COMPARATIVE ANALYSIS OF PACE 2018 MODIFIED and ORIGINAL CUTSCORE METHOD

Center for Assessment

January 22, 2019

Table of Contents

Introduction ..................................................................................................................................... 4
Standard Setting Methodology .................................................................................................. 4
Issues with the 2017-2018 PACE Standard Setting ................................................................. 5
Modified Standard Setting Approach for 2017-2018 ............................................................. 6
Issues and Concerns ................................................................................................................ 6
Investigating the 2017-2018 Results ....................................................................................... 7
Results/Discussion ...................................................................................................................... 8
Cohort Analysis ...................................................................................................................... 8
Longitudinal Analysis ............................................................................................................... 10
State Test Analysis ................................................................................................................ 11
Conclusion ................................................................................................................................ 13
Appendix A: Amherst ................................................................................................................ 16
Cohort Analysis Graphs ............................................................................................................ 16
Longitudinal Analysis Graphs .................................................................................................. 20
State Test Analysis Graphs ....................................................................................................... 23
Discussion ................................................................................................................................. 24
Appendix B: Bethlehem ............................................................................................................. 25
Cohort Analysis Graphs ............................................................................................................ 25
State Test Analysis Graphs ....................................................................................................... 29
Discussion ................................................................................................................................. 30
Appendix C: Concord ................................................................................................................ 31
Cohort Analysis Graphs ............................................................................................................ 31
Longitudinal Analysis Graphs .................................................................................................. 36
State Test Analysis Graphs ................................................................. 41
Discussion ............................................................................................ 42
Appendix D: Epping ................................................................................ 43
  Cohort Analysis Graphs ....................................................................... 43
  Longitudinal Analysis Graphs .............................................................. 48
  State Test Analysis Graphs ................................................................. 53
  Discussion ........................................................................................... 54
Appendix E: Laconia ............................................................................... 55
  Cohort Analysis Graphs ....................................................................... 55
  Longitudinal Analysis Graphs .............................................................. 58
  State Test Analysis Graphs ................................................................. 61
  Discussion ........................................................................................... 62
Appendix F: Newport ............................................................................. 63
  Cohort Analysis Graphs ....................................................................... 63
  Longitudinal Analysis Graphs .............................................................. 66
  State Test Analysis Graphs ................................................................. 68
  Discussion ........................................................................................... 69
Appendix G: Pittsfield ............................................................................ 70
  Cohort Analysis Graphs ....................................................................... 70
  Longitudinal Analysis Graphs .............................................................. 75
  State Test Analysis Graphs ................................................................. 80
  Discussion ........................................................................................... 81
Appendix H: Plymouth ........................................................................... 82
  Cohort Analysis Graphs ....................................................................... 82
  Longitudinal Analysis Graphs .............................................................. 87
  State Test Analysis Graphs ................................................................. 93
  Discussion ........................................................................................... 94
Appendix I: Rochester ............................................................................ 95
  Cohort Analysis Graphs ....................................................................... 95
  Longitudinal Analysis Graphs .............................................................. 100
  State Test Analysis Graphs ................................................................. 105
Discussion ................................................................................................................................................. 106
Appendix J: Sanborn ................................................................................................................................. 107
  Cohort Analysis Graphs ......................................................................................................................... 107
  Longitudinal Analysis Graphs ............................................................................................................. 111
  State Test Analysis Graphs .................................................................................................................. 116
  Discussion ............................................................................................................................................. 117
Appendix K: SAU23 ..................................................................................................................................... 118
  Cohort Analysis Graphs ......................................................................................................................... 118
  State Test Analysis Graphs .................................................................................................................. 123
  Discussion ............................................................................................................................................. 124
Appendix L: Seacoast Charter School ...................................................................................................... 125
  Cohort Analysis Graphs ......................................................................................................................... 125
  Longitudinal Analysis Graphs ............................................................................................................. 130
  State Test Analysis Graphs .................................................................................................................. 136
  Discussion ............................................................................................................................................. 137
INTRODUCTION

The PACE leaders and consultants have produced “comparable annual determinations” since the 2014-2015 school year. The methodology has worked well within the PACE context and has supported by the PACE technical advisory committee and the U.S. Department of Education. However, there were some issues and concerns raised about the achievement level reports produced for the 2017-2018 school year. This document presents information to try to better understand the issues associated with the PACE standard setting results for the 2017-2018 school year. The document starts with a very brief review of the basics of standard setting and summarizes the two standard setting methods used for PACE. The bulk of the document presents extensive analyses for each district, grade, and subject combinations.

Standard Setting Methodology

Determining the achievement levels on PACE (e.g., proficient, advanced) or any other test involves matching achievement (sometimes called performance) level descriptors with scores on a test or other distribution. Achievement level descriptors (ALDs) are the foundation of this process. There are many classes of methods used to connect ALDs with score distributions. Some methods focus on having panelists review and evaluate the difficulty of individual test items, while others have panelists review collections of student work or portfolios of scores. Still others rely on capturing expert judgments of examinees. We have been using an examinee-centered method called contrasting groups. This standard setting method involves using judgments from panelists (i.e. teachers) about the overall achievement of the examinees based on extensive knowledge of the examinee. To implement this method for the PACE pilot, we ask teachers at the end of the school year to make judgments about which achievement level best describes each of their students. This process relies heavily on a common understanding and interpretation of the achievement level descriptors (ALDs). The subject and grade specific ALDs are entered into an online survey where teachers can easily read the descriptions and match their students to the appropriate achievement level. The teacher judgment data is one component of the standard setting approach. These judgments must then be matched against the scores on competency-based assessments the students have completed throughout the year. One expects the students rated highest by their teachers to have the highest competency scores and the students rated lowest to have the lowest scores. That is generally true, but as one would expect, there is a fair amount of overlap.

