map GROWTH

MAP Growth College Readiness Benchmarks Technical Manual

MAP Growth College Readiness Benchmarks Technical Manual

Version 2026.1.0

Table of contents

List of Figures

List of Tables

1. Introduction

MAP Growth tests, and educational assessments generally, rely on two types of contextual references for meaningful interpretation of test scores: norms and benchmarks. These two references answer fundamentally different questions. Norms provide a descriptive reference, addressing "what typically occurs," while benchmarks provide a prescriptive reference, addressing "what should occur" (2025 MAP Growth Norms Technical Manual, 2025, p. 2). MAP Growth college readiness benchmarks address this second question.

Benchmarks are reference points used to judge performances against a standard or target. Applied to test scores, they provide interpretive meaning by assigning performance labels such as below basic, basic, proficient, and advanced, commonly used for state summative assessments. The MAP Growth college readiness benchmarks indicate whether students are *on track* or *not on track* to score at or above the college readiness thresholds on ACT and SAT reading and math tests, either directly in Grade 11 or, in earlier grades, under typical growth. These benchmarks enable educators to:

- label student performances as on track or not on track for college readiness,
- identify students who may benefit from supplemental learning opportunities, and
- support conversations by categorizing student performance within a meaningful context.

NWEA continuously updates benchmarks and norms to ensure they remain relevant and accurate. The 2026 college readiness benchmarks replace the previous set, which covered Grades 5 through 9 with ACT benchmarks anchored to scores of 22 and 24 (Thum & Matta, 2015) and preliminary SAT benchmarks issued as an addendum (Thum, 2017). The new benchmarks use post-pandemic test scores, making them more appropriate for understanding college readiness for current students. Table 1.1 summarizes the spring cut scores for ACT and SAT on track performance labels for MAP Growth Grade 6 to 11 Mathematics and Reading tests, and includes associated achievement percentiles derived using the 2025 MAP Growth norms conditioned on week 32.

This technical manual provides comprehensive documentation of the benchmarking and validation methodologies used to update the college readiness benchmarks. This methodology can be summarized in three main stages. First, Grade 11 college readiness cut scores were calculated using a sample of students with observed 11th grade MAP Growth

Table 1.1.: MAP Growth College Readiness Cut Scores and Achievement Percentiles

		Math		Reading	
Grade	Season	ACT	SAT	ACT	SAT
6	spring	237 (81.7)	241 (86.8)	237 (93.3)	223 (74.4)
7	spring	241 (81.3)	245 (86.4)	236 (89.3)	224 (70.5)
8	spring	246 (80.2)	250 (85.3)	235 (84.2)	225 (66.2)
9	spring	246 (79.2)	249 (83.1)	232 (79.1)	224 (64.5)
10	spring	247 (75.9)	250 (79.9)	230 (73.6)	224 (61.9)
11	spring	247 (72.7)	250 (76.9)	227 (68.8)	223 (60.9)

Note. Test scores at or above the cut scores are labeled as *on track* performances. Cut scores are on the RIT scale. Achievement percentiles in parentheses are based on the 2025 MAP Growth norms and conditioned on week 32. Percentiles differ across grades because normative growth at the high end of the distribution flattens in the upper grades; see Section 3.2.2.

and ACT/SAT scores in the spring of 2025. Second, Grade 6 to 10 MAP Growth readiness thresholds were established by backward projecting MAP Growth conditional growth norms. Third, validation analyses evaluated how accurately the benchmarks classify student performances from one year to the next, using a sample of more than 7,000 schools from all 50 states.

2. Sample

Establishing and validating the college readiness benchmarks required two samples that differed in purpose and design. The calibration sample included Grade 11 students from six districts in five states who took a MAP Growth assessment and either the ACT or SAT in spring 2025. We used it to establish the Grade 11 spring cut scores and to evaluate their classification accuracy. The validation sample comprised MAP Growth test events from students in more than 7,000 schools across the 2022-23 through 2025-26 school years. We used it to evaluate the validity of the Grade 6 to 10 benchmarks.

2.1. Calibration Sample

2.1.1. Data Collection

The calibration data set consisted of a sample of Grade 11 students who took MAP Growth assessments and either the ACT or SAT in spring 2025. NWEA recruited districts from across the United States, and participation required a signed data-sharing agreement. We excluded two schools from the mathematics calibration sample because descriptive analyses indicated invalid test administrations. A sensitivity analysis showed that retaining these schools shifted the math cut scores toward the high end of the normative performance distribution. Eligible students took MAP Growth assessments and college entrance exams 10 or fewer weeks apart.

2.1.2. Sample Characteristics

A separate benchmark was established for each combination of MAP Growth subject and college entrance exam: ACT Math, ACT Reading, SAT Math, and SAT Reading and Writing. Across these four combinations, the calibration sample comprised 26 schools in six districts and five states. Individual combinations drew on 15 to 17 schools and 1,030 to 1,634 students across three to five states.

The National Center for Education Statistics (NCES) provides comprehensive data on U.S. schools, including information on student demographics and school characteristics.

Table 2.1 draws on the Common Core of Data (CCD) to describe the characteristics of the schools in the calibration sample (NCES, 2024). The table also compares the sample to the population, namely schools that enroll Grade 11 students. Standardized mean differences (SMDs) between the sample and population are reported to compare the two groups. SMD values greater than or equal to 0.2, 0.5, and 0.8 are considered small, moderate, and large differences, respectively. Because participating districts agreed to share data rather than being randomly selected, the sample is not expected to represent the population. The table allows readers to compare it against both the population and their own local context.

Schools in the calibration sample differed from the population on several dimensions. Differences were large for urbanicity, enrollment, and Mountain region representation: sampled schools were more than twice as likely to be city schools (61.5% versus 26.3%), enrolled about twice as many students as the typical school serving Grade 11 students (1,385 versus 684), and were concentrated in the Mountain region (34.6% versus 9.4%). Differences were moderate for full-time equivalent teaching staff and for rural schools, which were nearly absent from the sample (3.8% versus 36.7%). Because these percentages are based on 26 schools, individual estimates are imprecise and several census divisions are unrepresented.

The calibration sample was used to estimate the relationship between MAP Growth scores and college entrance exam performance, not to estimate the prevalence of college readiness in the population. Differences in school composition affect the resulting cut scores only to the extent that this relationship varies across school characteristics. We assessed this variation and summarized the findings in Section 3.1 (see Appendix A for detailed methodology and results).

2.1.3. Descriptive Statistics

Whereas Table 2.1 describes the schools in the calibration sample, Table 2.2 describes the students and test scores that entered the analysis. The means and standard deviations (SDs) of MAP Growth and ACT/SAT test scores are reported for each combination included in the set of benchmarks. ACT scores are on a scale of 1 to 36, while SAT scores are on a scale of 200 to 800. MAP Growth test events and ACT/SAT tests occurred, on average, four weeks apart.

