

**Items that Require Additional Information or Revision in New Hampshire’s Innovative
Assessment Demonstration Authority Plan**

July 30, 2018

TABLE OF CONTENTS

Regulatory requirement.....	3
Consultation.....	3
Innovative assessment system (1).....	4
Innovative Assessment System (4)(i).....	9
Innovative assessment system (4)(ii).....	14
Innovative assessment system (5)(i).	15
Innovative assessment system (8).....	16
Innovative assessment system (9).....	17
Initial implementation in a subset of LEAs or schools.....	18
Application selection criteria	21
(a)(2).	21
(a)(3).	22
(b)(1).....	24
(b)(2).....	25
(c)(1).	27
(c)(2).	33
(d)(1).....	34
(d)(2).....	36
(d)(3).....	37
(d)(4).....	38
(e)(1).	40
(e)(2).	41
Appendix A: PACE District Achievement and Participation Rates 2016-17	42
Appendix B: Inter-Rater Reliability Analyses in 2015, 2016, and 2017.....	105
NH PACE: Inter-rater Reliability Analysis Report 2015	105
NH PACE: Inter-rater Reliability Analysis Report 2016	109
NH PACE: Inter-rater Reliability Analysis Report 2017	116
Appendix C: Generalizability Analysis in 2016 and 2017	125
Generalizability Analysis 2016	125
Generalizability Analysis 2017	131
Appendix D: Calibration Analysis in 2015, 2016, and 2017	141
Calibration Analysis 2015	141

Calibration Analyses 2016	154
Calibration Analyses 2017.....	163
Appendix E: Standard Setting Reports in 2015, 2016, and 2017	176
Standard Setting Report 2015	181
Standard Setting Report 2016	186
Standard Setting Report 2017	193
Appendix F: Body of Work Standards Validation 2016 and 2017	202
Body of Work (BOW) Standards Validation 2016.....	202
Body of Work (BOW) Standards Validation 2017	204
Appendix G: Concurrent Analyses 2016 and 2017	208
Concurrent Analyses 2016.....	208
Concurrent Analyses 2017.....	213
Appendix H: Non-Concurrent Analyses 2016 and 2017	223
Non-Concurrent Analyses 2016.....	223
Non-Concurrent Analyses 2017.....	227

Regulatory requirement	Required information from the SEA
<p>Consultation. Evidence that the SEA or consortium has developed an innovative assessment system in collaboration with--</p> <p>(1) Experts in the planning, development, implementation, and evaluation of innovative assessment systems, which may include external partners; and</p> <p>(2) Affected stakeholders in the State, or in each State in the consortium, including--</p> <p>(i) Those representing the interests of children with disabilities, English learners, and other subgroups of students described in section 1111(c)(2) of the Act;</p> <p>(ii) Teachers, principals, and other school leaders;</p> <p>(iii) Local educational agencies (LEAs);</p> <p>(iv) Representatives of Indian tribes located in the State;</p> <p>(v) Students and parents, including parents of children described in paragraph (a)(2)(i) of this section; and</p> <p>(vi) Civil rights organizations.</p>	<p>Provide a description for how the State has monitored the LEA consultation with those representing the interests of children (including children with disabilities, English learners, and other sub-groups of students described in section 1111(c)(2) of the ESEA).</p>

New Hampshire’s Response:

NH DOE realizes it can be more systematic in how it addresses this requirement. Therefore, going forward, NH DOE and the PACE leadership team will offer specific directions about the type of consultation required under Section 1204 and outline the types of evidence to be collected to allow NH DOE to monitor and improve these consultation efforts. These procedures will be presented to the district leaders at the first PACE leadership team meeting in early September 2018. Each district leader will be required to submit a report on their consultation efforts quarterly. These issues will be discussed at monthly PACE leadership meeting starting in October 2018 and will be subject to audit by NH DOE.

Regulatory requirement	Required information from the SEA
<p>Innovative assessment system (1). A demonstration that the innovative assessment system does or will--</p> <p>(1) Meet the requirements of section 1111(b)(2)(B) of the Act, except that an innovative assessment--</p> <p>(i) Need not be the same assessment administered to all public elementary and secondary school students in the State during the demonstration authority period described in 34 CFR 200.104(b)(2) or extension period described in 34 CFR 200.108 and prior to statewide use consistent with 34 CFR 200.107, if the innovative assessment system will be administered initially to all students in participating schools within a participating LEA, provided that the statewide academic assessments under 34 CFR 200.2(a)(1) and section 1111(b)(2) of the Act are administered to all students in any non-participating LEA or any non-participating school within a participating LEA; and</p> <p>(ii) Need not be administered annually in each of grades 3-8 and at least once in grades 9-12 in the case of reading/language arts and mathematics assessments, and at least once in grades 3-5, 6-9, and 10-12 in the case of science assessments, so long as the statewide academic assessments under 34 CFR 200.2(a)(1) and section 1111(b)(2) of the Act are administered in any required grade and subject under 34 CFR 200.5(a)(1) in which the SEA does not choose to implement an innovative assessment.</p>	<p>From the most recently available year of data, evidence that all students in participating PACE schools participated in either the PACE pilot assessment or the statewide assessment as required in section 1201(e)(2)(A)(x and xi) of the ESEA (i.e., a report that shows for each participating school, by grade, the participation rates in PACE and the participation rates in the statewide assessment for those grade/subjects not assessed with PACE).</p>

New Hampshire's Response:

In the 2015, 2016, and 2017 years, New Hampshire reported PACE participation rates at the district-level because every school within participating PACE districts implemented the PACE pilot. The district-level report for the most recently available year of data (2017) is provided in **Appendix A** of this document. PACE pilot assessment grades/subjects are not highlighted and the statewide assessment grades/subjects are highlighted in yellow. The overall participation rates for reading and math by district and grade are summarized below.

Bethlehem: 2017 Participation Rate

Grade	rea	mat
3	100%	100%
4	100%	100%
5	100%	100%
6	100%	100%
7	-	-
8	-	-
11	-	-
Overall	100%	100%

Concord: 2017 Participation Rate

Grade	rea	mat
3	100%	99%
4	99%	100%
5	92%	91%
6	100%	100%
7	98%	98%
8	98%	98%
11	93%	93%
Overall	97%	97%

Epping: 2017 Participation Rate

Grade	rea	mat
3	99%	99%
4	100%	100%
5	97%	99%
6	100%	100%
7	96%	100%
8	100%	95%
11	95%	95%
Overall	98%	98%

Lafayette: 2017 Participation Rate

Grade	rea	mat
3	100%	100%
4	100%	100%
5	100%	100%
6	94%	94%
7	-	-
8	-	-
11	-	-
Overall	99%	99%

Landaff: 2017 Participation Rate

Grade	rea	mat
3	**	**
4	-	-
5	-	-
6	-	-
7	-	-
8	-	-
11	-	-
Overall	**	**

Lisbon: 2017 Participation Rate

Grade	rea	mat
3	89%	95%
4	100%	100%
5	100%	100%
6	100%	100%
7	98%	100%
8	100%	100%
11	94%	94%
Overall	98%	99%

Monroe: 2017 Participation Rate

Grade	rea	mat
3	**	**
4	**	**
5	**	**
6	**	**
7	100%	100%
8	**	**
11	**	**
Overall	78%	78%

Pittsfield: 2017 Participation Rate

Grade	rea	mat
3	97%	100%
4	100%	98%
5	100%	100%
6	98%	100%
7	100%	74%
8	100%	95%
11	76%	76%
Overall	96%	93%

Profile: 2017**Participation Rate**

Grade	rea	mat
3	-	-
4	-	-
5	-	-
6	-	-
7	100%	100%
8	100%	100%
11	100%	100%
Overall	100%	100%

Rochester: 2017**Participation Rate**

Grade	rea	mat
3	100%	99%
4	99%	99%
5	100%	100%
6	95%	95%
7	96%	96%
8	98%	98%
11	91%	91%
Overall	97%	97%

Sanborn: 2017**Participation Rate**

Grade	rea	mat
3	100%	99%
4	98%	99%
5	100%	100%
6	100%	100%
7	99%	100%
8	100%	100%
11	94%	94%
Overall	99%	99%

Seacoast Charter School: 2017**Participation Rate**

Grade	rea	mat
3	100%	96%
4	100%	100%
5	94%	94%
6	100%	100%
7	95%	100%
8	100%	100%
11	-	-
0	98%	98%

Souhegan: 2017 PACE Participation Rate

Grade	rea	mat
3	-	-
4	-	-
5	-	-
6	-	-
7	-	-
8	-	-
11	97%	97%
0	97%	97%

***Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included).**

**** Note: Count is below cell size of 10**

For the 2017-18 school year, we will report participation rates at the school-level because there will be partial implementation within some PACE school districts. Where we have partial implementation, students will be reported as participating in either PACE or the NH SAS to ensure that all students will participate in NH's assessment system.

Further, NH DOE will monitor all participating schools and districts with a goal to ensure that at least 95% of students in each subgroup of students fully participates in PACE.

Regulatory requirement	Required information from the SEA
<p>Innovative Assessment System (4)(i).</p> <p>(4)(i) Generate results, including annual summative determinations as defined in paragraph (b)(7) of this section, that are valid, reliable, and comparable for all students and for each subgroup of students described in 34 CFR 200.2(b)(11)(i)(A)-(I) and sections 1111(b)(2)(B)(xi) and 1111(h)(1)(C)(ii) of the Act, to the results generated by the State academic assessments described in 34 CFR 200.2(a)(1) and section 1111(b)(2) of the Act for such students. Consistent with the SEA’s or consortium’s evaluation plan under 34 CFR 200.106(e), the SEA must plan to annually determine comparability during each year of its demonstration authority period in one of the following ways:</p> <p>(A) Administering full assessments from both the innovative and statewide assessment systems to all students enrolled in participating schools, such that at least once in any grade span (i.e., 3-5, 6-8, or 9-12) and subject for which there is an innovative assessment, a statewide assessment in the same subject would also be administered to all such students. As part of this determination, the innovative assessment and statewide assessment need not be administered to an individual student in the same school year.</p> <p>(B) Administering full assessments from both the innovative and statewide assessment systems to a demographically representative sample of all students and subgroups of students described in section 1111(c)(2) of the Act, from among those students enrolled in participating schools, such that at least once in any grade span (i.e., 3-5, 6-8, or 9-12) and subject for which there is an innovative assessment, a statewide assessment in the same subject would also be administered in the same school year to all students included in the sample.</p> <p>(C) Including, as a significant portion of the innovative assessment system in each required grade and subject in which both an innovative and statewide assessment are administered, items or performance tasks from the statewide assessment system that, at a minimum, have been previously pilot tested or field tested for use in the statewide assessment system.</p> <p>(D) Including, as a significant portion of the statewide</p>	<p>While the approach described seems responsive to the question and likely to result in the State evaluating whether the assessments provide comparable results, NH DOE must provide the results of the studies identified in its application (on pages 20-27), namely:</p> <ol style="list-style-type: none"> 1. Results of the Inter-Rater Reliability Analyses in 2015, 2016, and 2017. 2. Results of the Generalizability Analyses in 2016 and 2017. 3. Results of the contrasting group standard setting analyses from 2015, 2016, and 2017. 4. Results of the calibration audits during the PACE Summer Institute in 2015, 2016, and 2017. 5. Results of the body of evidence audits from 2015, 2016, and 2017. 6. Results of the analysis of the rigor of the performance standards across PACE and non-PACE assessment systems from 2015, 2016, and 2017. 7. Results of the concurrent comparability evaluations from 2016 and 2017. 8. Results of the non-concurrent comparability evaluations from 2016 and 2017.

<p>assessment system in each required grade and subject in which both an innovative and statewide assessment are administered, items or performance tasks from the innovative assessment system that, at a minimum, have been previously pilot tested or field tested for use in the innovative assessment system.</p> <p>(E) An alternative method for demonstrating comparability that an SEA can demonstrate will provide for an equally rigorous and statistically valid comparison between student performance on the innovative assessment and the statewide assessment, including for each subgroup of students described in 34 CFR 200.2(b)(11)(i)(A)-(I) and sections 1111(b)(2)(B)(xi) and 1111(h)(1)(C)(ii) of the Act;</p>	
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--

New Hampshire’s Response:

The following studies identified in New Hampshire’s IADA application (on pages 20-27 of the original application) were included in annual reports to the U.S. Department of Education in October 2015, December 2016, and December 2017. The requested analyses were extracted from those reports and provided as evidence here.

1. The inter-rater analysis results provide support for the degree of inter-rater consistency in the scoring of the common performance tasks. This evidence suggests that teachers within districts are able to successfully conduct calibration sessions and comparably evaluate student work. When analyses reveal potential scoring problems with the consistency of scoring, the Center for Assessment and NH DOE work closely with those schools and districts to better understand the possible sources for reduced inter-rater reliability statistics and to find ways to improve the scoring practices. Complete results of the Inter-Rater Reliability Analyses in 2015, 2016, and 2017 are in **Appendix B**.
2. Results from the NH PACE 2016 Generalizability Analysis Report suggested that classroom assessments may provide for reliable estimates of student achievement for use in a school accountability context like the NH PACE pilot project. The 2016 analysis used electronic grade book data from one school district (N=257) and found that approximately 15-20 assessments per year provided for an efficient trade-off while still ensuring a high degree of relative and absolute decision reliability. The 2017 Generalizability Analysis Report used electronic grade book data from three (of the nine) districts with strong experience implementing the PACE pilot in 2016-17 with a total of 3,348 students. As before, we examined the generalizability of the individual scores that go into achievement estimates (e.g., summative tests, quizzes, projects, performance tasks). Based on our limited analyses thus far, the results suggest that classroom assessments may provide reliable estimates of student achievement for use in a school accountability context like the NH PACE pilot project. Approximately 10-15 assessments per year provide for an efficient trade-off while still ensuring a high degree of relative and absolute decision reliability. Complete results of the Generalizability Analyses in 2016 and 2017 are in **Appendix C**.