The contrasting groups standard setting methodology then involves comparing the PACE scores with student placements into achievement levels in order to determine cut scores that would accurately classify the highest percentage of students into achievement levels.

Logistic regression is used to determine the point in the score distribution where examinees have a 50% chance of being classified in the next performance level or above (e.g., the probability that
a student is Level 3 or above is 50% at score X). A logistic regression analysis is run separately for each cut point—Level 2, Level 3, and Level 4—in each district, content area, and grade level. The figure below illustrates a hypothetical case for how logistic regression is used to estimate a cutscore between two distributions. Unfortunately, the real world is rarely as clean as the picture below, but the method has worked well through the years.

The participating PACE districts all rely on unique competencies and fairly unique assessments except for the PACE common task and perhaps reused PACE common tasks to make local competency determinations for students. Therefore, the logistic regression analyses are computed for every district/grade/subject combination.

**Issues with the 2017-2018 PACE Standard Setting**

Several new school districts and parts of school districts were included in PACE accountability system during the 2017-2018 school year. The assumptions underlying most statistical methods are challenged when dealing with small sample sizes. This challenge was significantly compounded by teacher judgment data that had considerably less variance than in previous years. The PACE achievement levels have always been more constrained than with tests like Smarter Balanced or NH SAS that are designed to spread the distribution of students, but we observed noticeably more spread in judgments in previous years. Some of the most egregious cases this year involved classes with all 20 or so students receiving the same rating (e.g., “3”). We are pretty sure this cannot be accurate. This example is an extreme case, but the combination of
generally more constrained distributions and smaller sample sizes made us less confident in the method we had been using since 2014-2015.

This played out by having too many cases where the logistic regression was unable to provide any estimate of a particular cutscore. In general, the cutscore between levels 2 and 3 was able to be estimated for almost all grade/subject/district combinations, but there were a substantial number of cases where the cutscores between levels 1 and 2 or between 3 and 4 could not be estimated. We are required to report student achievement in the same way (or as similar as possible) as the statewide assessment and in accordance with the ESSA technical requirements, which means that we need to produce three cutscores to yield four achievement levels. We observed very limited instances of these issues in the past and when we did, we “manually fit” a cutscore by, for example, taking half of the distance between the 2/3 cutscore and the highest possible score (e.g., 4.0) to produce a 3/4 cutscore. We have done this “manual fitting” in previous years with a very limited number of cases, but we became concerned when it looked like we might have to do this for many of the smaller schools.

**Modified Standard Setting Approach for 2017-2018**

Given these issues, we designed and implemented a new approach for establishing cutscores for the 2017-2018 school year. The small sample sizes were our biggest challenge so we designed an approach to ameliorate that concern. The modified approach involved standardizing all district/grade/subject scores such that the mean score for each district/grade/subject combination was 0 and the standard deviation was 1. This allowed us to pool data from all districts having that particular subject/grade combination (e.g., 7th grade math). These standardized scores (z-scores) allowed us to estimate all three cutscores for every grade/subject combination while maintaining the relative rigor of the grading practices in each district intact. We then translated the z-scores back to the metrics used in each district (e.g., 1-4 scale, 1-100 scale) for reporting purposes. We examined the results of this new method compared to the original method, but only did so in the aggregate (e.g., 7th grade math for all PACE districts) and, on average, things looked comparable and it seemed better than “manual fitting.”

**Issues and Concerns**

While the results appeared correct (again, on average), a call from a district leader at the end of November pointed out things did not work as well as it appeared on average. We first examined the results from typically high and low scoring districts and began to suspect that the pooling procedure generally pulled the high scorers down and boosted the lower scoring districts by depressing true differences in achievement across districts.
Investigating the 2017-2018 Results

We continued to investigate the results by using several analyses to evaluate the accuracy of the PACE 2018 modified (pooled) versus original cutscore methods:

- **Cohort analysis**: Examined how students in a given grade/subject in 2018 performed in comparison to students in the same grade/subject in that district for the previous year (2017) and any other years of data available;
- **Longitudinal analysis**: Compared how students in a given grade in 2018 performed in the previous grades (same subject) for the previous year (2017) and any other years of data available; and
- **State test analysis**: Compared proficiency rates between PACE and NH SAS in grades 3-8. All analyses were disaggregated by district, subject, and grade level in order to evaluate differences between the modified vs. original cutscore method at the lowest meaningful unit of analysis possible.

Results are compared using percent proficient or above, which means a student scored at Level 3 or 4. Data from the beginning of the PACE pilot in 2014-15 to the 2017-18 school year were used whenever available. Analyses focused on the PACE accountability grades 3-8 only, which include: Grades 4-7 ELA; Grades 3, 5-7 Math, and Grade 8 Science.

All district-level analyses are contained in the appendices and discussions of each district’s results on all three analyses are presented at the end of each district’s appendix. For example, Appendix A contains all of the results for Amherst School District’s cohort analysis, longitudinal analysis, and state test analysis using bar graphs and tables, if applicable. The last page of Appendix A has a paragraph synthesizing across the analyses conducted to evaluate the accuracy of the PACE 2018 modified vs. original cutscore methods in Amherst. The raw results underlying the graphs presented in the Appendices and discussed in this report can be found in the Excel document, “PACE proficiency results 2015 to 2018.” Some small districts (e.g., Monroe) are not included in this report because low student numbers in each grade makes it so percent proficient or above results are not available (i.e., suppressed) on the NH DOE website.