Table 2.1.: Calibration Sample Representativeness

	Population (N = 25,033)		Sample (N = 26)		
Variable	Mean	SD	Mean	SD	SMD
% Urban	26.3	44.0	61.5	49.6	0.801
% Suburban	24.7	43.1	30.8	47.1	0.142
% Town	12.4	32.9	3.8	19.6	-0.259
% Rural	36.7	48.2	3.8	19.6	-0.682
% East North Central	15.8	36.4	30.8	47.1	0.412
% East South Central	5.9	23.6	0.0	0.0	-0.251
% Middle Atlantic	10.7	30.9	0.0	0.0	-0.346
% Mountain	9.4	29.1	34.6	48.5	0.866
% New England	3.9	19.3	0.0	0.0	-0.201
% Pacific	15.9	36.5	11.5	32.6	-0.118
% South Atlantic	13.6	34.3	0.0	0.0	-0.398
% West North Central	11.2	31.5	0.0	0.0	-0.355
% West South Central	13.7	34.4	23.1	43.0	0.272
% AI/AN	2.6	11.6	0.9	0.9	-0.147
% Asian	3.0	7.1	4.0	5.5	0.137
% Black	14.1	22.2	11.2	15.9	-0.131
% Hispanic	25.8	27.5	35.0	24.3	0.335
% Multiple Races	4.2	4.2	5.2	2.8	0.237
% NH/PI	0.3	2.6	0.6	1.1	0.096
% White	49.8	33.1	43.1	27.8	-0.203
% Alternative/CTE/SPED	18.4	38.7	15.4	36.8	-0.077
% Charter	13.1	33.7	0.0	0.0	-0.388
Enrollment	683.6	780.3	1384.6	765.3	0.898
Full-Time Equivalents	43.7	44.3	76.1	44.3	0.732
Pupil-Teacher Ratio	16.0	15.0	17.8	4.2	0.122

Note. SMD is the standardized mean difference. AI/AN refers to American Indian and Alaska Native students. NH/PI refers to Native Hawaiian and Pacific Islander students.

Table 2.2.: Counts and Descriptive Statistics for Calibration Sample

	Counts				MAP Growth		ACT or SAT	
Benchmark	States	Districts	Schools	Students	Mean	SD	Mean	SD
ACT Math	5	5	15	1,030	231.7	29.7	16.6	5.0
ACT Reading	4	4	17	1,634	216.5	20.4	16.2	6.1
SAT Math	3	6	16	1,122	240.1	22.1	471.1	101.6
SAT Reading and Writing	3	6	16	1,331	221.9	16.2	493.8	102.9

2.2. Validation Sample

2.2.1. Data Collection

The validation data set was drawn from NWEA's research database of MAP Growth test records. Because these records were already available rather than collected through recruitment for this study, the validation sample spans substantially more schools than the calibration sample and more closely resembles the national population of public schools. Eligible students had MAP Growth test events in both a starting grade and a later grade within the study window. School years prior to 2022-23 were excluded because many students and schools experienced intermittent quarantines and closures during the 2021-22 school year (Zviedrite et al., 2024).

2.2.2. Cohort Design

The 2026 MAP Growth college readiness benchmarks used MAP Growth assessment data from the 2022-23 to the 2025-26 academic years to validate Grade 6 to 10 benchmarks. We implemented a cohort-based design to maximize longitudinal information from students who took assessments in multiple years.

Figure 2.1 illustrates the cohort design. Seven cohorts, labeled A through G, contribute test events across the four school years, with each cohort advancing one grade per year. Because a cohort must be observed in both the starting and ending grade, the number of contributing cohorts declines as the projection horizon lengthens: all seven cohorts support one-grade comparisons, five support two-grade comparisons, and three support three-grade comparisons. Grade 6 illustrates the pattern, with cohorts A, F, and G contributing Grade 6 performance labels compared against observed Grade 7 labels, cohorts A and F against Grade 8, and cohort A against Grade 9.

The horizon is further bounded by Grade 11, the highest grade with a benchmark. Grade 6 through 8 benchmarks are therefore validated one, two, and three grades ahead, Grade 9 benchmarks one and two grades ahead, and Grade 10 benchmarks one grade ahead.

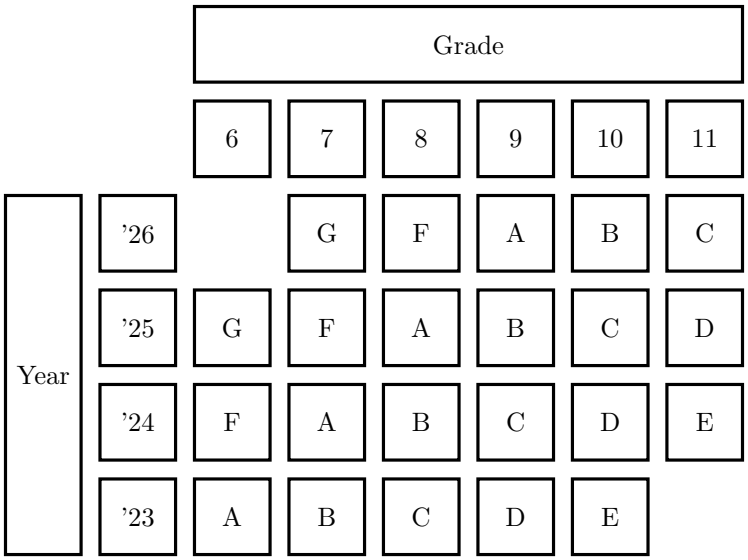

Figure 2.1.: Cohort Design for the Validation Sample

2.2.3. Sample Characteristics

Across the validation analyses, the sample comprised 176,002 students from 7,124 schools in all 50 states. Individual analyses drew on 852 to 4,467 schools and 10,226 to 47,885 students.

Table 2.3 draws on the CCD to describe school characteristics in the validation sample compared to the population (NCES, 2024). The population included all public schools in the United States serving students in Grades 6 to 11. SMDs are interpreted using the thresholds described in Section 2.1.2. Schools in the validation sample differed from the population in minor ways including being more likely to be located in the East North Central region of the United States, more likely to be charter schools, and less likely to be alternative, career and technical education, or special education campuses. Otherwise, the validation sample was not systematically different from the U.S. public school population.

Table 2.3.: Validation Sample Representativeness

	Population (N = 58,207)		Sample (N = 7,124)		
Variable	Mean	SD	Mean	SD	SMD
% Urban	27.7	44.8	30.5	46.1	0.062
% Suburban	27.6	44.7	28.2	45.0	0.013
% Town	10.9	31.2	10.3	30.4	-0.019
% Rural	33.8	47.3	31.0	46.2	-0.060
% East North Central	15.9	36.6	25.2	43.4	0.249
% East South Central	5.6	23.1	3.2	17.6	-0.109
% Middle Atlantic	10.5	30.7	6.0	23.7	-0.152
% Mountain	10.4	30.5	12.8	33.4	0.079
% New England	4.4	20.5	3.2	17.6	-0.058
% Pacific	17.2	37.7	11.9	32.4	-0.142
% South Atlantic	12.8	33.4	10.9	31.1	-0.059
% West North Central	10.2	30.2	8.5	27.9	-0.055
% West South Central	13.0	33.6	18.3	38.7	0.156
% AI/AN	2.2	10.7	2.4	12.1	0.017
% Asian	3.7	8.3	3.5	7.6	-0.019
% Black	14.2	22.8	15.1	23.3	0.041
% Hispanic	27.1	28.4	27.7	28.0	0.021
% Multiple Races	4.6	4.4	4.7	4.1	0.030
% NH/PI	0.4	2.6	0.3	1.7	-0.038
% White	47.8	33.4	46.2	32.9	-0.047
% Alternative/CTE/SPED	9.3	29.0	2.6	15.9	-0.239
% Charter	11.8	32.3	19.3	39.4	0.224
Enrollment	569.4	575.9	655.3	584.5	0.149
Full-Time Equivalents	36.9	33.2	41.6	32.2	0.141
Pupil-Teacher Ratio	15.8	12.4	15.9	9.1	0.007

Note. SMD is the standardized mean difference. AI/AN refers to American Indian and Alaska Native students. NH/PI refers to Native Hawaiian and Pacific Islander students.