3. The PACE innovative assessment system uses common performance tasks across districts to evaluate the degree of comparability in local scoring. These analyses rest on the assumption that patterns in scoring for the common tasks are representative of district's relative stringency and leniency in scoring of the local performance tasks and assessments, without directly evaluating the scoring quality and consistency of local tasks. The calibration audit is intended to uncover differences in scoring among districts that can be used to support decision-making about any adjustments to cut scores that may be needed due to systematic cross-district differences. The scores of student work samples on PACE performance tasks that result from this audit serves as the "calibration weights" so that more generalized inferences about relative leniency or stringency of district scoring practices can be made. The calibration and comparability analyses have been one of the most successful technical aspects of PACE. Each year—2015, 2016, and 2017—for which the analyses have been conducted has demonstrated that both the process is effective for evaluating differences in leniency/rigor by which districts score student work and for making the very few adjustments when necessary to ensure that the results are comparable. Complete results of the Calibration Audits in 2015, 2016, and 2017 are in **Appendix D**.
4. The purpose of the standard setting is to determine where in the score distributions the appropriate "cut points" lie for establishing achievement levels. To establish cut points we used an examinee-centered judgmental method called contrasting groups. This standard setting method involves judgments from panelists about the qualifications of the examinees based on prior knowledge of the examinee. To implement this method for the PACE pilot, we asked PACE teachers to make judgments about which achievement level best described each of their students from the previous year. This process relied on the achievement level descriptors (ALDs) that are the same ALDs used on the statewide assessment. The subject and grade level specific ALDs were entered into an online survey where teachers could easily read the descriptions and match their former students to the appropriate achievement level. 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. The results for all three years have indicated high degrees of comparable annual determinations with the statewide assessment (Smarter Balanced in 2015-2017). Complete descriptions of the methods and results of the Contrasting Group Standard Setting Analyses from 2015, 2016, and 2017 are in **Appendix E**.
5. As part of validating the annual determinations produced via the contrasting groups standard setting approach, we have collected a "body of evidence" for a small sample of students from a sample of courses in each participating district. Each district collected assessment evidence throughout the academic year for a sample of nine students, representing the range of performance in that district, for one content area per grade level. Teachers were asked to collect samples of student work from those nine students for each of the competencies. As part of the summer calibration workshops, teachers from across the PACE districts came together to review the portfolios of student work to make judgments about student achievement relative to the Achievement Level Descriptors. There was some confusion about the nature of the work to include in the samples in 2016 which made the results difficult to interpret and use. The process improved considerably in 2017 and the body of evidence results were moderately related to the contrasting groups results. However, the process still suffered from some confusion in 2017 that we expect to be corrected for 2018. A complete

description of the methods and results of the body of evidence audits from 2016 and 2017 are in **Appendix F**. The Body of Work audit was not conducted in 2015.

6. The concurrent comparability analyses revealed that the percentage of students deemed proficient across the assessment systems is remarkably consistent. Secondly, by calculating “PACE annual determinations” for the students taking the Smarter Balanced in 2016, the state has both Smarter Balanced and PACE 2015-2016 annual determinations for students in grade 3 ELA, grade 4 math, grade 8 ELA and math, and grade 11 ELA and math. Though annual determinations were not reported for these subjects and grades for PACE and no common performance task was administered, the same procedure for producing annual determinations was used in these grade levels as for the PACE reported annual determinations. The degree of similarity between the distributions further supports the comparability of the interpretations of the reported achievement levels. For all four comparisons conducted, the classification accuracy is at least 70% agreement. While this agreement is high, there are a variety of reasons why there may be legitimate differences in the results produced by the different assessment systems. First, the degree of agreement is limited by the reliability of each assessment system. In other words, an assessment cannot correlate more with another assessment than it can with itself (i.e., reliability), so since both PACE and Smarter Balanced (or SAT) are not perfectly reliable, we are approaching the upper bound of the relationship between the two assessment systems. Additionally, New Hampshire’s PACE assessment system is in place to measure the state-defined learning targets differently than they are measured in the statewide assessment system. The purpose is to measure the standards more deeply and authentically through performance-based assessments. Additionally, the PACE assessment system is intended to measure the set of standards more completely (e.g., including the listening and speaking standards). Therefore, perfect agreement between the two assessment systems is not expected. The demonstrated 70% agreement in proficiency classification across the two systems should be considered acceptable given the competing objectives of attaining comparability while designing and implementing an innovative assessment system that is intended to create meaningful changes to teaching and learning. The classification accuracies for the reported subgroups do not vary greatly from the overall classification accuracy of approximately 70%. Some variation around 70% is natural due to sampling error associated with the small sample sizes of many of the subgroups. Complete methods and results of the analysis of the rigor of the performance standards across PACE and non-PACE assessment systems from 2016 and 2017 are provided in the Concurrent and Non-Concurrent Analyses **Appendix G & H**, respectively. Concurrent and Non-Concurrent Analyses were not conducted in 2015.
7. We conducted two non-concurrent comparability evaluations because students participate in Smarter Balanced (now NH SAS) once per grade span: Smarter Balanced 2016 to PACE 2017 and PACE 2016 to Smarter Balanced 2017. Non-concurrent analyses could not be conducted for the SAT given that PACE did not report annual determinations for high school students in 2016 or 2017 as per the requirements of the original waiver. As would be expected, the classification accuracies across years are slightly lower than the classification accuracies observed for the concurrent year comparisons, with the elementary grades having slightly higher levels of consistent classifications than middle school. The second analysis compares last years’ performance on PACE in grade 3 math and grade 7 ELA and math with this years’ performance on Smarter Balanced for students in grade 4 math and grade 8 ELA and math. Only students with a PACE achievement level in 2016 and Smarter Balanced

achievement level in 2017 were used for these analyses (N=2,344). In one out of the three grades and subject areas, the percent proficient rose from PACE 2016 to Smarter Balanced 2017 and in the other two grades and subject areas the percent proficient either went down or stayed about the same, indicating that PACE results are comparable. Complete discussion of the methods and results of the non-concurrent comparability evaluations from 2016 and 2017 are in **Appendix H**.

Regulatory requirement	Required information from the SEA
<p>Innovative assessment system (4)(ii). (4)(ii) Generate results, including annual summative determinations as defined in paragraph (b)(7) of this section, that are valid, reliable, and comparable, for all students and for each subgroup of students described in 34 CFR 200.2(b)(11)(i)(A)-(I) and sections 1111(b)(2)(B)(xi) and 1111(h)(1)(C)(ii) of the Act, among participating schools and LEAs in the innovative assessment demonstration authority. Consistent with the SEA's or consortium's evaluation plan under 34 CFR 200.106(e), the SEA must plan to annually determine comparability during each year of its demonstration authority period;</p>	<p>See information required under 4(i) above.</p>

New Hampshire's Response: See New Hampshire's response under 4(i) above.

Regulatory requirement	Required information from the SEA
<p>Innovative assessment system (5)(i). (5)(i) Provide for the participation of all students, including children with disabilities and English learners; (ii) Be accessible to all students by incorporating the principles of universal design for learning, to the extent practicable, consistent with 34 CFR 200.2(b)(2)(ii); and (iii) Provide appropriate accommodations consistent with 34 CFR 200.6(b) and (f)(1)(i) and section 1111(b)(2)(B)(vii) of the Act;</p>	<p>See information requested in requirement (1) above.</p>

New Hampshire's Response: See response under requirement (1) above.

Regulatory requirement	Required information from the SEA
<p>Innovative assessment system (8). (8) Provide disaggregated results by each subgroup of students described in 34 CFR 200.2(b)(11)(i)(A)-(I) and sections 1111(b)(2)(B)(xi) and 1111(h)(1)(C)(ii) of the Act, including timely data for teachers, principals and other school leaders, students, and parents consistent with 34 CFR 200.8 and section 1111(b)(2)(B)(x) and (xii) and section 1111(h) of the Act, and provide results to parents in a manner consistent with paragraph (b)(4)(i) of this section and part 200.2(e);</p>	<p>A report which demonstrates specifically the disaggregated results of all students in participating PACE schools in the PACE assessment is required.</p>

New Hampshire’s Response:

New Hampshire currently disaggregates PACE results for each subgroup of students at the district-level. These results include both PACE assessment system and statewide assessment system results. Going forward, New Hampshire will disaggregate results for each subgroup of students separately for the PACE assessment system and statewide assessment system results at the school-level as long as the N-count is above 10. **Appendix A** contains the district-level reports for all students in participating PACE districts overall and for each subgroup. PACE grades/subjects are not highlighted and non-PACE grades/subjects are highlighted in yellow.

Regulatory requirement	Required information from the SEA
<p>Innovative assessment system (9). (9) Provide an unbiased, rational, and consistent determination of progress toward the State’s long-term goals for academic achievement under section 1111(c)(4)(A) of the Act for all students and each subgroup of students described in section 1111(c)(2) of the Act and a comparable measure of student performance on the Academic Achievement indicator under section 1111(c)(4)(B) of the Act for participating schools relative to non-participating schools so that the SEA may validly and reliably aggregate data from the system for purposes of meeting requirements for-- (i) Accountability under sections 1003 and 1111(c) and (d) of the Act, including how the SEA will identify participating and non-participating schools in a consistent manner for comprehensive and targeted support and improvement under section 1111(c)(4)(D) of the Act; and (ii) Reporting on State and LEA report cards under section 1111(h) of the Act.</p>	<p>See information requested under (8) above.</p>

New Hampshire’s Response: See response provided under (8) above.

Regulatory requirement	Required information from the SEA
<p>Initial implementation in a subset of LEAs or schools.</p> <p>If the innovative assessment system will initially be administered in a subset of LEAs or schools in a State--</p> <p>(1) A description of each LEA, and each of its participating schools, that will initially participate, including demographic information and its most recent LEA report card under section 1111(h)(2) of the Act; and</p> <p>(2) An assurance from each participating LEA, for each year that the LEA is participating, that the LEA will comply with all requirements of this section.</p>	<p>NH DOE must provide an assurance from each LEA that that the LEA will comply with all requirements of the IADA, as applicable.</p>

New Hampshire’s Response:

As of June 7th, 2018 the NH DOE has received assurance from all participating PACE districts that the LEA will comply with all of the requirements of Section 1204 of the Every Student Succeeds Act. A list of the districts is provided below.

District	Status
Amherst & Souhegan Cooperative	Affirmative – letter and signature attached
Bethlehem	Affirmative – letter and signature attached
Concord	Affirmative – letter and signature attached
Epping	Affirmative – letter and signature attached
Haverhill Cooperative	Affirmative – letter and signature attached
Laconia	Affirmative – letter and signature attached
Monroe	Affirmative – letter and signature attached
Newport	Affirmative – letter and signature attached
Pittsfield	Affirmative – letter and signature attached
Plymouth	Affirmative – letter and signature attached
Rochester	Affirmative – letter and signature attached
Sanborn	Affirmative – letter and signature attached
Seacoast Charter School	Affirmative – letter and signature attached

Application selection criteria	Required information from the SEA
<p>(a)(1) The rationale for developing or selecting the particular innovative assessment system to be implemented under the demonstration authority, including--</p> <p>(i) The distinct purpose of each assessment that is part of the innovative assessment system and how the system will advance the design and delivery of large-scale, statewide academic assessments in innovative ways; and</p> <p>(ii) The extent to which the innovative assessment system as a whole will promote high-quality instruction, mastery of challenging State academic standards, and improved student outcomes, including for each subgroup of students described in section 1111(c)(2) of the Act;</p>	<p>NH DOE must provide:</p> <ol style="list-style-type: none"> 1. A specific description of how each component of PACE (local summative tests, common performance tasks and local performance assessments) contributes to the annual summative determination for each grade/subject in the pilot. 2. A clear description of how the PACE assessment design affords students multiple ways to demonstrate that they have mastered the content.

New Hampshire’s Response:

Figure 1 below is a graphical representation of NH’s PACE assessment system intended to support NH’s response to question 1 above. As seen in the figure, local summative assessments (performance tasks and other summative assessments) tied to specific competencies (and standards) are used to produce competency-level scores for each student in each school. The PACE technical consultants (Center for Assessment) collect these data from each district to produce district level competency scores for each student using an unweighted linear combination of the student-level competency scores. The PACE Common Performance Task counts in this overall competency according to the weight assigned within each district’s competency system. The PACE Common Task is used to support the calibration analyses and, depending on each district’s calibration results, district competency scores may be adjusted slightly to ensure that annual determinations (i.e., performance level designations) are comparable across PACE districts and among PACE and non-PACE districts.

The graphic below and the description of the system above should make clear that the rich combination of local and common assessments provides students with a wealth of diverse ways to demonstrate their knowledge and skills. This is the crux of competency-based education—the foundation of PACE—that, by design, provides students multiple opportunities and multiple approaches for demonstrating mastery of key competencies. These multiple assessment opportunities are provided through the local assessment components of the system, but the quality of the local assessment component is evaluated by NH DOE and PACE leadership (discussed elsewhere).