In this report, red bars indicate that results are from the statewide achievement test, which was Smarter Balanced from 2015-2017 and NH SAS in 2018. Green bars indicate results from the PACE modified cutscore method used for the first time in 2018. The modified (pooled) method involves converting every district’s end-of-year competency scores by grade and subject into z-scores, pooling them together with all other districts’ end-of-year competency scores in the same grade and subject, applying the contrasting groups standard setting procedure, and then adjusting cut scores post hoc by district, grade, and subject using the results from the calibration analyses (i.e., consensus scores) during the PACE summer institute. The purple bars indicate results from the PACE original cutscore method used from 2015-2017 and then re-applied to the raw data in 2018 to compare results with the modified method. The original cutscore method used end-of-year competency scores and teacher judgment surveys, applied the contrasting groups standard
setting procedure, and then adjusted cut scores post hoc using the results from the calibration analyses during the PACE summer institute only if there was systematic evidence of district stringency or leniency in scoring.

**Results/Discussion**

**Cohort Analysis**

We compared performance (percent proficient or above) in a given district, subject, and grade level over time. For example, we compare the percent of students deemed proficient or above in each district, subject, and grade level over four years (from 2015 to 2018), based on available data. We would expect some instability in the results from year-to-year given that the cohort analysis includes different groups of students, but keeps the grade level/subject area constant. The two methods under analyses (modified vs. original method) can be directly compared by examining the extent to which the results for each district, subject, and grade level stays more consistent from 2015 to 2018.

Results suggest that effects of the modified vs. original cutscore method vary by district, grade, and subject. In some districts, grades, and subjects, the modified method seemed to drastically increase proficiency rates—defined as a 30% or greater increase in proficiency from 2017 results to 2018 modified results (e.g., Concord Math Gr 7; Epping ELA Gr 5; Laconia Math Gr 3; etc.). In other districts, grades, and subjects, the modified method seemed to drastically decrease proficiency defined as a 30% or greater decline in proficiency from 2017 to 2018 modified method (e.g., Amherst ELA Gr 5-7; Epping Math Gr 3 & 7; Pittsfield ELA Gr 6; etc.). Similarly, there are a few district, grade, and subject combinations where the original cutscore method appeared to support a drastic increase in proficiency (e.g., Pittsfield Math Gr 5; Plymouth ELA Gr 5) or decrease in proficiency (e.g., Plymouth Math Gr 7).

One way to compare the relative differences from year-to-year performance is to calculate the absolute value of the percent proficient or above differences in PACE grades/subjects between adjacent years for students in the same grade/subject and then sum up all the absolute values in any given adjacent year category. Absolute values remove the effect of the positive and negative sign and allow a comparison of the magnitude of difference across years. Table 1 portrays the sum of the absolute values of percent proficient or above differences between adjacent years for all districts and disaggregated by district. The differences in percent proficient or above using the original method are more similar to the 2015 to 2016 and 2016 to 2017 results for all districts combined because they are based on the same method. **The differences in percent proficient or above using the 2018 modified method are more than 50% greater in comparison to prior years than they would have been if we had used our original method.**
Table 1.
Sum of the Absolute Values of Percent Proficient or Above Differences between Adjacent Years for All Districts and Disaggregated by District

<table>
<thead>
<tr>
<th>District</th>
<th>2015 to 2016 Results</th>
<th>2016 to 2017 Results</th>
<th>2017 to 2018 Modified (Pooled) Method Results</th>
<th>2017 to 2018 Original Method Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Districts</td>
<td>831</td>
<td>814</td>
<td>1505</td>
<td>956</td>
</tr>
<tr>
<td>Amherst</td>
<td>22</td>
<td>43</td>
<td>151</td>
<td>52</td>
</tr>
<tr>
<td>Concord</td>
<td>95</td>
<td>45</td>
<td>101</td>
<td>55</td>
</tr>
<tr>
<td>Epping</td>
<td>99</td>
<td>115</td>
<td>223</td>
<td>99</td>
</tr>
<tr>
<td>Laconia</td>
<td>22</td>
<td>21</td>
<td>73</td>
<td>40</td>
</tr>
<tr>
<td>Newport</td>
<td>17</td>
<td>49</td>
<td>115</td>
<td>39</td>
</tr>
<tr>
<td>Pittsfield</td>
<td>99</td>
<td>131</td>
<td>130</td>
<td>103</td>
</tr>
<tr>
<td>Plymouth</td>
<td>112</td>
<td>110</td>
<td>135</td>
<td>165</td>
</tr>
<tr>
<td>Rochester</td>
<td>98</td>
<td>68</td>
<td>243</td>
<td>55</td>
</tr>
<tr>
<td>Sanborn</td>
<td>68</td>
<td>68</td>
<td>124</td>
<td>83</td>
</tr>
<tr>
<td>SAU 23</td>
<td>53</td>
<td>89</td>
<td>69</td>
<td>158</td>
</tr>
<tr>
<td>Seacoast</td>
<td>146</td>
<td>75</td>
<td>141</td>
<td>107</td>
</tr>
</tbody>
</table>

*Note.* Bethlehem and Monroe are not included because of small sample sizes and unavailable data.

The problem is that differences in percent proficient or above are not all in the same direction when examining the absolute value results by district as shown in Table 1. Some districts ‘benefited’ through use of the modified method because they saw an increase in percent proficient or above in some grade and subject combinations; whereas other districts did not benefit and saw a significant decline in percent proficient or above in some grade and subject combinations. They critical question remains: which results are more trustworthy? The next analyses aims to build upon the evidence supporting the trustworthiness of a particular method (modified vs. original method) through a longitudinal examination of student classes over time.
Longitudinal Analysis

The longitudinal analysis compares the performance of students over time within a particular district in the same subject area. For example, we compared the percent of students designated proficient or above in math for the graduating class of 2023 as they progressed in a district from Grade 4 in 2015 to Grade 7 in 2018. We would expect less instability in the results from year-to-year as compared to the cohort analysis given that the longitudinal analysis includes the same class of students over time. The two methods under analysis (modified vs. original cutscore method) can be directly compared by examining the extent to which the results for each district and subject area stay more consistent from 2015 to 2018.