Table 2.4.: Counts and Descriptive Statistics for the Mathematics Validation Sample

Grade		Counts			% OT ACT		% OT SAT	
T1	T2	Districts	Schools	Students	T1	T2	T1	T2
6	7	2,432	4,077	33,058	20.0	22.3	15.2	17.3
6	8	2,292	3,837	30,806	19.4	25.1	14.7	19.5
6	9	879	1,476	13,305	17.4	27.4	13.0	22.7
7	8	2,424	4,326	47,302	19.9	23.5	15.3	18.2
7	9	1,055	2,058	27,888	17.8	27.1	13.6	22.2
7	10	852	1,618	12,949	16.0	29.7	12.3	25.5
8	9	1,150	2,418	42,307	19.9	24.6	14.9	19.8
8	10	989	2,009	26,687	17.8	28.5	13.2	24.3
8	11	637	852	10,226	15.1	28.8	10.6	24.5
9	10	1,337	2,470	41,506	22.0	28.1	17.2	23.7
9	11	987	1,293	21,143	18.5	30.6	13.8	26.4
10	11	991	1,312	21,229	25.3	30.5	20.8	26.3

Note. T1 and T2 refer to spring test events. % OT refers to the percent of students performing on track for college readiness at each time point.

2.2.4. Descriptive Statistics

Whereas Table 2.3 describes the schools in the validation sample, Tables 2.4 and 2.5 describe the students and test scores that entered each validation analysis, for mathematics and reading, respectively. Each student took a MAP Growth test event at time 1 (T1) and time 2 (T2). Sample sizes decline as the interval between T1 and T2 lengthens, reflecting the number of cohorts available at each horizon. Across mathematics and reading, one-grade comparisons drew on 21,229 to 47,885 students, two-grade comparisons drew on 21,143 to 31,572 students, and three-grade comparisons drew on 10,226 to 13,511 students.

Table 2.5.: Counts and Descriptive Statistics for the Reading Validation Sample

Grade		Counts			% OT ACT		% OT SAT	
T1	T2	Districts	Schools	Students	T1	T2	T1	T2
6	7	2,437	4,114	33,268	5.0	10.0	30.2	36.4
6	8	2,316	3,907	31,572	5.0	18.1	30.1	43.7
6	9	1,113	1,897	13,511	4.5	29.2	28.6	52.3
7	8	2,426	4,467	47,885	9.4	16.7	34.9	41.6
7	9	1,317	2,646	28,216	8.4	28.6	32.8	50.7
7	10	1,048	2,047	12,718	8.2	38.2	31.5	53.9
8	9	1,433	3,152	42,726	14.2	25.8	37.3	47.0
8	10	1,232	2,610	26,885	12.6	35.5	34.7	50.9
8	11	788	1,020	10,593	10.6	43.4	31.4	52.9
9	10	1,615	3,134	40,853	23.0	33.3	42.9	48.7
9	11	1,169	1,540	21,180	19.4	42.6	38.3	52.0
10	11	1,170	1,565	21,611	29.8	42.3	45.1	51.7

Note. T1 and T2 refer to spring test events. % OT refers to the percent of students performing on track for college readiness at each time point.

3. Establishing College Readiness Benchmarks

College readiness benchmarks for MAP Growth Mathematics and Reading tests were established in two steps. First, we established Grade 11 spring classification thresholds using the calibration sample and receiver operating characteristic (ROC) curve methods to identify MAP Growth scores that optimally classified students according to college readiness criteria for both ACT and SAT. Second, we used MAP Growth conditional growth norms to set college readiness cut scores for MAP Growth Mathematics and Reading tests in prior grades. The final set of benchmarks is summarized in Table 1.1.

3.1. Calibrating Grade 11 Classification Thresholds

The goal of using ROC curves to set cut scores is to optimize classification accuracy against a "gold standard" criterion, which for this application was performing at or above the college readiness thresholds on the ACT and SAT. Table 3.1 shows the college ready score ranges for each test. On the ACT, these ranges correspond to a 75% to 80% chance of earning a C or better and a 50% chance of earning a B or better in credit-bearing courses (Allen, 2013). On the SAT, they correspond to a 75% chance of earning a C or better in first-semester, credit-bearing college courses (CollegeBoard, n.d.).

ROC curves identify the MAP Growth score that best reproduces a student's college ready classification on the ACT or SAT. The method requires two measures: (a) a continuous

Table 3.1.: ACT and SAT College Ready Test Score Ranges and Labels

Test	Score Range	Criterion Label
ACT Math	22 to 36	College Readiness
ACT Reading	22 to 36	College Readiness
SAT Math	530 to 800	College and Career Readiness
SAT Reading and Writing	480 to 800	College and Career Readiness

variable, here MAP Growth test scores, and (b) an external criterion, here whether a student scored at or above the college readiness threshold on the ACT or SAT in the same grade and season. This is conceptually similar to test scale linking, as both involve relationships between measures, but it serves a different purpose and is not a form of test scale linking. Test scale linking relates scores on two or more measurement scales using methods such as linear equating, equipercentile linking, and score concordance tables, whereas ROC curves relate scores on a single scale to an external criterion.

Comparing MAP Growth scores against that criterion produces four classification outcomes, shown in Figure 3.1. Each panel plots MAP Growth RIT scores against ACT or SAT scores, with dashed lines marking the cut score on each scale; those cut scores are the benchmarks established later in this section. Students who scored at or above both cut scores are correctly labeled on track (true positive, upper right quadrant), and students who scored below both are correctly labeled off track (true negative, lower left quadrant). A student's MAP Growth test performance is mislabeled as on track when they actually perform in the off track score range on the ACT or SAT (false positive, upper left quadrant). Likewise, a student's MAP Growth performance is mislabeled as off track when they actually score in the college ready range on the ACT or SAT (false negative, lower right quadrant).

Figure 3.2 shows ROC curves for the empirical MAP Growth and ACT or SAT score distributions. Each point on a curve represents a different MAP Growth cut score, with its own true positive and false positive rates. The curves were constructed by calculating the true positive and false positive rates at every possible cut score rather than by fitting a regression-based prediction model.

True positive rates (TPR) are shown on the y-axes and false positive rates (FPR) on the x-axes. The *true positive rate* is the proportion of students who met the ACT or SAT college readiness threshold that the cut score labels on track; the *false positive rate* is the proportion of students who did not meet it that the cut score labels on track. Both rates are computed among students grouped by their actual ACT or SAT performance, whereas the predictive values in Table 3.2 are computed among students grouped by their MAP Growth label.

The ROC curves also display the *area under the curve (AUC)*, the proportion of the plot area that falls below the curve. AUC condenses the performance of all potential cut scores into a single measure of discrimination, and values of 0.80 or above generally indicate that the cut scores distinguish well between on track and off track test performances. All four AUC values exceed 0.80.

