. Some use two-tailed tests; others use one-tailed. In each case, we applied the method the state reported using for calculating the confidence interval.

States have some leeway to make changes in their plans, subject to approval by the U.S. Department of Education. These changes may include the setting the trajectory for proficiency improvement rates, defining minimum subgroup sizes, and employing confidence intervals. We used the state accountability plans that were in place as of February 2008 (U.S. Department of Education 2008) as the primary form of documentation and applied the rules in place at that time to conduct the analysis.

Because schools report much of the information about subgroups to NWEA separately from their reports to the state, the subgroup identifiers supplied to us for this study are not always identical to those supplied to the state, particularly in terms of student ethnicity. This is one reason we caution that this study does not attempt a formal replication of any particular school's state test results and AYP status. Instead, we aim to illustrate how a school with the particular data supplied to us might perform relative to some of the various states' standards and AYP rules.

For this analysis, then, we attempted to determine the

AYP status of a fixed group of students at a single point in time against the AYP targets for 2008. We included all subgroups that exceeded the minimum size in the analysis and applied confidence intervals for those states in which it was appropriate. We didn't evaluate safe harbor

status, participation rates in state testing, growth models, or average daily attendance in this study, nor did we attempt to evaluate whether a school had met NCLB requirements for bringing adequate numbers of highly qualified teachers on board.

APPENDIX B

Table B-1. Estimated state test proficiency cut scores in reading using MAP*

State	3rd grade	4th grade	5th grade	6th grade	7th grade	8th grade
Arizona	23	25	25	32	30	36
California	61	43	53	56	52	56
Colorado	7	11	11	13	17	14
Delaware	28	32	23	27	23	20
Florida	33	40	53	34	37	50
Georgia	16	16	12	7	12	8
Idaho	33	32	32	34	37	36
Illinois	35	27	32	25	32	22
Indiana	27	27	29	32	34	33
Kansas	35	29	40	32	32	33
Maine	37	43	44	46	43	44
Massachusetts	55	65	50	43	46	31
Michigan	16	20	23	21	25	28
Minnesota	26	34	32	37	43	44
Montana	26	25	27	30	32	36
Nevada	46	40	53	34	40	39
New Hampshire	33	34	34	43	40	48
New Jersey	15	25	16	27	23	n/a
New Mexico	33	32	30	43	32	33
North Dakota	22	29	34	37	30	33
Ohio	21	21	21	25	23	22
Rhode Island	33	34	34	43	40	48
South Carolina	43	58	64	62	69	71
Texas	12	23	30	21	32	n/a
Vermont	33	34	34	43	40	48
Washington	37	23	27	40	49	36
Wisconsin	14	16	16	16	17	14
Wyoming	49	49	44	52	43	44
28-state median	33	29	32	34	32	36

*In percentile ranks; n/a = not available

APPENDIX C

Table C-1. Estimated state test proficiency cut scores in math using MAP*

State	3rd grade	4th grade	5th grade	6th grade	7th grade	8th grade
Arizona	30	28	33	40	36	42
California	46	55	57	62	59	64
Colorado	6	8	9	16	19	25
Delaware	25	26	24	29	36	36
Florida	30	40	46	52	43	32
Georgia	8	23	10	33	22	15
Idaho	30	34	35	38	41	47
Illinois	20	15	20	20	19	20
Indiana	35	32	31	27	26	34
Kansas	30	34	35	33	45	38
Maine	43	46	46	52	54	53
Massachusetts	68	77	70	67	70	67
Michigan	6	13	21	27	35	32
Minnesota	30	43	54	52	52	51
Montana	43	43	40	45	43	60
Nevada	50	46	46	35	36	38
New Hampshire	41	35	34	44	44	53
New Jersey	13	23	26	40	43	n/a
New Mexico	46	49	54	60	61	56
North Dakota	20	27	23	32	39	41
Ohio	20	31	40	33	32	31
Rhode Island	41	35	34	44	44	53
South Carolina	71	64	72	65	68	75
Texas	30	34	24	35	41	n/a
Vermont	41	35	34	44	44	53
Washington	36	46	48	57	59	56
Wisconsin	29	29	26	21	21	23
Wyoming	36	43	43	42	45	51
28-state median	32.5	34.5	34.5	40	43	42

*In percentile ranks; n/a = not available