Similar to the cohort analysis, results between the modified vs. original cutscore methods vary by district, subject, and student class (i.e., Class of 2023, Class of 2024, etc.). In some districts, grades, and subjects the modified method seemed to underestimate student achievement in comparison to longitudinal trends; whereas in other districts, grades, and subjects the opposite occurred. The same is true for the original method.

To compare the relative differences from year-to-year for the same group of students by subject area, we calculated the absolute value of the percent proficient or above differences in adjacent years by subject area and then compared the means and standard deviations for each subject area. We did not sum across the absolute values as in Table 1 because the number of comparisons varies across years and therefore could lead to misleading results. Districts were aggregated together in order to increase the frequency counts of each analysis—district analyses can be found in the Appendices.

Table 2 presents descriptive statistics (mean, median, frequency count, and standard deviation) for the absolute values of percent proficient or above longitudinal student class differences between adjacent years for all districts by subject area. **Results suggest that the modified cutscore method is less consistent overall in comparison to the original method in both subject areas.** For example, the mean (or average) absolute value difference between 2017 math percent proficient or above results and the 2018 modified method results is approximately 19% and the same method comparison in ELA is about 17%. The original method has a mean change of about 15% in math and 12% in ELA.
Table 2.
*Descriptive Statistics of the Absolute Values of Percent Proficient or Above Longitudinal Student Class Differences between Adjacent Years for All Districts by Subject*

<table>
<thead>
<tr>
<th>District</th>
<th>Subject</th>
<th>2015 to 2016 Results</th>
<th>2016 to 2017 Results</th>
<th>2017 to 2018 Modified (Pooled) Method Results</th>
<th>2017 to 2018 Original Method Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Districts</td>
<td>Math</td>
<td>Mean</td>
<td>12.56</td>
<td>12.88</td>
<td>19.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
<td>11.5</td>
<td>12.5</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>16</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>8.438</td>
<td>8.473</td>
<td>13.356</td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td>Mean</td>
<td>9.33</td>
<td>11.31</td>
<td>17.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
<td>7</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>15</td>
<td>26</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>6.411</td>
<td>8.064</td>
<td>11.176</td>
</tr>
</tbody>
</table>

The longitudinal analysis in the Appendix provides more nuance as it is organized by the year students will graduate from high school (i.e., Class of 2023, Class of 2024, Class of 2025, or Class of 2026). Some districts will not have analyses for a particular class of students depending on the grade levels in which they are implementing PACE for accountability purposes. The x-axis is labeled by year of results, grade level, and method used for 2018 (modified or original cutscore method). The y-axis is the percent of students proficient or above (labeled ‘mean’). Red bars indicate results are based on statewide assessment system (Smarter Balanced), green bars indicate results are based on PACE modified cutscore method, and purple bars indicate results are based on PACE original cutscore method.

**State Test Analysis**

The state test analysis compares the 2018 performance of students by subject area across state assessment systems (NH SAS and PACE). Grades 3-11 are included in order to incorporate as much information as possible. Figures 1 & 2 below illustrate the results for PACE districts using red bars for the 2018 NH SAS (statewide achievement test) results in Gr 4, 8, and 11 math and Gr 3, 8, and 11 ELA, respectively. Grade 11 math and ELA is the SAT. The green bars represent the modified (pooled) cutscore method and the purple bars represent the original (old) cutscore method.
Figure 1. Comparison of 2018 Mean Percent Proficient or Above Results from NH SAS and PACE (modified/pooled vs. original/old) for all PACE Districts in Math Grades 3-11

Figure 2. Comparison of 2018 Mean Percent Proficient or Above Results from NH SAS and PACE (modified/pooled vs. original/old) for all PACE Districts in ELA Grades 3-11
Results from analyzing the two figures suggest that the modified and original cutscore methods at the aggregate level are fairly consistent with the NH SAS results except for a couple grade/subject combinations. For example, the math results differ between methods in Grade 7 math where the modified method appears to slightly underestimate student achievement and the original method appears to slightly overestimate student achievement. Similarly, the ELA results suggest that the modified method tends to underestimate the percent of students deemed proficient or above in Grade 4 ELA; whereas the original method appears almost level with the NH SAS results from Gr 3 ELA.

As with the other analyses in this report, district-specific results by subject area can be found in the Appendices. Results for the state test analyses vary by district, subject area, and grade level so it is important to consider all analyses together when trying to make a judgment about which method is a more accurate reflection of student achievement.

**Conclusion**

The three analyses in this report evaluate the 2018 PACE modified (pooled) vs. original (old) cutscore methods to evaluate which method is a more trustworthy reflection of student achievement based on cohort, longitudinal, and state test analyses. **When considering all districts together, the original cutscore method appears to produce more stable results across cohorts and years than the modified method.** Effects of the two methods do vary, however, based on district, grade, and subject area. Some districts will likely ‘benefit’ from using the original method and other districts will not.

*In addition, there is also evidence from the district-level analyses synthesized at the end of each Appendix and collated in Table 3 that the modified method tended to pull historically high achieving districts down towards the mean (e.g., Amherst) and historically low achieving districts up towards the mean (e.g., Newport & Laconia).* This is an artifact of the modified method where we pooled all districts together to do standard setting.