An optimal cut score maximizes the true positive rate and minimizes the false positive rate. A perfect classifier would sit at the upper left corner of the graph, with a true positive rate of 1.00 and a false positive rate of 0.00. Because perfect classifiers do not exist in practice, we used Euclidean distance to find the point on each curve closest to that corner. Equation 3.1 gives the distance D for each possible cut score, and we chose the cut score

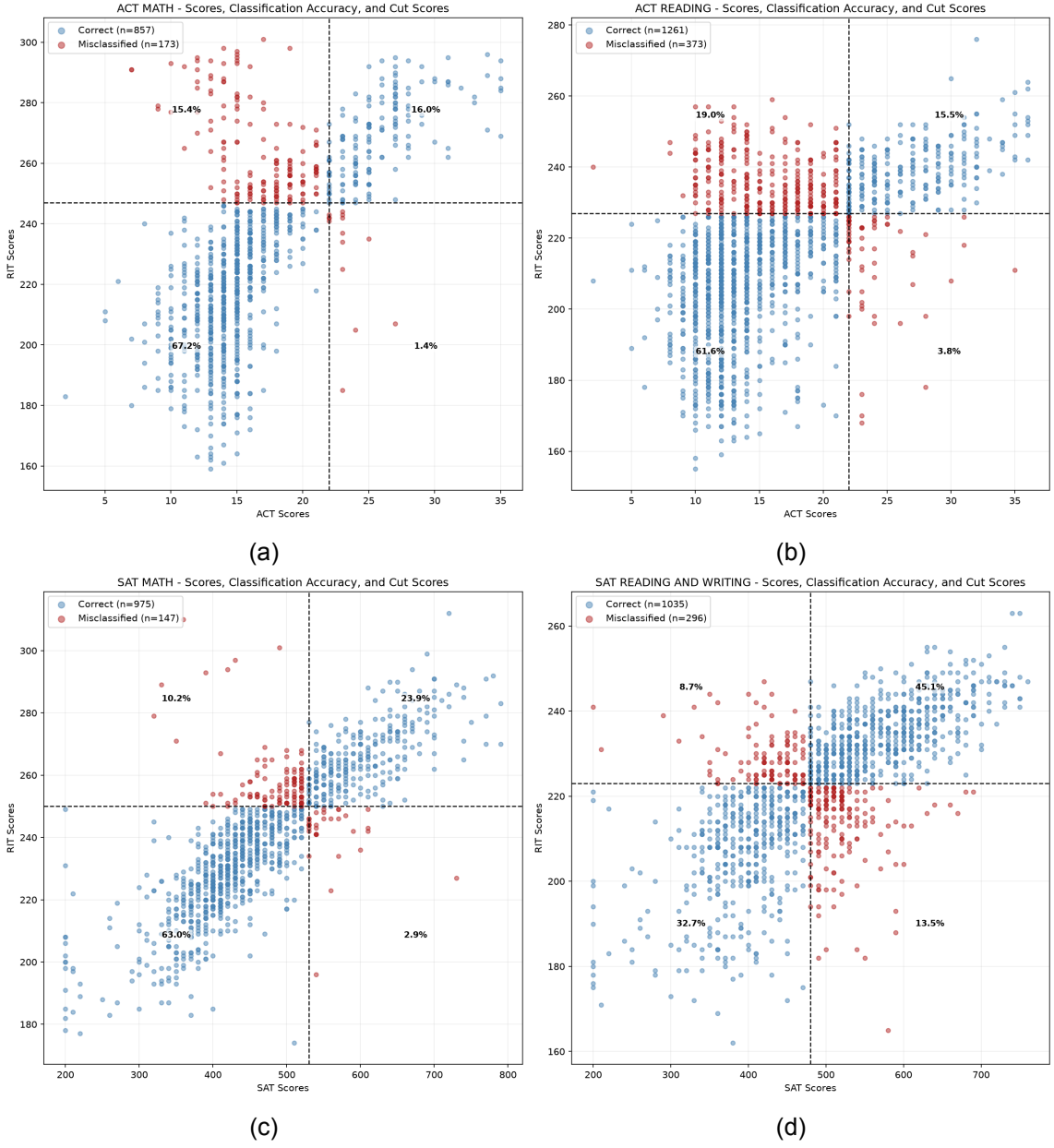

Figure 3.1.: Scatter Plots of Correctly and Incorrectly Classified Students in the Calibration Sample

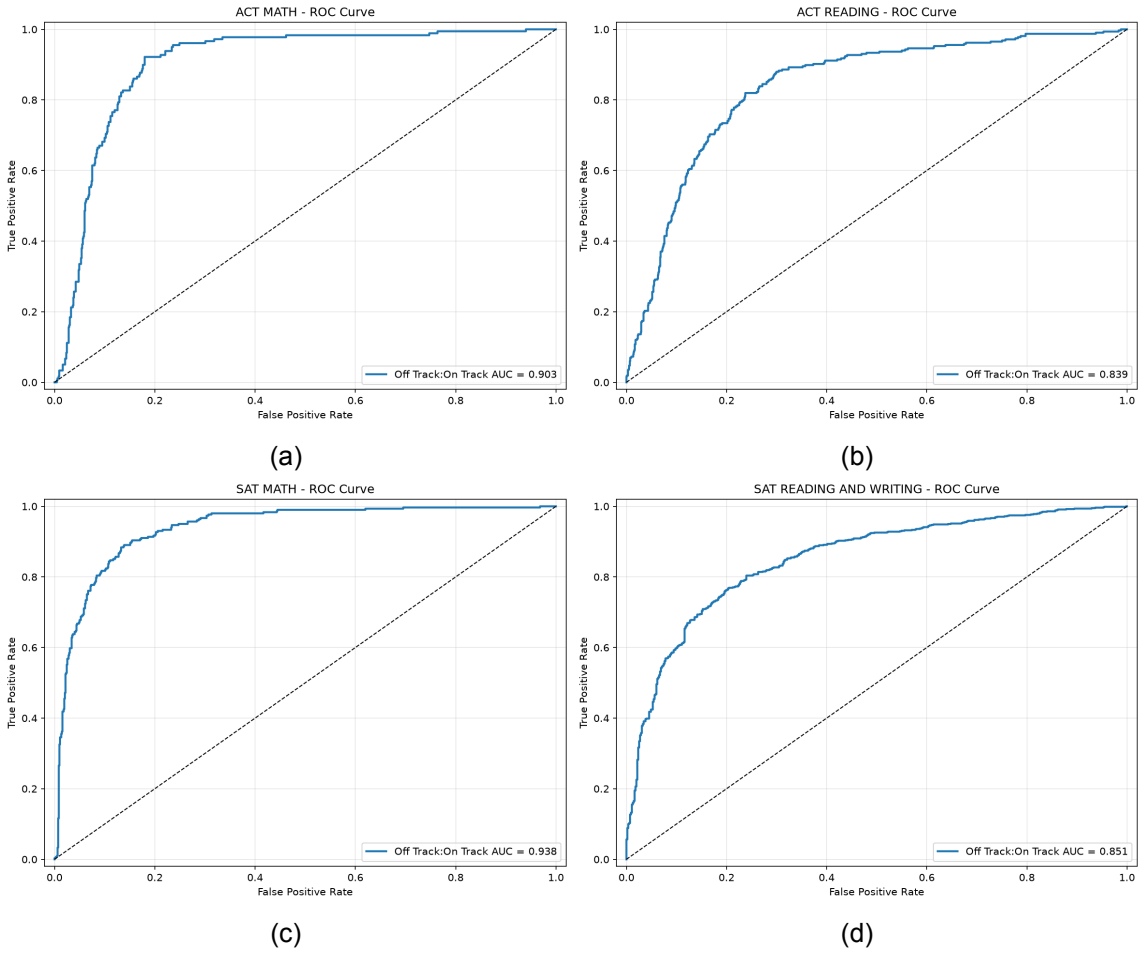

Figure 3.2.: ROC Curves for MAP Growth Grade 11 College Readiness Cut Scores

Table 3.2.: MAP Growth Grade 11 Spring College Readiness Classification Metrics

	Math		Reading	
Metric	ACT	SAT	ACT	SAT
Sensitivity	0.922	0.890	0.804	0.769
Specificity	0.813	0.861	0.764	0.789
Positive Predictive Value	0.509	0.702	0.450	0.838
Negative Predictive Value	0.980	0.955	0.942	0.707
Accuracy	0.832	0.869	0.772	0.778

with the smallest D as the college readiness benchmark for each MAP Growth test and college entrance exam combination.