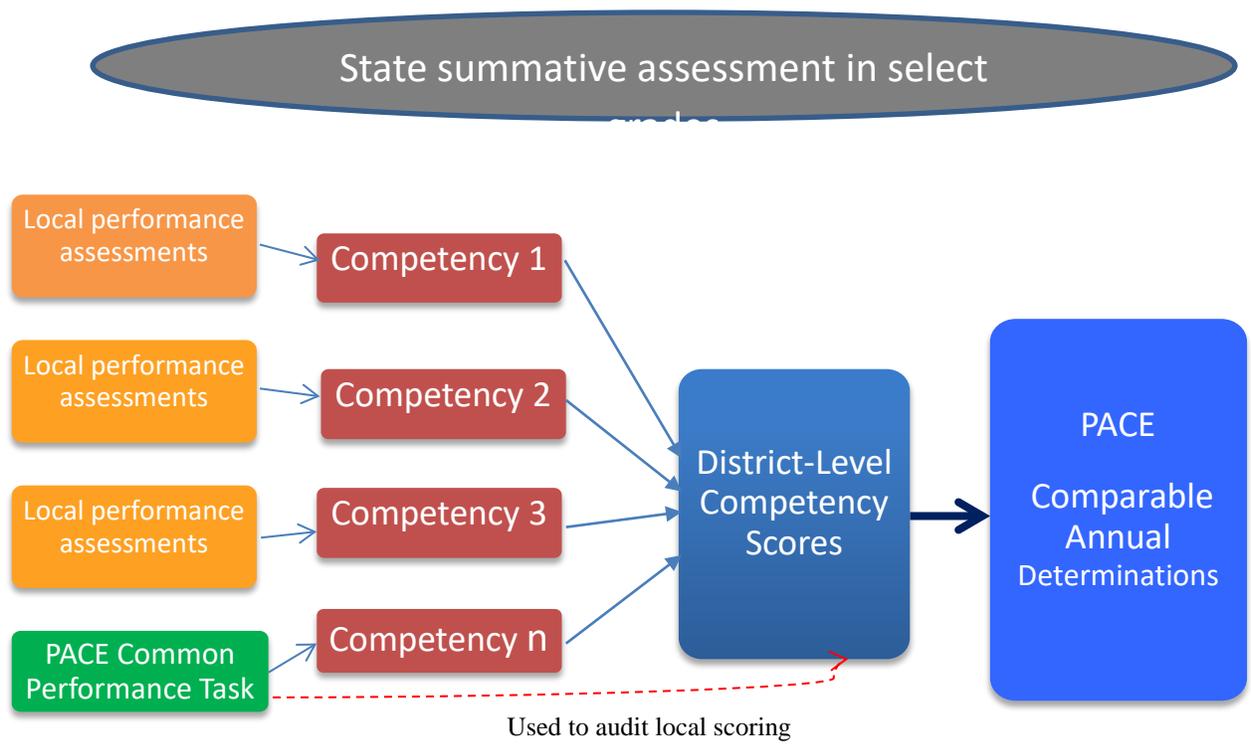


Figure 1. Graphical representation of the PACE assessment system.

Application selection criteria	Required information from the SEA
<p>(a)(2). The plan the SEA or consortium, in consultation with any external partners, if applicable, has to--</p> <p>(i) Develop and use standardized and calibrated tools, rubrics, methods, or other strategies for scoring innovative assessments throughout the demonstration authority period, consistent with relevant nationally recognized professional and technical standards, to ensure inter-rater reliability and comparability of innovative assessment results consistent with 34 CFR part 200.105(b)(4)(ii), which may include evidence of inter-rater reliability; and</p> <p>(ii) Train evaluators to use such strategies, if applicable;</p>	<p>A plan to continue, for all participating PACE LEAs, during the period of the demonstration authority:</p> <ol style="list-style-type: none"> 1. Performance standards validations. 2. Local scoring audit activities (known as body of work samples).

New Hampshire’s Response:

While the comments on the strengths of New Hampshire’s approach to training and scoring are overwhelmingly positive and the peers note how the practices employed adhere to relevant professional and technical standards, two of the five peer reviewers of New Hampshire’s IADA application questioned New Hampshire’s plan to discontinue the use of the Body of Work method to validate the performance standards. Multiple standard setting methods are often employed when setting cut scores for educational or psychological assessments. For PACE, the contrasting groups standard setting is the first and primary method of setting cutscores. The second method, known as Body of Work, is designed to help validate the standards derived from the contrasting groups methodology. In its initial submission for the IADA, New Hampshire indicated that we may consider phasing out the Body of Work standards validation after a number of years of demonstrated success. This quality control procedure is intended to verify our own processes and procedures for producing annual determinations, and therefore may not be necessary to continue once the validity of the processes for creating PACE annual determinations has been well documented. Additionally, the Body of Work standards validation places a heavy burden on participating teachers, schools, and districts. As PACE scales statewide, we need to make programmatic decisions that can maintain the technical quality of the innovative assessment system while also trying to minimize the additional administrative burden of the innovative system on schools over time. In order to address the reviewers’ concerns we will continue using of the Body of Work standards validation method for all new districts entering PACE until we have gathered credible evidence supporting the validity of the PACE annual determinations.

To clarify, the Body of Work method is the standards validation method while the local scoring audit is covered in the calibration studies discussed throughout this document.

Application selection criteria	Required information from the SEA
<p>(a)(3). If the system will initially be administered in a subset of schools or LEAs in a State--</p> <p>(i) The strategies the SEA, including each SEA in a consortium, will use to scale the innovative assessment to all schools statewide, with a rationale for selecting those strategies;</p> <p>(ii) The strength of the SEA’s or consortium’s criteria that will be used to determine LEAs and schools that will initially participate and when to approve additional LEAs and schools, if applicable, to participate during the requested demonstration authority period; and</p> <p>(iii) The SEA’s plan, including each SEA in a consortium, for how it will ensure that, during the demonstration authority period, the inclusion of additional LEAs and schools continues to reflect high-quality and consistent implementation across demographically diverse LEAs and schools, or contributes to progress toward achieving such implementation across demographically diverse LEAs and schools, including diversity based on enrollment of subgroups of students described in section 1111(c)(2) of the Act and student achievement. The plan must also include annual benchmarks toward achieving high-quality and consistent implementation across participating schools that are, as a group, demographically similar to the State as a whole during the demonstration authority period, using the demographics of initially participating schools as a baseline.</p>	<p>A projected schedule for the inclusion of additional LEAs into the PACE pilot assessment that includes specific targets/goals for expansion during each year of the demonstration period.</p>

New Hampshire’s Response:

There are currently 13 LEAs fully implementing the PACE assessment system and another 12 school districts evaluating implementation within the next year or two. While NH has approximately 175 individual school districts, many smaller districts are organized into school administrative unions (SAU) to hire a single superintendent and employ a common central office staff. Since the participation of the superintendent is critical to the success of PACE, we count participation at the SAU level. Since all participating districts employ a mix of the NH State Assessment System (NH SAS) and PACE (see Figure 1 above), NH DOE will count a district as a participant in the PACE assessment system if it is implementing PACE in at least one grade span for at least one content area. When a school / district has such a limited implementation, the remaining schools in the district will be using NH SAS. Additionally, schools and districts will be encouraged to try PACE by using PACE common tasks with students even if the grade/subject at that school is participating in NH SAS.

NH DOE has been waiting until the Innovative Assessment Demonstration Authority opportunity to begin trying to scale PACE. Operating under a waiver from NCLB and ESSA did not provide enough security for NH and its school districts to begin large-scale expansion. Operating under the IADA will provide that opportunity for NH. That said, we propose

continuing to expand gradually in the first years to ensure that we have the necessary processes and tools in place to support full expansion.

One of the ways to facilitate the expansion of PACE is to offer PACE tasks as an alternative means of demonstrating competency for students in schools that are otherwise participating in the NH SAS. This option will likely require students to complete multiple PACE tasks if the school pursues this option in order to produce a stable estimate of the student's achievement for that grade/subject.

Application selection criteria	Required information from the SEA
<p>(b)(1). The extent and depth of prior experience that the SEA, including each SEA in a consortium, and its LEAs have in developing and implementing the components of the innovative assessment system. An SEA may also describe the prior experience of any external partners that will be participating in or supporting its demonstration authority in implementing those components. In evaluating the extent and depth of prior experience, the Secretary considers—</p> <p>(i) The success and track record of efforts to implement innovative assessments or innovative assessment items aligned to the challenging State academic standards under section 1111(b)(1) of the Act in LEAs planning to participate; and</p> <p>(ii) The SEA’s or LEA’s development or use of--</p> <p>(A) Effective supports and appropriate accommodations consistent with 34 CFR part 200.6(b) and (f)(1)(i) and section 1111(b)(2)(B)(vii) of the Act for administering innovative assessments to all students, including English learners and children with disabilities, which must include professional development for school staff on providing such accommodations;</p> <p>(B) Effective and high-quality supports for school staff to implement innovative assessments and innovative assessment items, including professional development; and</p> <p>(C) Standardized and calibrated tools, rubrics, methods, or other strategies for scoring innovative assessments, with documented evidence of the validity, reliability, and comparability of annual summative determinations of achievement, consistent with 34 CFR part 200.105(b)(4) and (7).</p>	<p>See information requested under (a)(3) above.</p>

New Hampshire’s Response:

Please see (a)(3) above regarding the additional information requested.

Application selection criteria	Required information from the SEA
<p>(b)(2). The extent and depth of SEA, including each SEA in a consortium, and LEA capacity to implement the innovative assessment system considering the availability of technological infrastructure; State and local laws; dedicated and sufficient staff, expertise, and resources; and other relevant factors. An SEA or consortium may also describe how it plans to enhance its capacity by collaborating with external partners that will be participating in or supporting its demonstration authority. In evaluating the extent and depth of capacity, the Secretary considers--</p> <p>(i) The SEA’s analysis of how capacity influenced the success of prior efforts to develop and implement innovative assessments or innovative assessment items; and</p> <p>(ii) The strategies the SEA is using, or will use, to mitigate risks, including those identified in its analysis, and support successful implementation of the innovative assessment.</p>	<p>Provide specific examples of successful risk mitigation (from previous PACE experience) or provide descriptions of strategies for mitigating the risks associated with implementing the innovative assessment system.</p>

New Hampshire’s Response:

There are generally few surprises with PACE because of the monthly meetings with district leadership and the even more regular check-in opportunities with the “content-leads” or those teachers leading the task development work. As noted above, the monthly district leadership meetings are a collaborative way to get regular updates on implementation progress and to discuss any challenges. This approach allows us to head off any problems well before the end of the year. There are two main classes of risks with PACE and while we have not had issues thus far, we discuss each to describe the ways in which we preemptively address potential threats.

- ✓ **Student Participation in the Assessment System:** As a reminder, PACE annual determinations are based on assessment information gathered throughout the school year, so if a student moves into a school district late in the year, she/he would not have enough data to calculate an annual determination. Therefore, the PACE leadership team determined that such students participate in the NH SAS. Schools may also decide to use NH SAS for students in PACE schools as another means of demonstrating competence. This may be especially beneficial for students who struggled with the required knowledge and skills earlier in the school year or otherwise struggle to demonstrate competence through the use of a performance assessment.

- ✓ **Data Collection:** The PACE innovative assessment system requires a fairly robust data collection from districts in order to evaluate comparability and produce annual determinations. A key aspect of the data collection is the samples of student work used for both the calibration analyses and the Body of Work cutscore audit (different samples of work). Initially, districts were responsible for scanning and uploading the work samples, but this was too much of a burden for many participating districts. We then shifted to having the work shipped to the Center for Assessment, but that quickly threatened to overwhelm the Center's scanning capabilities. We have shifted this past year to contracting with Measured Progress, a major test vendor, to receive all documents and scan all materials using commercial scanners. This process has improved the quality of the scanned and uploaded documents considerably and has proven very efficient.
- ✓ Finally, by administering the NH SAS in certain years provides NH DOE a degree of assurance that the student performance is on track.

These cases are illustrative of how the PACE leadership team closely monitors implementation risks and is able to head them off before they threaten the quality of the project.

Application selection criteria	Required information from the SEA
<p>(c)(1). The extent to which the timeline reasonably demonstrates that each SEA will implement the system statewide by the end of the requested demonstration authority period, including a description of--</p> <ul style="list-style-type: none"> (i) The activities to occur in each year of the requested demonstration authority period; (ii) The parties responsible for each activity; and (iii) If applicable, how a consortium’s member SEAs will implement activities at different paces and how the consortium will implement interdependent activities, so long as each non-affiliate member SEA begins using the innovative assessment in the same school year consistent with 34 CFR part 200.104(b)(2); 	<p>NH DOE must provide:</p> <ul style="list-style-type: none"> 1. A timeline for activities during the demonstration authority period designed to scale up the number of districts toward a statewide implementation of the innovative assessment system was provided (e.g., recruitment activities). 2. A plan and timeline for conducting research studies in response to the recommendations from the external evaluation was provided. (This may also be addressed in the information requested in (e)(1) below.)

New Hampshire’s Response:

1. The NH DOE and the PACE leadership believe that in a small state like NH, the best recruitment and onboarding strategies involve personal conversations among district and school leaders, PACE leaders, and NH DOE leadership. The Commissioner of Education and other key leaders from the NH DOE meet monthly with all NH district superintendents to update the field on various initiatives.
 - a. NH DOE, along with its partners at NEA NH, NHLI, and the Center for Assessment, have expanded the availability of state workshops focusing on performance assessment for deeper learning to teams of educators beyond the PACE consortium schools. These in-person workshops will be offered multiple times throughout the first two years while NH DOE and its partners, particularly the Center for Assessment, work to move these workshops online.
 - b. Similarly, the Center for Assessment is creating a “performance assessment toolkit” that will be widely shared within existing PACE schools as well as across the state to help non-PACE schools learn how to design and use rich performance tasks.
 - c. One of the keys to increasing the reach of PACE in an affordable manner is a technology platform that will support asynchronous task development, scoring, calibration, data collection, and psychometric analyses. The New Hampshire Learning Initiative (NHLI) is currently in discussions with a technology company that should lead to a platform to support the tasks outlined above.
2. As discussed on the recent telephone conference call with ED, HumRRO completed an evaluation of the theory of action and processes supporting PACE in March 2017. We present HumRRO’s recommendations in italics below and then following each HumRRO

recommendation, we present the actions being taken by the NH DOE and the PACE leadership team.

- a. Considering that the PACE leadership team is still working through the recommendations from the 2017 evaluation, we argue that it is not prudent to engage in another formal evaluation until we have a few more years under our belt and we feel confident that the HumRRO recommendations have been addressed.
- b. Participating PACE districts and NH DOE recognize that they will need to participate in a standards and assessment peer review of PACE, which will serve as a thorough evaluation of many of the PACE processes and outcomes.
- c. That said, NH DOE and the Center for Assessment conduct numerous analyses each year and provide feedback to school districts on such things as interrater reliability, cross-district comparability, Body of Work audits, local assessment reviews, assessment map reviews, and regular feedback on Common Task quality. These analyses are memorialized in a yearly technical report, but the most important aspect of this work is the feedback provided to district leadership and educators so they are able to improve their practices. We have evidence that districts have improved their practices dramatically, particularly interrater reliability and cross-district comparability, as a result of yearly feedback provided to each district.