Table 3 below is our attempt to synthesize the effects of the two methods for each of the PACE districts. In the end, some districts will ‘benefit’ from a change to the original method in some grade/subject areas because the percent of students deemed proficient or above will increase, but other districts will ‘not benefit’ depending on the grade/subject area. **The three analyses used in this report to evaluate the two cutscore methods supports the general conclusion that the original method is more consistent with both previous grade and student performance, but as noted throughout this report, there is considerable variability by district, grade, and subject.**
### Table 3.

*Effects of the Modified and/or Original Cutscore Methods Based on District-Level Analyses*

<table>
<thead>
<tr>
<th>District</th>
<th>Effect of Modified and/or Original Cutscore Method</th>
<th>Comparison of Modified (green) vs. Original (purple) Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amherst</td>
<td>Modified method tended to significantly reduce the percentage of students deemed proficient or above.</td>
<td></td>
</tr>
<tr>
<td>Bethlehem</td>
<td>Small sample sizes.</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>Concord</td>
<td>Modified method tended to increase the percent of students deemed proficient or above.</td>
<td></td>
</tr>
<tr>
<td>Epping</td>
<td>Modified method tended to reduce the percent of students proficient or above and the original method appears to increase the percent (and vice versa), depending on grade/subject.</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>Laconia</td>
<td>Both methods tended to inflate student proficiency. The modified method tended to increase the percent of students deemed proficient or above slightly more than the original method.</td>
<td></td>
</tr>
<tr>
<td>Newport</td>
<td>Modified method tended to significantly increase the percent of students deemed proficient or above.</td>
<td></td>
</tr>
<tr>
<td>Pittsfield</td>
<td>Small sample sizes.</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>Plymouth</td>
<td>Small sample sizes.</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>Rochester</td>
<td>In some grades in math the original method tends to overestimate student achievement results and the modified method tends to underestimate results. Alternatively, in some grades in ELA the modified method tends to overestimate student achievement results and the original method tends to underestimate results. Both methods are inconsistent with the state test results in most grade/subject combinations, although an argument could be made that the original method</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>Location</td>
<td>Description</td>
<td>Conclusion</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Sanborn</td>
<td>In math, the original method tended to overestimate the percent of students proficient or above in a couple grades and the modified method tends to underestimate the percent of students in a couple grades. In ELA, the original method tended to underestimate achievement in one particular grade and both methods appeared to underestimate achievement in another grade.</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>SAU23</td>
<td>Small sample sizes.</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>Seacoast</td>
<td>Small sample sizes</td>
<td>Inconclusive</td>
</tr>
</tbody>
</table>
APPENDIX A: AMHERST

Amherst started implementing the PACE pilot in the 2017-18 school year at grades 5-7. Amherst offers a unique context with which to evaluate the modified vs. original cutscore method because there are three years of Smarter Balanced test results available prior to implementation of PACE in the 2017-18 school year (~130-180 students/grade).

Cohort Analysis Graphs
Longitudinal Analysis Graphs

Class of 2023

District: Amherst, Subject: Math

Mean

- 2015 % Prof+ GRADE 4
- 2016 % Prof+ GRADE 5
- 2017 % Prof+ GRADE 6
- 2018 Pooled Method % Prof+ GRADE 7
- 2018 Old Method % Prof+ GRADE 7

District: Amherst, Subject: Reading

Percent Prof-

- 2015 % Prof+ GRADE 4
- 2016 % Prof+ GRADE 5
- 2017 % Prof+ GRADE 6
- 2018 Pooled Method % Prof+ GRADE 7
- 2018 Old Method % Prof+ GRADE 7
State Test Analysis Graphs

District: Amherst, Subject: Math

- 2019%Prof+ STATE
- 2018 Pooled Method %Prof+
- 2018 Old Method %Prof+

Grade

Mean

Grade

District: Amherst, Subject: Reading

- 2019%Prof+ STATE
- 2018 Pooled Method %Prof+
- 2018 Old Method %Prof+

Mean

Grade
Discussion

Amherst’s analysis offers a unique lens into the evaluation of the modified vs. original cutscore method because they started implementing the PACE pilot in the 2017-18 school year at grades 5-8. As a result, there is information available from the statewide achievement test (Smarter Balanced) from 2014-15 to provide context for district trends over time. In general, the various analyses for Amherst indicate that the original method was more consistent than the modified method based on the cohort, longitudinal, and state test analyses. The modified method tended to significantly reduce the percentage of students deemed proficient or above.
Bethlehem started implementing the PACE pilot in the 2016-17 school year at grades 3-6. Bethlehem is included as part of the consolidated SAU35 school district for standard setting (Lisbon, Profile, Lafayette, Landaff), which implemented PACE in the 2016-17 school year at all PACE accountability grades 3-8. The SAU35 school district (other than Bethlehem) left the PACE pilot after the 2016-17 school year. Bethlehem is a small school district with typically around 20-30 students per grade level. Due to small sample sizes, results may fluctuate from year-to-year due to only a few students moving proficiency classifications and therefore results should be interpreted with caution and in the context of other district trends.

**Cohort Analysis Graphs**

![Cohort Analysis Graphs](image-url)
State Test Analysis Graphs

District: Bethlehem, Subject: Math

- 2018% Prof+ STATE
- 2018 Pooled Method % Prof+
- 2018 Old Method % Prof+

District: Bethlehem, Subject: Reading

- 2018% Prof+ STATE
- 2018 Pooled Method % Prof+
- 2018 Old Method % Prof+
Discussion

The difficulty with interpreting results of the various analyses from Bethlehem is small sample sizes. Instability in results across cohorts, years, or assessment systems therefore may be an artifact of the uncertainty associated with small samples. In general, it appears that the modified method was more stringent than the original cutscore method—but effects do vary by grade and subject.
APPENDIX C: CONCORD

Concord started implementing the PACE pilot in the 2015-16 school year at all PACE accountability grades (grades 3-8).