$$D = \sqrt{\text{FPR}^2 + (1 - \text{TPR})^2} \tag{3.1}$$

where

- FPR is the false positive rate, plotted on the x-axis, and
- TPR is the true positive rate, plotted on the y-axis.

Table 3.2 reports the sensitivity and specificity of the cut scores established using this methodology, along with three further metrics. *Sensitivity* is the true positive rate, and *specificity* is the true negative rate, equal to one minus the false positive rate. *Positive predictive value (PPV)*, also referred to as precision, is the proportion of students labeled on track who actually scored at or above the college readiness threshold on the ACT or SAT. *Negative predictive value (NPV)* is the proportion of students labeled off track who actually scored below it. *Accuracy* is the proportion of all students correctly labeled as on or off track.

Across the four benchmarks, sensitivity ranged from 0.77 to 0.92, specificity from 0.76 to 0.86, and accuracy from 0.77 to 0.87. Because the Euclidean distance criterion weights sensitivity and specificity equally, each cut score balances the two. Unlike the predictive values, these metrics do not depend on how many students met the college readiness criterion.

The proportion of students who actually met the college readiness criterion differed substantially across the four benchmarks, from 17% for ACT Math to 59% for SAT Reading and Writing. This proportion drives PPV and NPV. When few students meet the criterion, those who did not far outnumber those who did, so even a modest false positive rate yields about

as many false positives as there are true positives. In the ACT Math sample, applying a true positive rate of 0.92 to the 17% who met the threshold and a false positive rate of 0.19 to the 83% who did not yielded almost equal numbers of true and false positives, giving a PPV of 0.51.

PPV was therefore lowest for the two ACT benchmarks (0.51 for math and 0.45 for reading) and highest for SAT Reading and Writing (0.84), while NPV ran the opposite way (0.98 and 0.94 for the ACT benchmarks and 0.71 for SAT Reading and Writing). These differences reflect how demanding each college entrance exam criterion is rather than how well MAP Growth cut scores discriminate; AUC ranged from 0.84 to 0.94 across the four benchmarks. For the ACT benchmarks, an off track label is therefore rarely wrong while an on track label is correct about half the time, and for SAT Reading and Writing the pattern reverses.

Because each cut score was selected to minimize D within the calibration sample, Table 3.2 summarizes classification performance in the same sample used to establish the cut scores. These values are therefore expected to be somewhat higher than those obtained when the benchmarks are applied to a new sample.

As noted in Section 2.1.2, schools in the calibration sample differed from the population on several demographic dimensions, including enrollment, urbanicity, and Mountain region representation. To evaluate whether these differences affected the cut scores, we conducted a stability analysis comparing cut scores and classification metrics across demographic subgroups within the calibration sample using bootstrap resampling. The analysis found that cut scores and classification metrics remained consistent across nearly all demographic comparisons, with 95% confidence intervals overlapping substantially. One exception emerged for ACT Math in Mountain region schools, where the estimated cut score of 253 was notably higher than both the full-sample cut score of 247 and the cut score for non-Mountain schools, though classification metrics remained comparable across groups. The overlapping confidence intervals in the remaining comparisons indicate that observed differences could reasonably be attributed to sampling variability rather than systematic differences in the relationship between MAP Growth and college entrance exam performance, supporting the validity of applying these benchmarks to schools with varying demographic characteristics. Full details of the stability analysis, including bootstrap methodology and results for all demographic comparisons, appear in Appendix A.

3.2. Setting Grade 6 to 10 Cut Scores

We set cut scores for Grades 6 through 10 by backward projection through the MAP Growth conditional growth norms. This gives the Grade 6 through 10 benchmarks a different basis than the Grade 11 benchmarks: the Grade 11 cut scores classify students against observed

ACT and SAT performance, whereas the lower-grade cut scores identify the scores from which students reach that Grade 11 threshold under typical growth. Backward projection inverts the usual growth question. Instead of asking what score follows a given starting score after typical growth, it asks what starting score would reach a known later score under typical growth. Equation 3.2 gives that starting score at the 50th conditional growth percentile.

$$y_{w_1} = \frac{y_{w_2} - \mu_\Delta + a\hat{\mu}_{w_1}}{a+1} \tag{3.2}$$

where

- w_1 and w_2 denote the week of instruction in the projected (earlier) grade and the known (later) grade, respectively;
- y_{w_1} is the backward-projected MAP Growth score for week w_1 based on the 50th conditional growth percentile;
- y_{w_2} is the known MAP Growth score for week w_2;
- μ_Δ is the marginal growth mean, the expected score difference between w_1 and w_2 (Equation 3.4);
- $\hat{\mu}_{w_1}$ is the marginal mean MAP Growth score at week w_1 (Equation 3.4); and
- a is the ratio of marginal growth covariance terms defined in Equation 3.6.

We applied Equation 3.2 recursively, starting from the Grade 11 spring benchmarks established in Section 3.1. Those benchmarks were projected back to Grade 10 spring, the resulting Grade 10 cut scores were projected back to Grade 9 spring, and the process continued until all cut scores reported in Table 1.1 were established.

Each projection in that sequence drew on the 2025 norms model coefficients and a design matrix to calculate the marginal growth means and covariances. The week of instruction was fixed at 32 for spring cut scores, and coefficients came from the model for the grade being projected rather than the grade already known. Grade 10 projections, for example, used Grade 10 model coefficients, the known Grade 11 cut scores, and the design matrix in Equation 3.3.

$$\mathbf{X}_j = \mathbf{Z}_j = \begin{bmatrix} 1 & w_1 & 0 & 0 \\ 1 & 0 & 1 & w_2 \end{bmatrix} = \begin{bmatrix} 1 & 32 & 0 & 0 \\ 1 & 0 & 1 & 32 \end{bmatrix} \tag{3.3}$$

3.2.1. Deriving Backward Conditional Growth Projections

The derivation begins with two elements of the MAP Growth conditional growth norms (see 2025 MAP Growth Norms Technical Manual, 2025, ch. 3.3): the marginal growth mean vector (Equation 3.4) and the marginal growth covariance matrix (Equation 3.5). From these, the conditional growth percentile is set to the median, both sides of the resulting expression are rewritten in known quantities, and the expression is solved for the earlier score.