HumRRO Recommendations (2017)

Our evaluation found that PACE is currently functioning largely as intended. The recommendations included here call for additional monitoring or minor improvements to current processes. As the system expands, more substantial changes may become necessary, but this evaluation does not indicate a need for major modifications at this time.

Recommendation 1: Monitor and Support District Engagement

PACE should regularly gauge local leadership support and target interventions when district leaders voice concerns or reduce their district's involvement with the program. PACE has done this for one district by helping support a PACE coordinator within the district with experienced consultants. As the program expands, these checks and interventions should become more routinized to ensure that all districts maintain adequate support for the educators implementing the program.

✓ Ongoing

- The monthly PACE Leadership meetings provide a regular check on district engagement. If any concerns or issues are detected, more directed actions are taken with the district.

Recommendation 2: Evaluate Effectiveness of Collaboration Methods

PACE should evaluate the effectiveness of the new collaboration methods. While task development meetings with teachers from all Tier 1 districts were becoming unwieldy, one of the attributes teachers reported as positive was having direct input into the program. Findings from

the survey indicate that those teachers who had not participated in cross-district collaborations tended to have less favorable ratings of PACE. If the new collaboration methods reduce opportunities for cross-district collaborations, then teachers may perceive less personal value in PACE. Regular monitoring and adjustments can help safeguard against this potential issue.

✓ Ongoing

- New collaboration methods have not yet been introduced in light of the caution called for by this recommendation. However, as PACE expands and new technology-based collaboration approaches are required, the PACE leadership team will closely monitor through surveys and focus groups the engagement of participating educators.

Recommendation 3: Consider Additional Training/Supports for Teachers Not Directly Involved in Common Task Development

As the percentage of PACE participants directly involved in future common task development decreases (either through including a smaller number of teachers in a meeting or by expanding into additional districts), the professional development and training stemming from those activities may need to be supplemented with additional training.

✓ This year

- PACE Teacher Leaders, content leads, and task developers have been provided instructions and supports to better transmit institutional knowledge to all teachers in their respective districts.
- The Libguides have been used to share broadly all key documents and resources.

✓ Next year

- Expanding opportunities for performance assessment development training for all interested NH schools and districts.
- Developing set of common resources for assessment literacy across all levels of PACE participation.

Recommendation 4: Infuse Equity and Accommodations Training into PACE Activities

Include training on scaffolding and accommodations as part of the regular schedule of PACE activities. Despite quality documentation and training, teachers continued to report uncertainty regarding equity issues, especially for accommodating students with disabilities (SWD). Scaffolding should be available to all students, including SWD, and is currently built into task development activities.

✓ Ongoing

- This is a continuing area of work and emphasis for the PACE leadership. All content leads (the teacher leads responsible for task development) have been trained on the use of Universal Design for Learning (UDL) and the use of accommodations and/or other supports are listed on the task templates.

Additionally, the project assessment leaders have provided additional training tools on the use of UDL to support increased fairness and accessibility.

Recommendation 5: Investigate the Impact of Reading/Writing Requirements on Accessibility

Investigate the impact of the reading and writing demands of the PACE tasks on accessibility and student performance. If, for instance, we are interested in knowing whether students understand and can perform computations associated with a mathematics concept, including a long reading passage to set up the task might interfere with a student demonstrating her math abilities. We recommend examining score patterns among the PACE tasks, course grades, and performance on comparison measures (e.g., Smarter Balanced) for students with and without disabilities as one way to investigate whether the reading and writing requirements may be impacting students' scores.

✓ Ongoing

- Similar to the response to recommendation #4 above, this is a continuing area of work and emphasis for the PACE leadership and relies on thoughtful employment of Universal Design for Learning principles and techniques.

Recommendation 6: Routinize Timely Reviews of Local Performance Tasks

Evaluate the quality of the locally developed performance tasks and rubrics. As the pool of locally developed tasks expands, it is important to ensure that the tasks and rubrics are of sufficient quality to be used to generate student scores and annual determinations. Teachers report that their skill level in developing these tasks improves with each year of PACE participation, so it stands to reason that the validity and reliability of students' scores should improve with time.

✓ This year

- The Center for Assessment provides on-going training to build the cadre of experts available to review a sample of tasks from each participating district.

✓ Next year

- Expand the use of the peer and expert review approach and work to move this online so it can be completed asynchronously.

Recommendation 7: Plan for Future Research on the Impact of PACE on Teaching and Learning

The positive impacts of PACE on teaching and learning should continue to be externally verified beyond this evaluation. This may be part of a future research agenda when it becomes possible to evaluate the predictive strength of PACE results on college and career performance. In the interim, it may be possible to compare PACE versus non-PACE student performance on Smarter Balanced assessments, college entrance exams, or other measures.

✓ On-going

- Annual evaluation of student performance on standardized assessments for both achievement and growth.
- ✓ Subsequent years
 - Seeking funding from philanthropies to more deeply understand the connection between learning and engagement in complex performance assessments.
 - Begin to longitudinally track trends in career and college readiness (e.g., persistence in college), but this is dependent upon being able to gather quality data from NH's Institutions of Higher Education.

Recommendation 8: Evaluate the Benefit of Time in Program on Outcomes

As the system expands, it may be possible to investigate the benefits of time in the program on instructional practice and student learning. It would not be surprising if there was a direct correlation between years in the program and benefits, both perceived and realized, on assessment practice and student learning. We would not expect this correlation to be perfect, however. Contextual factors such as district size, fidelity of implementation, and the effectiveness of district or school teams could certainly impact the effects of time in the program.

- ✓ This year
 - We have begun conducting research into the potential influence of time in PACE on student outcomes and initial results are promising, especially for students with disabilities (Evans, 2017). However, due to the non-random inclusion of districts/schools in PACE, we must approach such analyses cautiously.

Recommendation 9: Consider Systematically Recycling Tasks

After the operational year, common tasks may still be used in place of, or in addition to, local tasks. PACE should consider some method of systematically repeating tasks across years as another check on the consistency of scoring. If tasks were repeated, previously scored “check sets” of student work from the prior year could be included in the current year. Score consistency across years could then be checked in a more systematic way.

- ✓ This year
 - We will be working with the PACE content leads to develop plans for task recycling. This includes relying on the larger number of teachers involved in task development to develop and field test multiple tasks for each subject/grade combination during this year's task development cycle.
- ✓ Subsequent years
 - We will continue this process of adding to the task bank each year in order to continue to grow the number of tasks available for local use. Such tasks will include the rubrics, teacher materials, and annotate samples of student work. The highest quality tasks will be reserved from the main task bank for potential reuse as operational tasks.

Recommendation 10: Begin Tracking Performance from Year to Year

The PACE system has the potential for variability across years. Comparing performance across years will allow PACE to see where there are large changes in the proportions of students at each achievement level in any district and to investigate potential reasons for those changes. Early reports to USED comparing student performance on PACE with performance on Smarter Balanced within and across years, as well as the data analyses completed for this evaluation, should be repeated annually. This will allow for continuous monitoring and by investigating anomalous results, PACE may be better able to identify potential threats to reliability and validity. Note: These analyses have now been conducted and are discussed on pages 8-9 of this document and are explained in great detail in Appendices G & H.

✓ On-going

- This has become a regular part of our analyses, both in terms of tracking student longitudinal performance, especially as students move from PACE to the state summative assessment and vice versa, as well as changes in cohort performance at the school and district levels.

End Goal: Students are College and Career Ready

Graduating students who are college and career ready is the ultimate goal of PACE. While we have found considerable evidence supporting the interim goals of PACE, it is still too early to evaluate college and career readiness. Once PACE has matured sufficiently and there are students who experienced both the PACE program and at least one year of college or career, we recommend that PACE support an ongoing research agenda to investigate claims under this ultimate goal.

Application selection criteria	Required information from the SEA
<p>(c)(2). The adequacy of the project budget for the duration of the requested demonstration authority period, including Federal, State, local, and non-public sources of funds to support and sustain, as applicable, the activities in the timeline under paragraph (c)(1) of this section, including--</p> <p>(i) How the budget will be sufficient to meet the expected costs at each phase of the SEA’s planned expansion of its innovative assessment system; and</p> <p>(ii) The degree to which funding in the project budget is contingent upon future appropriations at the State or local level or additional commitments from non-public sources of funds.</p>	<p>NH DOE must provide:</p> <ol style="list-style-type: none"> 1. A projected budget for each year of the demonstration authority period considered in the application. 2. A projected budget for planned evaluation activities (see also (e)(1) below).

New Hampshire’s Response:
See attached Excel file.

Application selection criteria	Required information from the SEA
<p>(d)(1). The extent to which the SEA or consortium has developed, provided, and will continue to provide training to LEA and school staff, including teachers, principals, and other school leaders, that will familiarize them with the innovative assessment system and develop teacher capacity to implement instruction that is informed by the innovative assessment system and its results;</p>	<p>NH DOE must provide:</p> <ol style="list-style-type: none"> 1. A description of the training or support that is provided to PACE teachers regarding their making appropriate linkages between the student performance on the assessment tasks and instruction in class. 2. A description of the specific training requirements that all participating PACE teachers must complete prior to administering pilot assessments. This description should include information regarding teachers who do not complete required training in terms of PACE participation.

New Hampshire’s Response:

New Hampshire appreciates the careful review of the peer reviewers, in particular their acknowledgement of the extensive and comprehensive investment into high-quality professional development and training of participating teachers. New Hampshire recognizes the central importance of this work to the success of improving instruction and student outcomes through PACE. One reviewer was unclear about the particular support provided to teachers regarding making appropriate linkages between student performance on the assessment tasks and instruction in class. While using student work to make personalized adjustments to instruction is a clear benefit of performance-based assessments (as opposed to highly secure and less informative standardized assessments), this particular use is just one of the instructional benefits of PACE. The primary theory of action is that by implementing high-quality, complex performance assessments throughout the school year, teachers will need to transform their instruction so that students are better prepared to succeed on these types of authentic and extended assessment experiences. PACE teachers have received extensive training on Center for Assessment-developed protocols for examining and analyzing student work. Close examination of student work reveals areas where students clearly understand the required knowledge and skills as well as areas where they may still be struggling with the content. Such assessment-instruction connections can be made only with assessments where teachers have the opportunity to interrogate student work rather than looking at multiple-choice options. To this end, the PACE training in performance assessment design pays close attention to the instructional context in which the performance assessments will be embedded. In developing the performance tasks, PACE teachers include information for implementing teachers that is intended to help inform the instruction leading up to the performance tasks such as particular skills students should have had an opportunity to practice. Increasingly common is the inclusion of formative assessment ideas and materials that are provided directly within the performance task template to help all implementing teachers make the instructional shifts that are anticipated as a result of PACE.

All teachers who administer PACE tasks are required to be trained to do so either as a result of their work on task development committees or locally as part of locally-required training for administering the PACE tasks. Further, the task development committees write extensive teacher directions for preparing for and administering the task. These directions even include tips for embedding the task in appropriate instructional units. Each participating PACE district and

school agrees to ensure that all administering teachers receive training on PACE Common Tasks that they will administer.

Additionally, there is a comprehensive suite of training options and opportunities for PACE teachers with varying levels of depth and commitment. All participating districts send teachers to the PACE summer institute, and all districts are encouraged to have at least one teacher leader and content leader representative who are trained more deeply and can serve as PACE ambassadors in the local setting. Additionally, as of summer 2018, training opportunities in performance assessment are being extended to all New Hampshire educators. We appreciate the reviewers' positive feedback related to the comprehensiveness of the system of training as this is some of the most extensive and powerful work that the NH DOE is engaged in related to PACE. While the majority of these training opportunities are optional, at a minimum, all participating teachers must engage in their own district's scoring calibration sessions. Any teacher who is administering and scoring a PACE common performance task must contribute student work and participate in calibrating their scoring practices. Even in small districts where teachers may be the only teacher for a given course, is the clear expectation that all teachers are dedicating professional time to look at student work and engage in calibration sessions with their colleagues. To audit this practice, the state requires that a sample of the common tasks be double-blind scored within district to monitor inter-rater reliability of within-district scoring. Results of the inter-rater reliability analyses provide overwhelming support for the degree of inter-rater consistency in scoring of the PACE Common Tasks with the average exact agreement on the scores for each rubric dimension of the common task greater than 75 percent. This evidence suggests that teachers within districts are able to successfully conduct calibration sessions and comparably evaluate student work.

Application selection criteria	Required information from the SEA
(d)(2). The strategies the SEA or consortium has developed and will use to familiarize students and parents with the innovative assessment system;	NH DOE must provide: <ol style="list-style-type: none"> 1. A description of standardized collateral materials about PACE and standardized recommendations to support LEAs in communicating with parents about PACE. This information should reference the information requested under (a)(1) above. 2. A description of how the State and LEAs will familiarize students with the PACE, in terms of both how the tasks and rubrics work in practice as well as how their performance on the tasks accrues to an annual proficiency score. This information should reference the information requested under (a)(1) above.

New Hampshire’s Response:

In compiling the IADA application and in receiving the reviewers’ feedback, New Hampshire has realized that communication with parents and students is an area in need of attention and improvement. New Hampshire acknowledges the helpful suggestions of the reviewers in establishing common templates and resources that LEAs can use when communicating with their parents and students. Pending Section 1204 approval, the NH DOE will be launching a new, public-facing website landing page for the PACE assessment system. On this website the NH DOE is committed to providing:

- a video explaining what PACE is as an assessment system and its role in changing instruction;
- a downloadable PDF brief paper that is written in non-technical language that explains more detail about the mechanics of the PACE assessment system and its role within the state accountability system ;
- a PowerPoint template that schools can use to familiarize students and parents with PACE;
- a performance assessment toolkit that provides common resources on task design, task quality, and rubrics; and
- the PACE annual technical report detailing all of the analyses to support the validity of the PACE assessment system.

Further, LEAs will be encouraged to offer “PACE nights” which would be open house-type of events where parents are invited in and PACE is explained to them. To further engage the parents, such “PACE nights” would allow parents the opportunity to participate in a PACE (perhaps with their child’s assistance) and/or to use tools for analyzing student work.

The NH DOE will continue to use the monthly district meetings to check in with districts regarding communication about the innovative assessment system to discuss common challenges, brainstorm solutions, and continue to build out and update the resources available on the website.