Cohort Analysis Graphs
Longitudinal Analysis Graphs

Class of 2023

District: Concord, Subject: Math

Mean

Year

2015 % Prof+ GRADE 4
2016 % Prof+ GRADE 5
2017 % Prof+ GRADE 6
2018 Pooled Method % Prof+ GRADE 7
2018 Old Method % Prof+ GRADE 7
Class of 2024
Class of 2025

District: Concord, Subject: Math

District: Concord, Subject: Reading
Class of 2026

District: Concord, Subject: Reading

Mean

2017 %Prof+ GRADE 3
2018 Pooled Method %Prof+ GRADE 4
2018 Old Method %Prof+ GRADE 4
State Test Analysis Graphs

District: Concord, Subject: Math

District: Concord, Subject: Reading
Discussion

In general, the various analyses for Concord indicate that the original method was more consistent than the modified method based on cohort, longitudinal, and state test analyses. The modified method tended to increase the percent of students deemed proficient or above in Concord, particularly in Gr 3, 6-7 math and Gr 5 & 7 ELA.
APPENDIX D: EPPING

Epping started implementing the PACE pilot in the 2014-15 school year at all PACE accountability grades (grades 3-8).

Cohort Analysis Graphs
District: Epping, Subject: Reading, Grade: 7

Percent Prof+:
- 2015 %Prof+
- 2016 %Prof+
- 2017 %Prof+
- 2018 Pooled Method %Prof+
- 2018 Old Method %Prof+
Longitudinal Analysis Graphs

Class of 2023

District: Epping, Subject: Math
Class of 2024

District: Epping, Subject: Math

District: Epping, Subject: Reading
Class of 2025

District: Epping, Subject: Math

District: Epping, Subject: Reading
State Test Analysis Graphs

District: Epping, Subject: Math

- 2016 % Prof+ STATE
- 2018 Pooled Method % Prof+
- 2018 Old Method % Prof+

Grade

District: Epping, Subject: Reading

- 2016 % Prof+ STATE
- 2018 Pooled Method % Prof+
- 2018 Old Method % Prof+

Grade
Discussion

In general, results in Epping are inconclusive when comparing the efficacy of the modified vs. original cutscore method based on the cohort, longitudinal, and state test analysis. In some grades/subjects, the modified method appears to underestimate the percent of students proficient or above and the original method appears to overestimate (e.g., Gr 3 & 7 math; Gr 4 ELA); whereas, in other grades/subjects the modified method appears to overestimate the percent of students proficient or above and the original method appears to underestimate (e.g., Gr 5 ELA).
Laconia started implementing the PACE pilot in the 2017-18 school year at grades 3-5. Laconia offers a unique context with which to evaluate the modified vs. original cutscore method because there are three years of Smarter Balanced test results available prior to implementation of PACE in the 2017-18 school year (~140-170 students/grade).

**Cohort Analysis Graphs**

---

**District: Laconia, Subject: Math, Grade: 3**

- 2015 %Prof+
- 2016 %Prof+
- 2017 %Prof+
- 2018 Pooled Method %Prof+
- 2018 Old Method %Prof+
Longitudinal Analysis Graphs

Class of 2025

District: Laconia, Subject: Math

District: Laconia, Subject: Reading
District: Laconia, Subject: Reading

Mean

2017 %Prof+ GRADE 3  |  2018 Pooled Method %Prof+ GRADE 4  |  2018 Old Method %Prof+ GRADE 4

Class of 2026
Discussion

In general, results from Laconia suggest that the modified method overestimated the percent of student proficient or above in comparison to the original method; however, both methods do not appear as consistent with cohort, longitudinal, or state trends as expected. Both the modified and original method appear to inflate student proficiency in Laconia in most grade/subject combinations.
APPENDIX F: NEWPORT

Newport started implementing the PACE pilot in the 2017-18 school year at grades 3-5. Newport offers a unique context with which to evaluate the modified vs. original cutscore method because there are three years of Smarter Balanced test results available prior to implementation of PACE in the 2017-18 school year, although there are fewer students per grade than Amherst and Laconia (~60-90 students/grade).

Cohort Analysis Graphs
Class of 2025

Longitudinal Analysis Graphs

District: Newport, Subject: Math

Class of 2025

District: Newport, Subject: Reading

66
Class of 2026
Discussion

Results in Newport suggest that the modified method tends to significantly overestimate the percent of students deemed proficient or above in most grade/subject combinations based on cohort, longitudinal, and state test analyses.
APPENDIX G: PITTSFIELD

Pittsfield started implementing the PACE pilot in the 2015-16 school year at all PACE accountability grades 3-8. Pittsfield is a small school district with typically around 30-40 students per grade level. Due to small sample sizes, results may fluctuate from year-to-year due to only a few students moving proficiency classifications and therefore results should be interpreted with caution and in the context of other district trends.