$$\hat{\mu}_{\Delta} = [\hat{\mu}_{w_1},\ \hat{\mu}_{w_2} - \hat{\mu}_{w_1}]' \tag{3.4}$$

$$\hat{\Omega}_{\Delta} = \begin{bmatrix} \hat{\omega}_{y_{w_1}} & \hat{\omega}_{y_{w_1},\Delta(\mathrm{y})} \\ \hat{\omega}_{\Delta(\mathrm{y}),y_{w_1}} & \hat{\omega}_{\Delta(\mathrm{y})} \end{bmatrix} \tag{3.5}$$

The second element of Equation 3.4 is the marginal growth mean, $\mu_{\Delta} = \hat{\mu}_{w_2} - \hat{\mu}_{w_1}$, the expected score difference between w_1 and w_2. The ratio a is formed with elements from the marginal growth covariance matrix (Equation 3.5).

$$a = \left(\frac{\hat{\omega}_{y_{w_1},\Delta(\mathrm{y})}}{\hat{\omega}_{y_{w_1}}} \right) \tag{3.6}$$

Treating the elements of Equations 3.4 through 3.6 as known, Equation 3.7 gives the conditional growth percentile.

$$z_{\Delta} = \frac{\Delta(\mathrm{y}) - \hat{\mu}_{\Delta}(y_{w_1})}{\sqrt{\hat{\omega}_{\Delta}(y_{w_1})}} \tag{3.7}$$

Because $z_{\Delta} = 0$ for the 50th conditional growth percentile, Equation 3.7 simplifies to Equation 3.8.

$$\hat{\mu}_{\Delta}(y_{w_1}) = \Delta(\mathrm{y}) \tag{3.8}$$

Both sides of Equation 3.8 can be written in known quantities. $\hat{\mu}_{\Delta}(y_{w_1})$ is the growth mean conditioned on the value of the first RIT score (2025 MAP Growth Norms Technical Manual, 2025, ch. 3.3.2), and $\Delta(\mathrm{y})$ is the difference between scores from two test occasions.

$$\hat{\mu}_{\Delta}(y_{w_1}) = \mu_{\Delta} + \left(\frac{\hat{\omega}_{y_{w_1},\Delta(y)}}{\hat{\omega}_{y_{w_1}}} \right) \times \ (y_{w_1} - \hat{\mu}_{w_1}) \tag{3.9}$$

$$\Delta(y) = y_{w_2} - y_{w_1} \tag{3.10}$$

Substituting Equations 3.9 and 3.10 into Equation 3.8, and substituting a for the ratio of marginal growth covariance terms results in Equation 3.11.

$$\mu_{\Delta} \ + \ a(y_{w_1} - \hat{\mu}_{w_1}) = y_{w_2} - y_{w_1} \tag{3.11}$$

Moving the known quantities to the right side and collecting the y_{w_1} terms yields Equation 3.2, which returns the MAP Growth score at the earlier week from which a student growing at the 50th conditional growth percentile would reach the known later score. Each Grade 6 through 10 benchmark therefore marks the score from which typical growth reaches the Grade 11 college readiness threshold.

3.2.2. Conditional Growth Normative Patterns for High-performing Students

In the MAP Growth norms, spring-to-spring growth declines as starting achievement rises, and in the upper grades it turns negative at the top of the distribution. The aggregate pattern is visible in the achievement norms, where mean spring reading scores rise steadily through Grade 8 and then change little through Grade 12 (2025 MAP Growth Norms Technical Manual, 2025, app. A). Within grades it is concentrated among higher performing students: students in Grade 8 at or above the 75th achievement percentile tend to score two or three points lower the following spring, and the decline steepens through Grade 10. This reflects normative growth in the U.S. student population rather than any property of the MAP Growth tests or the college readiness benchmarks.

Because the college readiness cut scores sit at the high end of the achievement distribution, this normative pattern shapes the cut scores reported in Table 1.1. Math cut scores rise from Grade 6 to Grade 8 and are essentially unchanged from Grade 8 to Grade 11. Reading ACT cut scores decline across grades, most steeply between Grades 8 and 11, while reading SAT cut scores are essentially unchanged. Where growth at the top of the distribution is flat or negative, a student must be higher in the distribution in earlier grades to remain at the same standard later. The ACT reading benchmark therefore falls at the 93rd achievement percentile in Grade 6 and the 69th in Grade 11, and all four benchmarks sit at successively lower percentiles as grade increases.

4. Validation

The Grade 11 benchmarks were evaluated in Section 3.1 against students' observed ACT and SAT performance. This chapter evaluates the Grade 6 to 10 benchmarks, which were not derived from observed college entrance exam performance but were projected backward from the Grade 11 thresholds. We implemented a cohort design to compare each benchmark against students' observed MAP Growth performance one, two, and three grades later, in the validation sample of more than 7,000 schools described in Section 2.2. Because that sample is independent of the one used to establish the benchmarks and the comparisons run forward in time, these results describe how the benchmarks perform when applied to new students. Results are reported in Tables 4.1 and 4.2.

Across all 48 comparisons, AUC ranged from 0.81 to 0.94 in 47 cases and was 0.799 in the remaining case (ACT Reading, Grade 8 to Grade 11). Because AUC does not depend on where a cut score falls or on how many students meet it, this indicates that the benchmarks order students better than chance with respect to later performance in every subject, exam, and projection interval examined.

Classification accuracy declined as the projection interval lengthened, as expected when more time separates the benchmark from the performance it anticipates. For mathematics the decline was slight. Accuracy ranged from 0.83 to 0.91 one grade ahead and from 0.82 to 0.85 three grades ahead, and negative predictive values remained at or above 0.81 throughout. For reading the decline was steeper, from 0.76 to 0.93 one grade ahead to 0.65 to 0.75 three grades ahead, with negative predictive values falling to between 0.62 and 0.74.

Much of the reading pattern followed from the normative growth described in Section 3.2.2 rather than from the benchmarks' ability to discriminate. Because reading cut scores fall from the 93rd achievement percentile in Grade 6 to the 69th in Grade 11, a student who holds the same position in the distribution moves from off track to on track without any change in relative standing. The validation data show this directly. The share of students labeled on track rose by an average of 29 percentage points over three grades in reading, compared with 12 points in mathematics. Students who crossed the threshold in this way were counted as misclassifications, which lowered negative predictive value and accuracy while leaving discrimination intact.

Table 4.1.: Classification Accuracy of Benchmarks for MAP Growth Mathematics

Grade		ACT				SAT			
T1	T2	PPV	NPV	Acc.	AUC	PPV	NPV	Acc.	AUC
6	7	0.80	0.92	0.90	0.94	0.79	0.94	0.91	0.94
6	8	0.82	0.89	0.87	0.91	0.80	0.91	0.89	0.91
6	9	0.81	0.84	0.83	0.86	0.80	0.86	0.85	0.85
7	8	0.82	0.91	0.89	0.93	0.79	0.93	0.91	0.93
7	9	0.84	0.85	0.85	0.87	0.82	0.87	0.87	0.86
7	10	0.86	0.81	0.82	0.85	0.86	0.83	0.83	0.84
8	9	0.80	0.89	0.87	0.89	0.78	0.90	0.88	0.88
8	10	0.84	0.84	0.84	0.86	0.84	0.85	0.85	0.85
8	11	0.85	0.81	0.82	0.85	0.85	0.83	0.83	0.84
9	10	0.80	0.87	0.85	0.88	0.80	0.88	0.87	0.88
9	11	0.83	0.81	0.81	0.85	0.82	0.83	0.83	0.85
10	11	0.77	0.85	0.83	0.86	0.75	0.87	0.84	0.86