Application selection criteria	Required information from the SEA
<p>(d)(3). The strategies the SEA will use to ensure that all students and each subgroup of students under section 1111(c)(2) of the Act in participating schools receive the support, including appropriate accommodations consistent with 34 CFR part 200.6(b) and (f)(1)(i) and section 1111(b)(2)(B)(vii) of the Act, needed to meet the challenging State academic standards under section 1111(b)(1) of the Act;</p>	<p>NH DOE must clearly describe teachers will receive training and support in implementing appropriate accommodations when administering performance tasks.</p>

New Hampshire’s Response:

The reviewer’s comments were overwhelmingly positive on the strengths of New Hampshire’s approach to ensuring that all students and each subgroup of students in participating PACE schools receive the support, including appropriate accommodations consistent with the federal regulations, needed to meet the challenging State academic standards. Two of the five reviewers of New Hampshire’s IADA application inquired about how all teachers in participating PACE schools receive training and support in Universal Design for Learning and implementing appropriate accommodations when administering performance tasks.

PACE Common Tasks are developed using a principled assessment design approach that incorporates the principles of Universal Design for Learning in the task template and task development process. Content leads along with the teachers involved in task development are trained in this process of principled assessment design which includes Universal Design for Learning. The task template also specifies what should be included in the teacher instructions that accompanies the performance tasks, including a description of the accommodations for students with disabilities and English learners. All teachers implementing a PACE Common Task are instructed to read the teacher instructions prior to administration and there is also a PACE Accommodations Manual that is identical to the accommodation standards on the statewide academic assessment (NH SAS).

In addition, participating PACE schools and districts indicate which students received accommodations on the PACE Common Task when they upload their scores into the state system. Starting in the 2018-19 school year, participating PACE schools and districts will be asked to specify exactly what accommodations were provided on the PACE Common Task using the same allowable list of accommodations on the NH SAS.

Application selection criteria	Required information from the SEA
<p>(d)(4). If the system includes assessment items that are locally developed or locally scored, the strategies and safeguards (e.g., test blueprints, item and task specifications, rubrics, scoring tools, documentation of quality control procedures, inter-rater reliability checks, audit plans) the SEA or consortium has developed, or plans to develop, to validly and reliably score such items, including how the strategies engage and support teachers and other staff in designing, developing, implementing, and validly and reliably scoring high-quality assessments; how the safeguards are sufficient to ensure unbiased, objective scoring of assessment items; and how the SEA will use effective professional development to aid in these efforts.</p>	<p>NH DOE must provide:</p> <ol style="list-style-type: none"> 1. Evidence that sufficient quality control procedures exist for the scoring of local tasks which are equivalent to quality control processes used for scoring common tasks (this may be partially addressed by information requested under (4)(i) and (a)(2) above). 2. Evidence of a process where all locally developed tasks and assessments are reviewed for quality (such as by another educator). This evidence should address how the local task review process is consistent with professional standards and practice for student assessment.

New Hampshire’s Response:

All of the evidence gathered to support the PACE system is related to the validity and reliability of the PACE annual determinations. Students participating in PACE do not receive scale scores, but annual determinations that are designed to be comparable to those offered by the statewide assessment. Of course, more nuanced information about student performance is available in an on-going way at the local level based on student performance on the common and local performance assessments and other local assessment. The state gathers validity evidence to support the scores it reports, which for PACE, is the annual determination. As noted by Reviewer 2, “Overall, the application provides sufficient evidence to conclude that the NH DOE PACE processes, if followed with fidelity, will produce results that are valid and reliable for their intended purposes.” This sentiment is appreciated and reflects the intense amount of work that the state and districts do together to gather the necessary evidence to support the PACE system. However, even Reviewer 2, who expressed confidence in the processes we have in place to support the validity and reliability of our reported annual determinations, questioned New Hampshire’s methods for reviewing local performance tasks and their scores. All five reviewers shared this concern.

New Hampshire is confident in the adequacy of the quality controls in place to ensure the reported annual determinations are reliable, valid, and comparable. The HumRRO evaluation report recommended several areas where quality control processes could be improved, including rigor around locally developed tasks. The NH DOE and PACE Leadership team have begun to implement responses to those recommendations in order to continue to improve the quality of the PACE processes. Part of the evidence supporting the quality of the annual determinations are expert and peer reviews of a sample of local performance tasks from each district (i.e., local

assessment audit) and the scoring practices on local tasks (i.e., generalizability studies). These audits are designed to support inferences about the “population” of tasks based on a sampling of tasks. Alignment studies for traditional standardized assessments also rely on samples of items and forms to make inferences about the quality of the assessment program for measuring the intended content. The PACE system would be untenable for both the state and the participating schools if every task contributing to student competency scores needed to be reviewed by the state. Fortunately, we have strong evidence that the sampling techniques we have employed are working—see technical manual for high degrees of concurrent and non-concurrent comparability with the statewide assessment and more than sufficient reliability in student competency scores as documented by the generalizability analyses, and the HumRRO external evaluation report for strong convergent and discriminant validity evidence.

Though New Hampshire believes the evidence we have to support the quality of our reported PACE annual determinations is sufficient, we acknowledge the concerns of all five reviewers. Therefore, New Hampshire is committed to engaging in a discussion with the participating PACE districts about how local assessments that contribute to students’ competency scores could be peer-reviewed within and/or across districts. We believe this type of discussion could lead to fruitful new professional practices that have the potential to more quickly and effectively raise the assessment literacy of PACE teachers, which is a primary intended outcome of the PACE pilot.

Application selection criteria	Required information from the SEA
(e)(1). The strength of the proposed evaluation of the innovative assessment system included in the application, including whether the evaluation will be conducted by an independent, experienced third party, and the likelihood that the evaluation will sufficiently determine the system’s validity, reliability, and comparability to the statewide assessment system consistent with the requirements of 34 CFR part200.105(b)(4) and (9);	NH DOE must provide a specific plan and timeline to conduct an external evaluation of the innovative assessment system during the course of the demonstration period.

New Hampshire’s Response:

As noted in our response to the request for additional information under Application Criteria (c)(1), HumRRO conducted an evaluation of the theory of action and processes supporting PACE in March 2017. As noted in the response to Application Criteria (c)(1), NH DOE and the PACE leadership are systematically working to address the major recommendations in the HumRRO evaluation report. It would be premature to engage in another evaluation for at least another few years. Also as noted in the response to Application Criteria (c)(1), NH DOE and its partners have developed an extensive and systematic continuous improvement process to feed useable data back to district leaders and educators. Independent evaluations are an important check on the system, but it is the yearly data collection, analyses, and feedback that will lead to greatest improvements in PACE implementation. NH DOE will work to try to raise external funds to support an external evaluation in 2021-2022, which will be approximately halfway through the IADA.

Application selection criteria	Required information from the SEA
<p>(e)(2). The SEA’s or consortium’s plan for continuous improvement of the innovative assessment system, including its process for--</p> <ul style="list-style-type: none"> (i) Using data, feedback, evaluation results, and other information from participating LEAs and schools to make changes to improve the quality of the innovative assessment; and (ii) Evaluating and monitoring implementation of the innovative assessment system in participating LEAs and schools annually. 	<p>NH DOE must provide:</p> <ul style="list-style-type: none"> 1. A description of how it will monitor how continuous improvement feedback is implemented by participating PACE LEAs (this includes feedback from activities requested under (a)(2) above). 2. A description of how it will annually assess the satisfaction and attitudes of educators in participating PACE LEAs regarding PACE activities (this may be part of the external evaluation plan requested in (e)(1) above).

New Hampshire’s Response:

- 1. As indicated above, NH DOE and its partners have engaged in a systematic continuous improvement process that feeds useful data back to district and school leaders, as well as teachers in the system. NH DOE and its partners are able to monitor the effectiveness of this feedback by observing and documenting the outcomes in the year following the feedback to observe improvements, for example, in cross district comparability.
- 2. NH DOE will conduct an annual survey of leaders and educators from all PACE participating school districts to gain insight into such things as training and preparedness to implement PACE, perceived changes in teaching practices, perceived improvements in student engagement and learning, and suggestions for improving PACE implementation. These data will be used as part of the continuous improvement process for the PACE leadership team to ensure that we continue to best need the needs of participating districts.

Appendix A: PACE District Achievement and Participation Rates 2016-17

Note: these results include a combination of PACE, SBAC, DLM and Science results.

Non-PACE grades and subject areas are highlighted in yellow.

Bethlehem School District 53

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	4%	38%	25%	33%	58%
4	8%	38%	54%	0%	54%
5	0%	37%	47%	17%	63%
6	14%	27%	55%	5%	59%
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	6%	35%	44%	16%	60%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	8%	21%	63%	8%	71%
4	15%	38%	23%	23%	46%
5	7%	27%	63%	3%	67%
6	9%	27%	64%	0%	64%
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	9%	27%	57%	7%	64%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	8%	0%	85%	8%	92%
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
0	8%	0%	85%	8%	92%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	**	**	**	**	**
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	59%	63%	92%	100%	100%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	40%	48%	**	100%	100%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	**	**	**	**	**
WaiverSubgroup - SWD and EconDis - Not EL	**	**	**	**	**

WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	60%	64%	92%	100%	100%

Note: ** Count is below cell size of 10

2017 PACE District Participation Participation Rate

Grade	rea	mat
3	100%	100%
4	100%	100%
5	100%	100%
6	100%	100%
7	-	-
8	-	-
11	-	-
0	100%	100%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included).

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	24%	23%	29%	24%	53%
4	8%	29%	62%	1%	63%
5	10%	23%	60%	8%	67%
6	8%	44%	41%	8%	48%
7	14%	35%	43%	8%	50%
8	21%	23%	40%	16%	56%
9	4%	55%	35%	6%	41%
10	21%	38%	36%	5%	41%
11	15%	19%	47%	19%	66%
0	14%	28%	46%	12%	58%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	12%	41%	44%	3%	47%
4	17%	31%	35%	16%	52%
5	10%	29%	50%	10%	60%
6	9%	36%	35%	20%	55%
7	17%	31%	49%	3%	52%
8	27%	25%	25%	24%	48%
9	15%	38%	42%	5%	47%
10	16%	25%	39%	20%	59%
11	17%	43%	32%	8%	40%
0	16%	34%	39%	12%	51%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	5%	34%	60%	0%	60%
8	3%	34%	52%	12%	64%
9	16%	31%	44%	10%	53%
10	-	-	-	-	-
0	4%	34%	56%	6%	62%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	54%	44%	65%	99%	98%
Race - American Indian or Alaskan Native (Non Hispanic)	42%	58%	**	100%	100%
Race - Asian (Non Hispanic)	50%	52%	62%	98%	97%
Race - Black or African American (Non Hispanic)	27%	19%	35%	98%	96%
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	62%	54%	64%	97%	97%
Race - Two or more races	**	**	**	**	**
Waiver Subgroup - EconDis and EL - Not SWD	15%	20%	24%	97%	95%

WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	47%	40%	55%	96%	96%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	32%	38%	**	97%	89%
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	34%	32%	32%	92%	92%
WaiverSubgroup - SWD and EconDis - Not EL	11%	13%	9%	95%	95%
WaiverSubgroup - SWD and EconDis and EL	23%	15%	**	100%	100%
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	58%	51%	62%	97%	97%

** Count is below cell size of 10

2017 PACE District Participation

Grade	Participation Rate	
	rea	mat
3	100%	99%
4	99%	100%
5	92%	91%
6	100%	100%
7	98%	98%
8	98%	98%
11	93%	93%
0	97%	97%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	18%	25%	30%	27%	57%
4	3%	40%	54%	4%	58%
5	6%	44%	36%	14%	50%
6	3%	45%	52%	0%	52%
7	1%	53%	41%	4%	46%
8	13%	26%	47%	14%	62%
9	7%	28%	38%	28%	66%
10	8%	31%	29%	32%	61%
11	16%	29%	48%	7%	55%
0	8%	38%	44%	10%	54%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	1%	12%	67%	19%	87%
4	14%	49%	28%	10%	38%
5	3%	42%	47%	8%	56%
6	5%	35%	42%	18%	60%
7	4%	32%	63%	1%	64%
8	26%	35%	23%	16%	39%
9	3%	39%	45%	13%	58%
10	4%	10%	66%	20%	86%
11	20%	45%	30%	5%	36%
0	10%	36%	43%	11%	54%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	3%	21%	70%	6%	76%
8	6%	35%	44%	14%	58%
9	0%	11%	67%	22%	89%
10	10%	43%	19%	28%	46%
0	5%	28%	57%	10%	67%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	44%	63%	**	100%	100%
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	54%	54%	67%	98%	98%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**

WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	48%	52%	85%	97%	97%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	27%	25%	57%	98%	98%
WaiverSubgroup - SWD and EconDis - Not EL	0%	20%	38%	93%	93%
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	54%	54%	67%	98%	98%

Note: ** Count is below cell size of 10

2017 PACE District

Participation

Participation Rate

Grade	rea	mat
3	99%	99%
4	100%	100%
5	97%	99%
6	100%	100%
7	96%	100%
8	100%	95%
11	95%	95%
0	98%	98%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	11%	16%	26%	47%	74%
4	9%	27%	50%	14%	64%
5	0%	33%	52%	14%	67%
6	6%	12%	53%	29%	82%
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	6%	23%	46%	25%	71%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	0%	21%	63%	16%	79%
4	5%	23%	32%	41%	73%
5	0%	29%	57%	14%	71%
6	6%	18%	47%	29%	76%
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	3%	23%	49%	25%	75%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	0%	23%	64%	14%	77%
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
0	0%	23%	64%	14%	77%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	**	**	**	**	**
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	71%	71%	72%	99%	99%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	57%	57%	**	100%	100%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	**	**	**	**	**
WaiverSubgroup - SWD and EconDis - Not EL	**	**	**	**	**

WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	71%	75%	77%	99%	99%

Note: ** Count is below cell size of 10

2017 PACE District

Grade	Participation Rate	
	rea	mat
3	100%	100%
4	100%	100%
5	100%	100%
6	94%	94%
7	-	-
8	-	-
11	-	-
0	99%	99%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	0%	33%	0%	67%	67%
4	-	-	-	-	-
5	-	-	-	-	-
6	-	-	-	-	-
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	0%	33%	0%	67%	67%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	33%	33%	0%	33%	33%
4	-	-	-	-	-
5	-	-	-	-	-
6	-	-	-	-	-
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	33%	33%	0%	33%	33%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
0	-	-	-	-	-

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	**	**	-	**	**
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	-	**	**
Race - Asian (Non Hispanic)	**	**	-	**	**
Race - Black or African American (Non Hispanic)	**	**	-	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	-	**	**
Race - White (Non Hispanic)	**	**	-	**	**
Race - Two or more races	**	**	-	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	-	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	**	**	-	**	**
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	-	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	**	**	-	**	**
WaiverSubgroup - SWD and EconDis - Not EL	**	**	-	**	**

WaiverSubgroup - SWD and EconDis and EL	**	**	-	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	-	**	**
All Students	**	**	-	**	**

Note: ** Count is below cell size of 10

2017 PACE District

Participation

Participation Rate

Grade	rea	mat
3	**	**
4	-	-
5	-	-
6	-	-
7	-	-
8	-	-
11	-	-
0	**	**

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	18%	47%	18%	18%	35%
4	0%	60%	32%	8%	40%
5	10%	5%	70%	15%	85%
6	5%	5%	84%	5%	89%
7	2%	55%	34%	9%	43%
8	10%	23%	40%	27%	67%
9	16%	24%	60%	0%	60%
10	4%	19%	78%	0%	78%
11	18%	29%	53%	0%	53%
0	8%	35%	45%	12%	57%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	12%	59%	12%	18%	29%
4	4%	24%	52%	20%	72%
5	20%	5%	55%	20%	75%
6	5%	21%	63%	11%	74%
7	9%	24%	64%	2%	67%
8	7%	17%	30%	47%	77%
9	32%	40%	16%	12%	28%
10	42%	8%	42%	8%	50%
11	18%	59%	24%	0%	24%
0	10%	27%	46%	17%	63%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	8%	28%	60%	4%	64%
8	10%	23%	37%	30%	67%
9	9%	52%	30%	9%	39%
10	4%	4%	72%	20%	92%
0	9%	25%	47%	18%	65%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	**	**	**	**	**
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	57%	63%	65%	98%	99%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	53%	59%	65%	100%	100%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	23%	38%	**	93%	93%
WaiverSubgroup - SWD and EconDis - Not EL	22%	37%	**	95%	100%

WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	57%	63%	65%	98%	99%

Note: ** Count is below cell size of 10

2017 PACE District

Participation

Participation Rate

Grade	rea	mat
3	89%	95%
4	100%	100%
5	100%	100%
6	100%	100%
7	98%	100%
8	100%	100%
11	94%	94%
0	98%	99%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	17%	33%	50%	0%	50%
4	0%	40%	40%	20%	60%
5	0%	60%	20%	20%	40%
6	-	-	-	-	-
7	15%	38%	38%	8%	46%
8	17%	33%	50%	0%	50%
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	10%	43%	38%	10%	48%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	0%	33%	50%	17%	67%
4	0%	0%	40%	60%	100%
5	0%	60%	20%	20%	40%
6	-	-	-	-	-
7	54%	8%	38%	0%	38%
8	33%	33%	33%	0%	33%
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	23%	28%	35%	15%	50%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	0%	0%	100%	0%	100%
8	0%	50%	50%	0%	50%
9	-	-	-	-	-
10	-	-	-	-	-
0	0%	27%	73%	0%	73%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	**	**	**	**	**
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	46%	49%	73%	80%	80%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	**	**	**	**	**
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	**	**	**	**	**

WaiverSubgroup - SWD and EconDis - Not EL	**	**	**	**	**
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	48%	50%	73%	78%	78%

Note: ** Count is below cell size of 10

**2017 PACE District
Participation**

Grade	Participation Rate	
	rea	mat
3	**	**
4	**	**
5	**	**
6	**	**
7	100%	100%
8	**	**
11	**	**
0	78%	78%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	33%	37%	17%	13%	30%
4	10%	53%	35%	3%	38%
5	4%	45%	45%	6%	51%
6	5%	28%	43%	25%	68%
7	12%	32%	47%	9%	56%
8	17%	33%	42%	8%	50%
9	26%	48%	23%	3%	26%
10	-	-	-	-	-
11	15%	27%	42%	15%	58%
0	13%	37%	39%	11%	50%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	7%	50%	23%	20%	43%
4	23%	30%	35%	13%	48%
5	6%	55%	38%	0%	38%
6	0%	39%	39%	22%	61%
7	4%	40%	44%	12%	56%
8	47%	29%	12%	12%	24%
9	7%	66%	12%	16%	28%
10	-	-	-	-	-
11	12%	69%	12%	8%	19%
0	14%	44%	30%	12%	42%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	3%	36%	56%	5%	62%
8	3%	32%	53%	12%	65%
9	16%	59%	22%	3%	24%
10	-	-	-	-	-
0	4%	34%	54%	8%	62%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	43%	38%	**	100%	93%
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	51%	42%	63%	96%	93%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	49%	49%	61%	96%	89%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	27%	18%	**	86%	86%

WaiverSubgroup - SWD and EconDis - Not EL	9%	17%	**	98%	100%
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	50%	42%	62%	96%	93%

Note: ** Count is below cell size of 10

**2017 PACE District
Participation**

Grade	Participation Rate	
	rea	mat
3	97%	100%
4	100%	98%
5	100%	100%
6	98%	100%
7	100%	74%
8	100%	95%
11	76%	76%
0	96%	93%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	-	-	-	-	-
4	-	-	-	-	-
5	-	-	-	-	-
6	-	-	-	-	-
7	3%	59%	38%	0%	38%
8	6%	26%	46%	23%	69%
9	0%	45%	42%	13%	55%
10	21%	38%	35%	6%	41%
11	17%	20%	51%	12%	63%
0	9%	35%	45%	11%	57%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	-	-	-	-	-
4	-	-	-	-	-
5	-	-	-	-	-
6	-	-	-	-	-
7	0%	21%	46%	33%	79%
8	26%	34%	17%	23%	40%
9	30%	13%	43%	13%	57%
10	16%	20%	32%	32%	64%
11	22%	34%	37%	7%	44%
0	16%	30%	34%	21%	55%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	-	-	-	-	-
8	3%	54%	31%	11%	43%
9	0%	46%	21%	33%	54%
10	5%	33%	48%	15%	63%
0	3%	54%	31%	11%	43%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	**	**	**	**	**
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	56%	54%	44%	100%	100%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	46%	42%	**	100%	100%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	**	**	**	**	**
WaiverSubgroup - SWD and EconDis - Not EL	**	**	**	**	**

WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	57%	55%	43%	100%	100%

Note: ** Count is below cell size of 10

2017 PACE District

Participation

Participation Rate

Grade	rea	mat
3	-	-
4	-	-
5	-	-
6	-	-
7	100%	100%
8	100%	100%
11	100%	100%
0	100%	100%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	30%	30%	23%	17%	40%
4	10%	34%	48%	8%	56%
5	9%	31%	53%	7%	60%
6	14%	38%	27%	21%	49%
7	12%	57%	10%	21%	31%
8	21%	33%	39%	7%	46%
9	13%	51%	0%	36%	36%
10	7%	41%	20%	32%	52%
11	23%	24%	46%	7%	53%
0	17%	35%	36%	12%	48%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	7%	18%	67%	8%	75%
4	19%	38%	32%	11%	43%
5	9%	14%	55%	22%	77%
6	15%	47%	20%	18%	37%
7	8%	43%	31%	18%	49%
8	37%	30%	17%	16%	32%
9	46%	27%	0%	26%	26%
10	3%	46%	14%	36%	50%
11	26%	44%	23%	7%	30%
0	18%	33%	35%	14%	49%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	1%	60%	0%	38%	39%
8	11%	60%	19%	9%	29%
9	18%	38%	9%	34%	44%
10	13%	66%	0%	21%	21%
0	6%	60%	10%	24%	34%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	30%	41%	26%	95%	95%
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	75%	70%	58%	100%	100%
Race - Black or African American (Non Hispanic)	37%	33%	**	97%	93%
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	49%	49%	34%	97%	97%
Race - Two or more races	43%	55%	28%	96%	95%
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	44%	47%	28%	98%	99%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	100%	100%
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	17%	22%	9%	89%	88%

WaiverSubgroup - SWD and EconDis - Not EL	10%	16%	8%	89%	89%
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	48%	49%	34%	97%	97%

Note: ** Count is below cell size of 10

2017 PACE District

Grade	Participation Rate	
	rea	mat
3	100%	99%
4	99%	99%
5	100%	100%
6	95%	95%
7	96%	96%
8	98%	98%
11	91%	91%
0	97%	97%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea: English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	17%	22%	32%	29%	61%
4	1%	47%	48%	4%	52%
5	12%	31%	49%	8%	57%
6	10%	26%	63%	1%	64%
7	11%	54%	27%	8%	35%
8	11%	19%	50%	19%	69%
9	1%	38%	51%	10%	62%
10	7%	49%	41%	3%	44%
11	26%	22%	44%	9%	53%
0	13%	32%	44%	11%	55%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	2%	45%	53%	0%	53%
4	10%	33%	30%	26%	57%
5	10%	34%	49%	7%	56%
6	6%	28%	58%	7%	66%
7	7%	27%	64%	3%	67%
8	28%	19%	26%	27%	53%
9	30%	42%	26%	2%	28%
10	5%	56%	37%	3%	39%
11	19%	43%	31%	6%	37%
0	13%	33%	44%	11%	55%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	1%	46%	53%	0%	53%
8	0%	48%	45%	7%	52%
9	1%	53%	36%	10%	46%
10	6%	52%	28%	13%	42%
0	0%	47%	48%	4%	52%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	45%	62%	**	97%	97%
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	79%	43%	**	100%	100%
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	55%	55%	53%	99%	99%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	44%	46%	36%	99%	99%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	16%	21%	32%	95%	95%
WaiverSubgroup - SWD and EconDis - Not EL	9%	20%	8%	100%	100%
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	55%	55%	52%	99%	99%

Note: ** Count is below cell size of 10

2017 PACE District

Participation	Participation Rate	
Grade	rea	mat
3	100%	99%
4	98%	99%
5	100%	100%
6	100%	100%
7	99%	100%
8	100%	100%
11	94%	94%
0	99%	99%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	16%	20%	28%	36%	64%
4	3%	36%	53%	8%	61%
5	17%	38%	45%	0%	45%
6	6%	42%	48%	3%	52%
7	5%	33%	62%	0%	62%
8	5%	14%	32%	50%	82%
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	9%	32%	45%	15%	60%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	0%	58%	42%	0%	42%
4	8%	39%	42%	11%	53%
5	17%	34%	41%	7%	48%
6	13%	23%	55%	10%	65%
7	9%	27%	59%	5%	64%
8	14%	27%	18%	41%	59%
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
0	10%	35%	43%	12%	55%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	3%	11%	86%	0%	86%
8	0%	32%	50%	18%	68%
9	-	-	-	-	-
10	-	-	-	-	-
0	2%	19%	72%	7%	79%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	**	**	**	**	**
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	60%	55%	79%	98%	98%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	50%	43%	**	100%	100%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	19%	25%	**	100%	100%
WaiverSubgroup - SWD and EconDis - Not EL	**	**	**	**	**

WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	60%	55%	79%	98%	98%

Note: ** Count is below cell size of 10

2017 PACE District

Participation

Participation Rate

Grade	rea	mat
3	100%	96%
4	100%	100%
5	94%	94%
6	100%	100%
7	95%	100%
8	100%	100%
11	-	-
0	98%	98%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Rea:English Language Arts: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	-	-	-	-	-
4	-	-	-	-	-
5	-	-	-	-	-
6	-	-	-	-	-
7	-	-	-	-	-
8	-	-	-	-	-
9	1%	46%	46%	8%	54%
10	1%	26%	64%	9%	74%
11	6%	13%	57%	25%	82%
0	6%	13%	57%	25%	82%

Note: Values may not sum to 100% due to rounding.

Mat: Mathematics: 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
3	-	-	-	-	-
4	-	-	-	-	-
5	-	-	-	-	-
6	-	-	-	-	-
7	-	-	-	-	-
8	-	-	-	-	-
9	28%	6%	47%	19%	66%
10	5%	38%	32%	25%	57%
11	9%	37%	41%	13%	54%
0	9%	37%	41%	13%	54%

Note: Values may not sum to 100% due to rounding.

Science 2017 PACE District Results by Grade and Level

Grade	Percent at Level 1: Does Not Meet the Achievement Level	Percent at Level 2: Approaching the Achievement Level	Percent at Level 3: Meets the Achievement Level	Percent at Level 4: Exceeds the Achievement Level	Percent at Level 3 & 4: Meets or Exceeds the Achievement Level
4	-	-	-	-	-
8	-	-	-	-	-
9	7%	55%	29%	8%	38%
10	10%	45%	22%	23%	45%
0	100%	0%	0%	0%	0%

Note: Values may not sum to 100% due to rounding.