Cohort Analysis Graphs
Longitudinal Analysis Graphs

Class of 2023

District: Pittsfield, Subject: Math

District: Pittsfield, Subject: Reading
Class of 2024

District: Pittsfield, Subject: Math

Mean

2015 %Prof+ GRADE 3 2016 %Prof+ GRADE 4 2017 %Prof+ GRADE 5 2018 Pooled Method %Prof+ GRADE 6 2018 Old Method %Prof+ GRADE 6

District: Pittsfield, Subject: Reading

Mean

2015 %Prof+ GRADE 3 2016 %Prof+ GRADE 4 2017 %Prof+ GRADE 5 2018 Pooled Method %Prof+ GRADE 6 2018 Old Method %Prof+ GRADE 6
Class of 2025

District: Pittsfield, Subject: Math

District: Pittsfield, Subject: Reading
Class of 2026

District: Pittsfield, Subject: Reading

Mean

2017 %Prof+ GRADE 3  2018 Pooled Method %Prof+ GRADE 4  2018 Old Method %Prof+ GRADE 4
State Test Analysis Graphs

**District: Pittsfield, Subject: Math**

- 2018 %Prof+ STATE
- 2018 Pooled Method %Prof+
- 2018 Old Method %Prof+

**District: Pittsfield, Subject: Reading**

- 2018 %Prof+ STATE
- 2018 Pooled Method %Prof+
- 2018 Old Method %Prof+
Discussion

The difficulty with interpreting results of the various analyses from Pittsfield is small sample sizes. Instability in results across cohorts, years, or assessment systems therefore may be an artifact of the uncertainty associated with small samples. In general, both the modified and original cutscore methods tend to inflate the percent of students deemed proficient or above in Pittsfield based on the cohort, longitudinal, and state test analyses—especially in math. The original method tends to overestimate student proficiency more than the modified method. The modified method appears unstable in ELA as there are large fluctuations in the percent of students deemed proficient or above between adjacent grades.
Plymouth started implementing the PACE pilot in the 2017-18 school year at all PACE accountability grades 3-8. There are three years of Smarter Balanced test results available prior to implementation of PACE in the 2017-18 school year. Plymouth is a small school district with typically around 40-50 students per grade level. Due to small sample sizes, results may fluctuate from year-to-year due to only a few students moving proficiency classifications and therefore results should be interpreted with caution and in the context of other district trends.

Cohort Analysis Graphs
District: Plymouth, Subject: Math, Grade: 7

District: Plymouth, Subject: Reading, Grade: 4
Longitudinal Analysis Graphs

Class of 2023

District: Plymouth, Subject: Math

District: Plymouth, Subject: Reading
Class of 2024

District: Plymouth, Subject: Math

District: Plymouth, Subject: Reading
Class of 2025

District: Plymouth, Subject: Math

District: Plymouth, Subject: Reading
Class of 2026

District: Plymouth, Subject: Reading

- 2017 %Prof+ GRADE 3
- 2018 Pooled Method %Prof+ GRADE 4
- 2018 Old Method %Prof+ GRADE 4

Mean
State Test Analysis Graphs

District: Plymouth, Subject: Math

Graph showing the mean scores for different grades and methods.

District: Plymouth, Subject: Reading

Graph showing the mean scores for different grades and methods.
Discussion

The difficulty with interpreting results of the various analyses from Plymouth is small sample sizes. Instability in results across cohorts, years, or assessment systems therefore may be an artifact of the uncertainty associated with small samples. In general, the original method tends to underestimate the percent of students proficient or above in comparison to the modified method based on cohort, longitudinal, and state trends. The one exception is Grade 5 ELA where the original method significantly overestimated student achievement in comparison to the modified method. The original method is particularly unstable in math with large fluctuations in percent proficient or above in adjacent grades.
APPENDIX I: ROCHESTER

Rochester started implementing the PACE pilot in the 2014-15 school year at all PACE accountability grades (grades 3-8).

Cohort Analysis Graphs
Longitudinal Analysis Graphs

Class of 2023

District: Rochester, Subject: Math

- 2015 %Prof+ GRADE 4
- 2018 %Prof+ GRADE 5
- 2017 %Prof+ GRADE 6
- 2018 Pooled Method %Prof+ GRADE 7
- 2018 Old Method %Prof+ GRADE 7

District: Rochester, Subject: Reading

- 2015 %Prof+ GRADE 4
- 2018 %Prof+ GRADE 5
- 2017 %Prof+ GRADE 6
- 2018 Pooled Method %Prof+ GRADE 7
- 2018 Old Method %Prof+ GRADE 7
Class of 2024

District: Rochester, Subject: Math

District: Rochester, Subject: Reading
Class of 2025

District: Rochester, Subject: Math

District: Rochester, Subject: Reading
Class of 2026
State Test Analysis Graphs

**District: Rochester, Subject: Math**

- 2018% Prof+ STATE
- 2018 Pooled Method % Prof+
- 2018 Old Method % Prof+

<table>
<thead>
<tr>
<th>Grade</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**District: Rochester, Subject: Reading**

- 2018% Prof+ STATE
- 2018 Pooled Method % Prof+
- 2018 Old Method % Prof+

<table>
<thead>
<tr>
<th>Grade</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion

In general, the various analyses for Rochester are inconclusive based on the cohort, longitudinal, and state trend analyses. In some grades in math the original method tends to overestimate student achievement results and the modified method tends to underestimate results (e.g., gr 3 & 5 math). Alternatively, in some grades in ELA the modified method tends to overestimate student achievement results and the original method tends to underestimate results (e.g., gr 6 & 7 ELA). Both methods are inconsistent with the state test results in most grade/subject combinations, although an argument could be made that the original method is more consistent with the exception of Gr 3 math.
APPENDIX J: SANBORN

Sanborn started implementing the PACE pilot in the 2014-15 school year at all PACE accountability grades (grades 3-8).