Table 4.2.: Classification Accuracy of Benchmarks for MAP Growth Reading

Grade		ACT				SAT			
T1	T2	PPV	NPV	Acc.	AUC	PPV	NPV	Acc.	AUC
6	7	0.80	0.94	0.93	0.94	0.83	0.84	0.84	0.91
6	8	0.90	0.86	0.86	0.89	0.88	0.75	0.79	0.89
6	9	0.95	0.74	0.75	0.84	0.90	0.63	0.70	0.85
7	8	0.80	0.90	0.89	0.91	0.85	0.81	0.82	0.90
7	9	0.91	0.77	0.78	0.85	0.89	0.68	0.75	0.86
7	10	0.94	0.67	0.69	0.81	0.88	0.62	0.70	0.83
8	9	0.83	0.84	0.84	0.87	0.85	0.76	0.79	0.88
8	10	0.89	0.72	0.74	0.83	0.86	0.68	0.74	0.84
8	11	0.92	0.62	0.65	0.80	0.85	0.62	0.69	0.81
9	10	0.80	0.81	0.81	0.86	0.82	0.76	0.79	0.87
9	11	0.87	0.68	0.72	0.82	0.83	0.67	0.74	0.83
10	11	0.81	0.74	0.76	0.83	0.81	0.72	0.76	0.83

The errors that remained fell in the more conservative direction. Reading benchmarks overidentified students as off track over long intervals, and their positive predictive values rose with the projection interval, from 0.80 to 0.85 one grade ahead to 0.85 to 0.95 three grades ahead. Setting classification thresholds requires weighing the relative consequences of false positive and false negative classifications (Metz, 1978). In education terms, a false positive may leave a student without additional support because they were characterized as on track when they were not, whereas a false negative directs support toward a student who may not have required it. The benchmarks are therefore most precise where the consequences of error are greatest, and least precise where the cost of overidentification is the provision of support a student may not need.

References

2025 MAP Growth Norms Technical Manual (2025.1.1). (2025). HMH Education Company. https://www.nwea.org/resource-center/white-paper/88182/MAP-Growth-Norms-Technical-Manual-1.pdf/

Allen, J. (2013). *Updating the ACT college readiness benchmarks*. ACT, Inc. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2013-6.pdf

CollegeBoard. (n.d.). *K-12 educator brief: The college and career readiness benchmarks for the SAT suite of assessments*. CollegeBoard. Retrieved August 5, 2026, from https://satsuite.collegeboard.org/media/pdf/educator-benchmark-brief.pdf

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

NCES. (2024). *Public Elementary/Secondary School Universe Survey Data*. https://nces.ed.gov/ccd/

Thum, Y. M. (2017). *MAP Growth College Readiness Benchmarks: An Addendum with Preliminary Results Keyed on the SAT* [NWEA Research Report]. NWEA.

Thum, Y. M., & Matta, T. (2015). *MAP College Readiness Benchmarks: A Research Brief* [NWEA Research Report]. NWEA. https://www.nwea.org/uploads/2020/10/MAP-College-Readiness-Benchmarks-Research-Brief.pdf

Zviedrite, N., Jahan, F., Moreland, S., Ahmed, F., & Uzicanin, A. (2024). COVID-19–Related School Closures, United States, July 27, 2020–June 30, 2022. *Emerging Infectious Diseases*, *30*(1), 58–69. https://doi.org/10.3201/eid3001.231215

A. Grade 11 Cut Score Stability Analysis

College readiness cut scores for Grade 11 were established using a convenience sample of schools that agreed to share data rather than a random sample of schools. As noted in Section 2.1.2, schools in the calibration sample differed from the population across several demographic dimensions including urbanicity, enrollment, and Mountain region representation. To evaluate whether these differences affected the cut scores, we conducted a stability analysis that compared cut scores and classification metrics across demographic subgroups within the calibration sample.

A.1. Methodology

The stability analysis examined whether the relationship between MAP Growth scores and college entrance exam performance varied systematically across school characteristics. We focused on three demographic dimensions along which the calibration sample differed most substantially from the population: city locale, enrollment above the median, and Mountain region representation. For each dimension, we compared schools in one demographic category (e.g., city schools) to all other schools in the sample (e.g., non-city schools). This produced three demographic comparisons, each conducted separately for the four college entrance exam scales: ACT Math, ACT Reading, SAT Math, and SAT Reading and Writing. One analysis was dropped, SAT Math comparing city to non-city schools, due to insufficient data, yielding 11 total analyses.

The analysis employed bootstrap resampling to account for the relatively small calibration sample and to quantify uncertainty in the estimated cut scores and classification metrics. For each demographic comparison and college entrance exam scale combination, we calculated point estimates for cut scores and classification metrics from the subsample of the full calibration data set corresponding to each demographic group. Then we constructed 95% confidence intervals for these metrics by drawing 10,000 bootstrap samples with replacement from the original data, stratified by demographic group. All bootstrap samples were drawn using a fixed random seed to ensure reproducibility.

Within each bootstrap sample, we followed a two step process. First, we applied the same ROC curve methodology described in Section 3.1 to identify the optimal cut score.

Specifically, we constructed an empirical ROC curve by calculating the true positive and false positive rates at every possible MAP Growth cut score, then selected the cut score that minimized Euclidean distance. This procedure generated a distribution of 10,000 cut scores for each demographic group. Second, we calculated classification metrics (e.g., sensitivity) for each bootstrap sample using the cut score identified in the first step. The 10,000 samples formed the bootstrap distributions that were used to construct 95% confidence intervals for cut scores and classification metrics, which allowed us to evaluate whether classification performance differed across demographic groups.

The bootstrap approach addresses two sources of uncertainty. First, it accounts for sampling variability in the calibration data, providing confidence intervals that reflect the precision of the cut score estimates given the available sample size. Second, by resampling within demographic groups, it allows direct comparison of cut scores and classification metrics across groups while preserving the structure of the data. If the confidence intervals for cut scores from two demographic groups overlap substantially, this indicates that any observed differences could reasonably be attributed to sampling variability rather than systematic differences in the relationship between 11th grade MAP Growth RIT scores and college entrance exam performance criteria.

A.2. Results

Tables A.1 through A.4 present the cut scores and classification metrics for each demographic comparison. Each table reports point estimates with 95% bootstrap confidence intervals in parentheses. The cut scores shown in the table captions are those established from the full calibration sample as reported in Table 1.1.

The stability analysis indicated that cut scores and classification metrics remained consistent across demographic subgroups, with confidence intervals overlapping substantially in nearly all comparisons. For ACT Math (Table A.1), cut scores ranged from 235 to 253 across subgroups, with all confidence intervals except that for Mountain region schools encompassing the full-sample cut score of 247 (see below for details). Sensitivity ranged from 0.80 to 0.89, specificity from 0.80 to 0.84, and accuracy from 0.80 to 0.85. Similar patterns emerged for ACT Reading (Table A.2), where cut scores ranged from 223 to 231, sensitivity from 0.75 to 0.92, specificity from 0.75 to 0.85, and accuracy from 0.75 to 0.82.