2017 PACE District Results by Subgroup (students are only counted in one (1) category)

PACE District Results by Race/Ethnicity	English Language Arts Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Mathematics Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	Science Percent at Level 3 & 4: Meets or Exceeds the Achievement Level	ELA Participation	Math Participation
	rea	mat	sci	rea	mat
Race - Hispanic	**	**	**	**	**
Race - American Indian or Alaskan Native (Non Hispanic)	**	**	**	**	**
Race - Asian (Non Hispanic)	**	**	**	**	**
Race - Black or African American (Non Hispanic)	**	**	**	**	**
Race - Native Hawaiian or Pacific Islander (Non Hispanic)	**	**	**	**	**
Race - White (Non Hispanic)	82%	55%	**	98%	98%
Race - Two or more races	**	**	**	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**	**	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	**	**	**	**	**
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	32%	32%	**	96%	96%
WaiverSubgroup - SWD and EconDis - Not EL	**	**	**	**	**
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**

WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All Students	82%	54%	**	97%	97%

Note: ** Count is below cell size of 10

2017 PACE District

Grade	Participation Rate	
	rea	mat
3	-	-
4	-	-
5	-	-
6	-	-
7	-	-
8	-	-
11	97%	97%
0	97%	97%

*Note: Grade 0 Represents District Total, but only includes accountability grades (i.e. grades 9 and 10 are not included)

Appendix B: Inter-Rater Reliability Analyses in 2015, 2016, and 2017

NH PACE: Inter-rater Reliability Analysis Report 2015

The purpose of analyzing the inter-rater reliability on the common PACE performance tasks is so that we may make judgements about the degree of consistency of a score given by any one scorer if it were to be scored by another scorer. Score consistency is desirable in the case of the PACE project to ensure that student scores can be fairly compared across classrooms and districts. One of the first steps in establishing this type of comparability is examining the degree of agreement on scores for teachers within schools. On top of that, score reliability is a necessary component within a validity argument. Rather than being able to calculate traditional reliability estimates such as coefficient alpha, because of the human judgement involved in the scoring process for the PACE performance tasks reliability must be examined through inter-rater reliability estimates. Just like coefficient alpha, inter-rater reliability analyses will estimate the proportion of “true score” variance within the observed score variance. To assess this kind of scoring consistency, all participating PACE districts were asked to have a sample of student work on the PACE performance tasks scored by two teachers independently, thereby producing double-scores for a sample of students.

After the data were cleaned, compiled and sorted, there were a total of 699 double-scores included in the inter-rater reliability analysis. The submitted double scores are broken down by grade, subject, and district in Table 1 below. Double scores have yet to be received from Epping.

Table 1
Number of Double Scores by Grade, Subject, and District

Grade	Frequency	Subject	Frequency	District	Frequency
3	24	ELA	170	Epping	0
4	48	Math	166	Rochester	167
5	48	Science	363	Sanborn	450
6	43	Total	699	Souhegan	82
7	44			Total	699
8	132				
9	110				
10	250				
Total	699				

For this report, inter-rater reliability is examined using three statistical indicators: percent agreement, Cohen’s Kappa, and intraclass correlations. Multiple indicators are used because each statistic provides unique information that is useful for making judgements about the degree of score reliability.

Percent Agreement

To calculate this first set of statistics, the analytic rubric scores were averaged across dimensions and rounded for each rater. Then, the percentage of cases where the average rounded rubric score is the same across raters was calculated to represent the “percent exact” match. The rounded rubric scores that were different only by one point fall into the “percent adjacent” category. This analysis reveals a very strong degree of agreement when all data is analyzed together, over 99% of all double scores fall into either the exact or adjacent categories. Of the 699 cases, 624 (89.3%) rounded average rubric scores were in exact agreement, and 72 (10.3%) of the scores were adjacent. That leaves just three scores (0.4%) where the scorers disagreed by more than one point. Table 2 shows these statistics disaggregated by district and content area.

Table 2
Exact Agreement & Adjacent

		%Exact	%Adjacent
Rochester	ELA	76.1	22.5
	Math	84.7	15.3
	Science	75.0	25.0
Sanborn	ELA	100.0	0.0
	Math	100.0	0.0
	Science	99.1	.9
Souhegan	ELA	58.6	34.5
	Math	46.7	53.3
	Science	56.5	43.5

As shown in Table 2, Sanborn has rather strong inter-rater reliability, but this is due to their consensus scoring approach.

Cohen’s Kappa

In addition to percent agreement, Cohen’s Kappa is another popular way for evaluating inter-rater reliability. The reason that Cohen’s Kappa is useful over and above the percent agreement measures shown above is twofold, first, the initial step of the analysis involves creating a cross-tabular presentation of the distribution of scores. Presenting the score distributions in this way illuminates the score points that may be more difficult than others to score consistently. Table 3 shows this distribution of double scores across the possible score points. As an example of how to interpret this table, the value of 10.2% in the upper left hand corner indicates that approximately 10% of the double-scored student work samples were rated an average of 1 by both raters. Moving across that same row, the table indicates that .7% of the double scored student work received a score of 1 by the first rater and a score 2 by the second rater.

Table 3
Distribution of Rubric Scores Across Raters

Score 1	Score 2			
	1	2	3	4
1	10.2%	.7%		
2	1.1%	26.6%	1.7%	.1%
3		3.9%	30.2%	1.0%
4		.3%	1.9%	22.3%

Table 3 confirms the information that was provided by the first analysis in that the majority of cases fall along the diagonals indicating good agreement. The percentages in the off-diagonal cells draw our attention to any scores that may be more difficult to score consistently. The cell highlighted in orange indicates that while scores are generally very consistent across the score scale, it is the difference between 2 and 3 that is more difficult to consistently distinguish. A second reason why Cohen’s Kappa is useful to calculate in addition to percent agreement is that it takes into account the possibility that two raters may arrive at the same score by chance alone. Cohen’s Kappa is calculated using the following formula:

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

where Pr(a) is observed agreement and Pr(e) is the probability of chance agreement. Across all districts, the Kappa estimate is .851, which according to Cohen’s rules of thumb, indicates almost perfect agreement. Table 4 shows the individual Kappa estimates by district and content area. Values can be interpreted in the following way: 0-.2 slight agreement, .21-.40 fair agreement, .41-.60 moderate agreement, .61-.80 substantial agreement, and 0.81-.1 represents almost perfect agreement.

Table 4
Cohen's Kappa by District and Subject

District		Kappa	Standard Error	Sig.
Rochester	ELA	.643	.076	.000
	Math	.789	.058	.000
	Science	.633	.133	.000
Sanborn	ELA	1.000	0.000	.000
	Math	1.000	0.000	.000
	Science	.987	.008	.000
Souhegan	ELA	.382	.123	.001
	Math	.204	.132	.083
	Science	.251	.179	.127

Table 4 reveals that all of the inter-rater reliability estimates show at least substantial agreement except for Souhegan. The level of agreement demonstrated in math and science in Souhegan may be particularly problematic in that the Kappa estimate is not significantly different than zero. This lack of statistical significance however, is likely in part due to lack of power from the reduced sample size given that this district only participated in the PACE project at the high school level for the 2014-2015 academic year.

Intraclass Correlations (ICC)

The most powerful way to estimate inter-rater reliability with the double-scored rubric data is with an interclass correlation. Rather than assuming the data are nominal or categorical as the prior two analyses have, the interclass correlation coefficient calls for the mean rubric scores (not rounded) to estimate the proportion of true score variance within the observed mean rubric scores. There are six ways to estimate the intraclass correlation coefficient and the appropriateness of each variation in estimation is dependent on both the study design and the nature of the scores. In this case, ICC (1,1)¹ is calculated indicating that the design is one-way random (i.e., each subject is scored by a different set of randomly selected raters) and the reliability is calculated from a single measure—average rubric score—rather than a mean of measures from different raters. Since the degree of absolute agreement has already been evaluated in the first two sets of analyses, this measure was chosen as an indicator of consistency (i.e., do fluctuations in the two sets of score move together). Overall, the ICC(1,1) for all districts together is .949. This reliability coefficient is remarkably high and indicates that almost 95% of the variance in the average rubric scores can be classified as true score variance. In other words, the degree of error due to individual rater differences is small. Table 5, below, shows how this estimate changes by district and subject.

Table 5
Intraclass Correlation Coefficients

District	Subject	ICC
Rochester	ELA	.858
	Math	.967
	Science	.913
Sanborn	ELA	.990
	Math	.996
	Science	.998
Souhegan	ELA	.575
	Math	.676
	Science	.598

¹[Shrout, Patrick E., and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin 86.2 \(1979\): 420.](#)

NH PACE: Inter-rater Reliability Analysis Report 2016

Critically, the PACE pilot promotes innovation at the local school district level. Rich discussions among educators about what competency looks like for every grade level and content area occurs in the adoption of a competency-based instructional model and the development of aligned performance assessments. Defining the expectations for student performance in a competency-based education model requires that the educators have shared definitions about both the content standards and the required evidence for evaluating student competence relative to the content standards. Therefore, baked within this model is within-district comparability in expectations for student performance.

The PACE pilot requires and audits the degree of consistency in educator scores of student work through the use of common performance tasks in two main ways: 1) within district calibration sessions resulting in annotated anchor papers, and 2) within district estimates of inter-rater reliability. First, all PACE districts hold grade-level calibration sessions for the scoring of the common task. Teachers bring samples of their student work from the common performance task representing the range of achievement in their classrooms. Teachers work together to come to a common understanding about how to use the rubrics to score papers and identify prototypical examples of student work for each score point on each rubric dimension. The educators annotate each of the anchor papers documenting the groups' rationale for the given score-point decision. These annotated anchor papers are then distributed throughout the district to help improve within-district consistency in scoring. Second, we externally audit the consistency in scoring by asking each district to submit a sample of papers from each common performance task that have been double-blind scored by teachers. The collection of double scores is then analyzed using inter-rater reliability methods to estimate within-district scoring consistency.

All participating PACE districts were asked to have 18 student work samples on each of the PACE performance tasks scored by two teachers independently, thereby producing double-scores for a sample of students. After the data were cleaned, compiled and sorted, there were a total of 2,337 double-scores included in the inter-rater reliability analysis. The submitted double scores are broken down by grade, subject, and district in Table 6.

Table 6.
Number of Double Scores by Grade, Subject, and District

Grade	Frequency	Subject	Frequency	District	Frequency
3	176	ELA	935	Concord	460
4	369	Math	885	Epping	337
5	373	Science	517	Monroe	89
6	282	Total	2,337	Pittsfield	520
7	271			Rochester	449
8	136			Sanborn	286
9	330			Seacoast	116
10	400			Souhegan	80
Total	2,337			Total	2,337

Inter-rater reliability is examined using two statistical indicators: percent agreement and Cohen’s Kappa. Two indicators are used because each statistic provides unique information that is useful for making judgments about the degree of score reliability.

Percent Agreement

Below we report rater consistency in two ways. First, we report percent agreement by task and rubric dimension (Table 7). As per the March 1, 2016 PACE Progress Report to USED, the target set for rater consistency is a 60% exact agreement rate for each dimension on the PACE Common Tasks. Exact agreement rates that did not meet this target are highlighted in green below. To calculate rater consistency by task and rubric dimension, scores on each rubric dimension were compared across raters. Then, the percentage of cases where the dimension score is the same across raters by task was calculated using a weighted average of data from all districts to represent the “percent exact” match. The dimension scores that were different only by one point fall into the “percent adjacent” category. This analysis reveals a strong degree of agreement when all data is analyzed together—about 98% of all double scores fall into either the exact or adjacent categories. Only two tasks had a rubric dimension that did not meet the 60% exact agreement—grade 6 ELA rubric dimension 3 and high school algebra rubric dimension 2.

Table 7.

Percent Exact Agreement & Adjacent by Task and Rubric Dimension for All Districts

Task	N	Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5	
		%Exa ct	%Ad j								
ELA											
4	189	80.4	19.6	80.4	19.0	83.1	16.9	76.4	22.9		
5	192	79.2	20.9	78.6	20.8	77.6	21.9	78.1	21.4		
6	158	68.4	26.6	77.8	20.9	58.9	35.4	74.0	25.9		
7	143	69.9	27.3	78.3	20.3	73.4	25.2	74.8	23.8		
9	123	72.4	26.0	72.4	26.8	77.2	21.1	76.4	22.8		
10	130	68.5	26.9	69.2	28.5	73.1	26.1	70.0	28.4		
Math											
3	176	83.0	15.9	83.5	16.0	84.7	15.4				
5	181	88.4	11.6	85.1	14.9	88.9	4.4				
6	124	69.4	27.4	66.9	27.4						
7	128	82.8	14.8	83.6	15.6	85.2	13.3				
Alg	143	65.7	32.2	58.0	33.6						
Geo	133	63.9	36.1	63.9	33.0	72.9	26.3				
Science											
4	180	71.7	27.7	75.0	25.0	73.3	26.1	75.6	23.9	74.9	24.1
8	136	80.1	19.9	75.0	25.0	72.8	25.7	71.3	25.7	69.4	27.4
Life	137	84.7	13.1	81.8	12.4	81.0	14.6	85.4	11.7	83.2	15.3
Phys	64	87.5	9.4	78.1	20.3	85.9	14.1	87.5	9.4	87.5	12.5

Second, we report rater consistency by district and subject area (Table 8). To calculate rater consistency by district and subject area, scores on each rubric dimension were compared across raters for each task. An average of the percent exact and percent adjacent for each task by district was calculated and then combined by subject area using a weighted average. This analysis reveals a strong degree of agreement for each district by subject area. However, Souhegan appears to have systematically lower rates of agreement in each subject area.

Table 8.
*Percent Exact Agreement & Adjacent
 by District and Subject Area*

District	Subject	%Exact	%Adj
Concord	ELA	78.59	20.23
	Math	76.37	20.91
	Science	75.00	24.50
Epping	ELA	66.50	28.75
	Math	64.42	32.41
	Science	84.30	15.45
Monroe	ELA	65.85	29.89
	Math	83.32	16.68
	Science	61.43	34.27
Pittsfield	ELA	72.72	26.92
	Math	70.89	28.79
	Science	79.98	19.58
Rochester	ELA	84.05	15.86
	Math	88.91	10.89
	Science	80.40	18.61
Sanborn	ELA	80.49	19.05
	Math	77.58	20.91
	Science	78.20	17.33
Seacoast	ELA	73.49	25.82
	Math	82.48	16.77
	Science	79.18	18.75
Souhegan	ELA	46.79	45.95
	Math	38.21	39.58
	Science	52.80	37.20

Cohen's Kappa

In addition to percent agreement, Cohen's Kappa is another way to evaluate inter-rater reliability. The reason that Cohen's Kappa is useful over and above the percent agreement measures is because it takes into account the possibility that two raters may arrive at the same score by chance alone. Cohen's Kappa is calculated using the following formula:

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

where Pr(a) is observed agreement and Pr(e) is the probability of chance agreement. Table 3.1 shows the individual Kappa estimates by task and rubric dimension for each subject calculated from a weighted average of Kappa estimates across districts. Values can be interpreted in the following way: 0-.2 slight agreement, .21-.40 fair agreement, .41-.60 moderate agreement, .61-.80 substantial agreement, and 0.81-.1 represents almost perfect agreement. Across all districts,

the Kappa estimates in ELA, math and science are between .41 and .85, which according to Cohen’s rules of thumb, indicates moderate to substantial agreement.