Cohort Analysis Graphs
Longitudinal Analysis Graphs

Class of 2023

District: Sanborn, Subject: Math
Class of 2025

District: Sanborn, Subject: Math

- 2015 %Prof+ GRADE 2
- 2016 %Prof+ GRADE 3
- 2017 %Prof+ GRADE 4
- 2018 Pooled Method %Prof+ GRADE 5
- 2018 Old Method %Prof+ GRADE 5

District: Sanborn, Subject: Reading

- 2015 %Prof+ GRADE 2
- 2016 %Prof+ GRADE 3
- 2017 %Prof+ GRADE 4
- 2018 Pooled Method %Prof+ GRADE 5
- 2018 Old Method %Prof+ GRADE 5
Class of 2026

District: Sanborn, Subject: Reading

Mean

2017 %Prof+ GRADE 3
2018 Pooled Method %Prof+ GRADE 4
2018 Old Method %Prof+ GRADE 4
State Test Analysis Graphs

District: Sanborn, Subject: Math

District: Sanborn, Subject: Reading
Discussion

In general, the various analyses for Sanborn are inconclusive based on the cohort, longitudinal, and state trend analyses. In math, the original method tended to overestimate the percent of students proficient or above in Gr 6 & 7 and the modified method tends to underestimate the percent of students in Gr 3 & 7. In ELA, the original method tended to underestimate achievement in Gr 5 and both methods appeared to underestimate achievement in Gr 4.
APPENDIX K: SAU23

SAU23 started implementing the PACE pilot in the 2017-18 school year at the following: Bath grades 3-6; Piermont grades 3-8; and Warren grades 3-8. Due to small sample sizes, results may fluctuate from year-to-year due to only a few students moving proficiency classifications and therefore results should be interpreted with caution and in the context of other district trends. It was difficult to ensure that state test comparisons only included results for those students in the PACE pilot given the consolidated nature of SAU23.

Cohort Analysis Graphs

![Cohort Analysis Graph](image-url)
State Test Analysis Graphs

District: SAU 23, Subject: Math

District: SAU 23, Subject: Reading
Discussion

The difficulty with interpreting results of the various analyses from SAU23 is small sample sizes and the difficulty with obtaining accurate aggregate results by grade and subject for only those schools (or smaller SAUs) in the PACE pilot. Instability in results across cohorts, years, or assessment systems therefore may be an artifact of the uncertainty associated with small samples or inclusion/exclusion of students in the PACE system. In general, the original method tends to overestimate the percent of students proficient or above in comparison to the modified method based on cohort, longitudinal, and state trends.
Seacoast Charter School started implementing the PACE pilot in the 2015-16 school year at all PACE accountability grades (grades 3-8). Due to small sample sizes (~20-40 students/grade), results may fluctuate from year-to-year due to only a few students moving proficiency classifications and therefore results should be interpreted with caution and in the context of other district trends.

Cohort Analysis Graphs
Longitudinal Analysis Graphs

Class of 2023

District: SeaCoast, Subject: Math

<table>
<thead>
<tr>
<th>Year</th>
<th>% Prof+ Grade 4</th>
<th>% Prof+ Grade 5</th>
<th>% Prof+ Grade 6</th>
<th>Pooled Method % Prof+ Grade 7</th>
<th>Old Method % Prof+ Grade 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>38</td>
<td>45</td>
<td>48</td>
<td>52</td>
<td>55</td>
</tr>
<tr>
<td>2016</td>
<td>40</td>
<td>50</td>
<td>55</td>
<td>60</td>
<td>58</td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td>55</td>
<td>65</td>
<td>70</td>
<td>68</td>
</tr>
<tr>
<td>2018 Pooled</td>
<td>45</td>
<td>55</td>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018 Old</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Class of 2024

**District: Seacoast, Subject: Reading**

<table>
<thead>
<tr>
<th>Year</th>
<th>Grade</th>
<th>Method</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Prof+</td>
<td>Grade 4</td>
<td>50</td>
</tr>
<tr>
<td>2016</td>
<td>Prof+</td>
<td>Grade 5</td>
<td>40</td>
</tr>
<tr>
<td>2017</td>
<td>Prof+</td>
<td>Grade 6</td>
<td>50</td>
</tr>
<tr>
<td>2018 Pooled</td>
<td>Prof+</td>
<td>Grade 7</td>
<td>40</td>
</tr>
<tr>
<td>2018 Old</td>
<td>Prof+</td>
<td>Grade 7</td>
<td>30</td>
</tr>
</tbody>
</table>

**District: Seacoast, Subject: Math**

<table>
<thead>
<tr>
<th>Year</th>
<th>Grade</th>
<th>Method</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Prof+</td>
<td>Grade 3</td>
<td>60</td>
</tr>
<tr>
<td>2016</td>
<td>Prof+</td>
<td>Grade 4</td>
<td>50</td>
</tr>
<tr>
<td>2017</td>
<td>Prof+</td>
<td>Grade 5</td>
<td>40</td>
</tr>
<tr>
<td>2018 Pooled</td>
<td>Prof+</td>
<td>Grade 6</td>
<td>60</td>
</tr>
<tr>
<td>2018 Old</td>
<td>Prof+</td>
<td>Grade 6</td>
<td>50</td>
</tr>
</tbody>
</table>
Class of 2025

District: Seacoast, Subject: Math

District: Seacoast, Subject: Reading
Class of 2026
State Test Analysis Graphs

District: Seacoast, Subject: Math

District: Seacoast, Subject: Reading
Discussion

The difficulty with interpreting results of the various analyses from Seacoast Charter School is small sample sizes. Instability in results across cohorts, years, or assessment systems therefore may be an artifact of the uncertainty associated with small samples. In general, both methods tend to inflate student achievement in math with the exception of Gr 3. The modified method tends to overestimate student proficiency in math more than the original method.