For SAT Math (Table A.3), the city locale comparison was excluded due to insufficient sample size (fewer than 100 students in city schools). Among the remaining comparisons, cut scores ranged from 250 to 251, with overlapping confidence intervals. Sensitivity ranged from 0.85 to 0.89, specificity from 0.85 to 0.90, and accuracy from 0.86 to 0.89. SAT

Table A.1.: Stability of MAP Growth Benchmark for ACT Math (Cut Score = 247)

	City		Enrollment		Mountain	
Estimate	Y	N	Y	N	Y	N
Cut Score	250 (247, 253)	242 (235, 248)	253 (247, 253)	235 (235, 248)	253 (250, 257)	242 (235, 248)
Sens.	0.89 (0.85, 0.95)	0.80 (0.67, 1.00)	0.85 (0.83, 0.94)	0.87 (0.67, 1.00)	0.85 (0.79, 0.91)	0.81 (0.68, 0.95)
Spec.	0.83 (0.79, 0.87)	0.82 (0.72, 0.91)	0.83 (0.77, 0.86)	0.80 (0.76, 0.91)	0.84 (0.79, 0.90)	0.82 (0.74, 0.88)
PPV	0.55 (0.49, 0.63)	0.29 (0.15, 0.47)	0.61 (0.53, 0.66)	0.17 (0.11, 0.34)	0.76 (0.70, 0.84)	0.14 (0.08, 0.21)
NPV	0.97 (0.96, 0.99)	0.98 (0.96, 1.00)	0.95 (0.94, 0.98)	0.99 (0.98, 1.00)	0.91 (0.87, 0.94)	0.99 (0.98, 1.00)
Acc.	0.84 (0.81, 0.87)	0.82 (0.73, 0.90)	0.84 (0.80, 0.86)	0.80 (0.77, 0.91)	0.85 (0.81, 0.88)	0.82 (0.74, 0.87)

Reading and Writing (Table A.4) showed cut scores ranging from 221 to 227, sensitivity from 0.71 to 0.81, specificity from 0.79 to 0.86, and accuracy from 0.76 to 0.85.

One notable exception emerged in the ACT Math analysis for schools in the Mountain region. The estimated cut score for Mountain region schools was 253 (95% CI [250, 257]), notably higher than both the full-sample cut score of 247 and the cut score for non-Mountain schools of 242 (95% CI [235, 248]). The confidence interval for Mountain region schools does not include the full-sample cut score, indicating that this difference cannot be attributed solely to sampling variability. Despite this difference in cut scores, classification metrics remained comparable across groups: sensitivity was 0.85 for Mountain schools versus 0.81 for non-Mountain schools, specificity was 0.84 versus 0.82, and accuracy was 0.85 versus 0.82. The higher cut score in Mountain region schools suggests that students in these schools may require higher MAP Growth scores to achieve the same probability of meeting the ACT Math college readiness threshold, though the relatively small subgroup sizes (421 students in Mountain schools versus 609 in non-Mountain schools) limit the precision of this estimate.

Apart from that exception, the overlapping confidence intervals across demographic groups indicate that observed differences in cut scores and classification metrics could reasonably be attributed to sampling variability rather than systematic differences in the relationship between MAP Growth and college entrance exam performance. This finding supports the validity of applying the benchmarks established from the full calibration sample to schools with varying demographic characteristics. The insufficient sample size for city schools in the SAT Math analysis reflects the concentration of SAT-taking students in non-city schools within the calibration sample rather than any fundamental limitation of the benchmarking methodology.

A. Grade 11 Cut Score Stability Analysis

Table A.2.: Stability of MAP Growth Benchmark for ACT Reading (Cut Score = 227)

	City		Enrollment		Mountain	
Estimate	Y	N	Y	N	Y	N
Cut Score	227 (224, 228)	223 (219, 233)	227 (223, 230)	228 (225, 229)	231 (224, 232)	227 (223, 228)
Sens.	0.81 (0.75, 0.86)	0.92 (0.77, 1.00)	0.82 (0.78, 0.92)	0.76 (0.70, 0.84)	0.76 (0.71, 0.90)	0.75 (0.72, 0.85)
Spec.	0.75 (0.70, 0.80)	0.78 (0.70, 0.96)	0.79 (0.72, 0.84)	0.74 (0.68, 0.79)	0.85 (0.71, 0.89)	0.78 (0.70, 0.81)
PPV	0.46 (0.41, 0.51)	0.27 (0.17, 0.62)	0.50 (0.43, 0.57)	0.39 (0.32, 0.46)	0.74 (0.62, 0.81)	0.35 (0.28, 0.40)
NPV	0.94 (0.92, 0.95)	0.99 (0.97, 1.00)	0.95 (0.93, 0.97)	0.94 (0.92, 0.96)	0.86 (0.83, 0.93)	0.95 (0.94, 0.97)
Acc.	0.76 (0.73, 0.80)	0.79 (0.72, 0.95)	0.80 (0.76, 0.83)	0.75 (0.70, 0.78)	0.82 (0.76, 0.85)	0.77 (0.71, 0.80)

Table A.3.: Stability of MAP Growth Benchmark for SAT Math (Cut Score = 250)

	Enrollment		Mountain	
Estimate	Y	N	Y	N
Cut Score	250 (249, 252)	251 (242, 256)	251 (243, 253)	250 (249, 252)
Sens.	0.89 (0.84, 0.92)	0.89 (0.80, 1.00)	0.85 (0.71, 1.00)	0.89 (0.85, 0.92)
Spec.	0.85 (0.83, 0.90)	0.90 (0.79, 0.94)	0.87 (0.71, 0.93)	0.86 (0.84, 0.91)
PPV	0.72 (0.68, 0.80)	0.55 (0.35, 0.70)	0.35 (0.16, 0.53)	0.73 (0.69, 0.80)
NPV	0.95 (0.92, 0.96)	0.98 (0.97, 1.00)	0.98 (0.97, 1.00)	0.95 (0.93, 0.96)
Acc.	0.86 (0.84, 0.89)	0.89 (0.80, 0.94)	0.87 (0.73, 0.92)	0.87 (0.85, 0.90)

Note. City analysis excluded due to fewer than 100 students in schools located in a city.

Table A.4.: Stability of MAP Growth Benchmark for SAT Reading and Writing (Cut Score = 223)

	City		Enrollment		Mountain	
Estimate	Y	N	Y	N	Y	N
Cut Score	227 (219, 234)	223 (221, 224)	223 (221, 225)	223 (221, 228)	221 (217, 228)	223 (221, 225)
Sens.	0.81 (0.67, 1.00)	0.77 (0.73, 0.82)	0.77 (0.73, 0.83)	0.75 (0.65, 0.85)	0.71 (0.63, 0.91)	0.78 (0.73, 0.82)
Spec.	0.86 (0.70, 0.98)	0.79 (0.74, 0.83)	0.79 (0.74, 0.83)	0.79 (0.71, 0.91)	0.79 (0.62, 0.90)	0.78 (0.74, 0.83)
PPV	0.68 (0.43, 0.93)	0.85 (0.82, 0.87)	0.85 (0.82, 0.88)	0.75 (0.66, 0.88)	0.63 (0.47, 0.79)	0.85 (0.82, 0.88)
NPV	0.93 (0.85, 1.00)	0.69 (0.66, 0.74)	0.69 (0.65, 0.74)	0.79 (0.72, 0.88)	0.84 (0.79, 0.95)	0.69 (0.65, 0.74)
Acc.	0.85 (0.73, 0.95)	0.78 (0.76, 0.80)	0.78 (0.76, 0.81)	0.77 (0.72, 0.85)	0.76 (0.68, 0.85)	0.78 (0.76, 0.80)