Table 9.

Cohen’s Kappa by Task and Rubric Dimension for All Districts

Task	Rubric Dimension 1			Rubric Dimension 2			Rubric Dimension 3			Rubric Dimension 4			Rubric Dimension 5		
	K	SE	Sig.												
<i>ELA</i>															
4	0.718	0.042	0.000	0.717	0.042	0.000	0.748	0.041	0.000	0.651	0.052	0.000			
5	0.683	0.045	0.000	0.683	0.044	0.000	0.679	0.044	0.000	0.670	0.045	0.000			
6	0.541	0.053	0.000	0.676	0.048	0.000	0.408	0.056	0.000						
7	0.584	0.052	0.000	0.689	0.049	0.000	0.639	0.050	0.000	0.652	0.051	0.000			
9	0.617	0.055	0.000	0.618	0.057	0.000	0.682	0.053	0.000	0.669	0.053	0.000			
10	0.573	0.056	0.000	0.571	0.057	0.000	0.622	0.054	0.000	0.583	0.056	0.000			
<i>Math</i>															
3	0.746	0.042	0.000	0.754	0.042	0.000	0.722	0.046	0.000						
5	0.834	0.034	0.000	0.799	0.035	0.000	0.851	0.031	0.000	0.721	0.041	0.000			
6	0.572	0.058	0.000	0.504	0.058	0.000				0.612	0.053	0.000			
7	0.770	0.044	0.000	0.783	0.043	0.000	0.786	0.045	0.000						
Alg	0.534	0.054	0.000	0.444	0.054	0.000									
Geo	0.475	0.062	0.000	0.453	0.062	0.000	0.628	0.053	0.000						
<i>Science</i>															
4	0.598	0.048	0.000	0.637	0.048	0.000	0.616	0.048	0.000	0.648	0.046	0.000	0.655	0.044	0.000
8	0.704	0.051	0.000	0.630	0.056	0.000	0.621	0.052	0.000	0.602	0.054	0.000	0.578	0.059	0.000
Life	0.803	0.040	0.000	0.765	0.043	0.000	0.750	0.045	0.000	0.812	0.039	0.000	0.785	0.041	0.000
Phys	0.834	0.054	0.000	0.697	0.071	0.000	0.789	0.064	0.000	0.826	0.057	0.000	0.830	0.056	0.000

Table 10 shows the individual Kappa estimates by rubric dimension and subject area for each district. The Kappa estimates for each subject area are a weighted average of Kappa estimates across tasks in that subject area. Any Kappa estimate lower than moderate agreement is highlighted in green.

Table 10.

Cohen's Kappa by District, Subject Area, and Rubric Dimension

Distr	Subj	Rubric Dimension 1			Rubric Dimension 2			Rubric Dimension 3			Rubric Dimension 4			Rubric Dimension 5		
		<i>K</i>	SE	Sig.												
CON	ELA	.678	.044	.000	.722	.043	.000	.598	.049	.000	.665	.046	.000			
	Math	.736	.040	.000	.607	.045	.000	.745	.050	.000	.549	.109	.000			
	SCI	.616	.069	.000	.570	.072	.000	.714	.064	.000	.694	.067	.000	.617	.069	.000
EPP	ELA	.539	.058	.000	.561	.057	.000	.529	.059	.000	.546	.060	.000			
	Math	.567	.055	.000	.431	.058	.000	.667	.064	.000	.355	.167	.014			
	SCI	.778	.056	.000	.801	.056	.000	.740	.063	.000	.795	.054	.000	.796	.054	.000
MON	ELA	.643	.102	.000	.590	.106	.000	.364	.110	.000	.323	.105	.001			
	Math	.766	.094	.000	.440	.180	.001	.616	.151	.000	.279	.192	.161			
	SCI	.421	.154	.014	.468	.266	.025	.197	.213	.291	.478	.175	.004	.197	.195	.268
PIT	ELA	.543	.044	.000	.575	.046	.000	.633	.042	.000	.672	.044	.000			
	Math	.515	.053	.000	.631	.048	.000	.751	.049	.000	.491	.139	.000			
	SCI	.740	.047	.000	.718	.047	.000	.726	.047	.000	.793	.041	.000	.698	.048	.000
ROC	ELA	.780	.039	.000	.745	.042	.000	.791	.038	.000	.778	.039	.000			
	Math	.874	.030	.000	.816	.035	.000	.864	.034	.000	.910	.050	.000			
	SCI	.819	.047	.000	.737	.057	.000	.653	.060	.000	.630	.061	.000	.784	.049	.000
SAN	ELA	.675	.056	.000	.771	.049	.000	.786	.048	.000	.695	.056	.000			
	Math	.625	.058	.000	.728	.052	.000	.788	.062	.000	1.00	.000	.000			
	SCI	.756	.064	.000	.688	.067	.000	.675	.071	.000	.720	.067	.000	.702	.069	.000
SEA	ELA	.650	.099	.000	.664	.103	.000	.523	.101	.000	.631	.100	.000			
	Math	.740	.066	.000	.840	.053	.000	.770	.075	.000	.838	.088	.000			
	SCI	.478	.232	.008	.840	.153	.000	.870	.117	.000	.355	.229	.035			
SOU	ELA	.154	.120	.144	.518	.114	.000	.217	.121	.036	.221	.112	.023			
	Math	.109	.127	.382	.187	.121	.084	.242	.212	.232						
	SCI	.348	.124	.002	.241	.125	.034	.303	.135	.009	.451	.134	.000	.402	.145	.001

This analysis reveals that all of the inter-rater reliability estimates show at least moderate agreement (and for many, substantial agreement) on all rubric dimensions except for a few districts. The level of agreement demonstrated in Souhegan and Monroe may be problematic in that the Kappa estimate is not significantly different than zero. The statistical non-significance, however, is likely in part due to lack of power from the reduced sample size given that Souhegan only participated at the high school level and the Monroe district is very small and unable to submit the requested number of student work samples.

The results of both analyses provide support for the degree of inter-rater consistency in the scoring of the common performance tasks. This evidence suggests that teachers within districts are able to successfully conduct calibration sessions and comparably evaluate student work. Both analyses point to a potential problem with the consistency of scoring in the

Souhegan school district. The Center for Assessment is working closely with Souhegan High School to better understand the possible sources for reduced inter-rater reliability in this district, and to find ways to improve the scoring practices.

NH PACE: Inter-rater Reliability Analysis Report 2017

Critically, the PACE pilot promotes innovation at the local school district level. Rich discussions among educators about what competency looks like for every grade level and content area occurs in the adoption of a competency-based instructional model and the development of aligned performance assessments. Defining the expectations for student performance in a competency-based education model requires that the educators have shared definitions about both the content standards and the required evidence for evaluating student competence relative to the content standards. Therefore, integrated within this model is within-district comparability in expectations for student performance.

The PACE pilot requires and audits the degree of consistency in educator scores of student work through the use of common performance tasks in two main ways: 1) within district calibration sessions resulting in annotated anchor papers, and 2) within district estimates of inter-rater reliability. First, all PACE districts hold grade-level calibration sessions for the scoring of the common task. Teachers bring samples of their student work from the common performance task representing the range of achievement in their classrooms. Teachers work together to come to a common understanding about how to use the rubrics to score papers and identify prototypical examples of student work for each score point on each rubric dimension. The educators annotate each of the anchor papers documenting the groups' rationale for the given score-point decision. These annotated anchor papers are then distributed throughout the district to help improve within-district consistency in scoring. Second, we externally audit the consistency in scoring by asking each district to submit a sample of double-blind scores from each common performance task. The collection of double scores is then analyzed using inter-rater reliability methods to estimate within-district scoring consistency.

All participating PACE districts were asked to have 18 student work samples on each of the PACE performance tasks scored by two teachers independently, thereby producing double-scores for a sample of students. After the data were cleaned, compiled and sorted, there were a total of 2,543 double-scores included in the inter-rater reliability analysis. The submitted double scores are broken down by grade, subject, and district in Table 6 below. Chemistry was not a required PACE Common Task in 2016-17 so only two districts submitted student work samples.

Table 6. *Number of Double Scores by Grade, Subject, and District*

Grade	Frequency	Subject	Frequency	District	Frequency
Grade 3	188	ELA	974	Concord	350
Grade 4	379	Math	927	Epping	318
Grade 5	399	Science	642	Monroe	86
Grade 6	330	Total	2543	Pittsfield	504
Grade 7	309			Rochester	462
Grade 8	156			Sanborn	292
Grade 9	143			SAU 35	190
Grade 10	104			Seacoast	221
Algebra	111			Souhegan	120
Geometry	108			Total	2543
Life	157				
Physical	121				
Chemistry	38				
Total	2543				

For this report, inter-rater reliability is examined using two statistical indicators: percent agreement and Cohen’s Kappa. Two indicators are used because each statistic provides unique information that is useful for making judgments about the degree of score reliability.

Percent Agreement

Below we report rater consistency in two ways. First, we report percent agreement by task and rubric dimension (Table 7). As per the March 1, 2016 PACE Progress Report to the USDOE, the target set for rater consistency is a 60% exact agreement rate for each dimension on the PACE Common Tasks. Exact agreement rates that did not meet this target are highlighted in green below. To calculate rater consistency by task and rubric dimension, scores on each rubric dimension were compared across raters. Then, the percentage of cases where the dimension score is the same across raters by task was calculated using a weighted average of data from all districts to represent the “percent exact” match. The dimension scores that differed by one-point fall into the “percent adjacent” category. This analysis reveals a strong degree of agreement when all data is analyzed together—about 98% of all double scores fall into either the exact or adjacent categories. Only two tasks had a rubric dimension that did not meet the 60% exact agreement—high school geometry rubric dimensions 5-6 and high school physical science rubric dimension 2—but they only fell short by a very minor amount.

Table 7. Percent Exact Agreement & Adjacent by Task and Rubric Dimension for All Districts

Task	Rubric Dimension 1		Rubric Dimension 2		Rubric Dimension 3		Rubric Dimension 4		Rubric Dimension 5		Rubric Dimension 6	
	Exact	Adj										
<i>ELA</i>												
4	79.9%	20.1%	69.9%	30.1%	79.8%	20.2%	72.0%	14.9%				
5	72.3%	27.2%	74.6%	24.9%	82.1%	17.4%	80.1%	12.7%				
6	70.5%	29.5%	71.6%	27.1%	75.6%	23.7%	74.2%	13.7%				
7	79.4%	20.6%	75.0%	25.0%	73.8%	25.0%	74.7%	17.8%				
9	73.4%	25.2%	74.1%	25.9%	75.5%	23.1%	77.6%	14.5%				
10	69.2%	29.8%	68.3%	28.8%	76.0%	22.1%	69.2%	15.0%				
<i>Math</i>												
3	82.7%	16.8%	88.3%	11.7%	80.1%	18.8%	81.9%	12.0%	85.3%	11.9%		
5	84.7%	14.7%	86.3%	13.2%	83.0%	15.5%	76.6%	14.5%	75.9%	16.2%		
6	79.3%	20.1%	73.1%	25.7%	75.9%	22.4%						
7	81.8%	17.6%	74.3%	25.7%	73.3%	25.3%						
Alg	65.8%	31.5%	76.0%	19.3%	68.5%	26.9%	72.4%	14.2%	67.9%	14.5%	67.3%	31.6%
Geo	61.5%	35.6%	72.8%	24.3%	70.4%	27.8%	69.7%	15.8%	58.8%	21.1%	59.8%	39.2%
<i>Science</i>												
4	77.1%	22.9%	79.9%	20.1%	77.7%	21.1%	81.3%	11.0%	76.7%	15.2%		
8	70.5%	28.8%	70.5%	27.6%	73.7%	26.3%	75.3%	15.0%	70.1%	16.8%		
Life	72.9%	25.2%	73.7%	25.0%	71.2%	24.8%	69.3%	17.9%	75.9%	14.4%		
Phys	73.1%	26.1%	55.1%	40.7%	72.3%	26.9%	67.8%	23.6%	63.8%	29.2%		
Chem	84.2%	15.8%	78.9%	15.8%	81.1%	18.9%	83.3%	6.0%	73.0%	11.6%		

Second, we report rater consistency by district and subject area (Table 8). To calculate rater consistency by district and subject area, scores on each rubric dimension were compared across raters for each task. An average of the percent exact and percent adjacent for each task by district was calculated and then combined by subject area using a weighted average. This analysis reveals a strong degree of agreement for each district by subject area, although Monroe and Souhegan appear to have systematically lower rates of agreement in each subject area along with the lowest number of student work samples when all three subjects are taken together.

Table 8. Percent Exact Agreement & Adjacent by District and Subject Area

District	Subject	N	Exact	Adj
Concord	ELA	135	68.5%	29.4%
	Math	120	82.4%	16.2%
	Science	89	74.1%	23.2%
Epping	ELA	118	79.2%	20.8%
	Math	120	73.5%	26.3%
	Science	80	72.3%	27.8%
Monroe	ELA	39	51.3%	45.5%
	Math	36	50.2%	44.2%
	Science	11	47.3%	49.1%
Pittsfield	ELA	201	72.4%	27.5%
	Math	138	73.9%	25.4%
	Science	165	76.2%	23.6%
Rochester	ELA	174	81.2%	18.8%
	Math	173	86.8%	12.7%
	Science	106	85.6%	14.2%
Sanborn	ELA	107	83.9%	15.9%
	Math	109	74.1%	23.5%
	Science	73	68.9%	28.4%
SAU 35	ELA	69	65.2%	33.0%
	Math	68	75.8%	23.5%
	Science	49	65.5%	32.8%
Seacoast	ELA	91	89.0%	11.0%
	Math	108	80.9%	18.9%
	Science	22	76.4%	21.8%
Souhegan	ELA	40	52.5%	44.4%
	Math	40	50.0%	41.7%
	Science	39	51.3%	40.0%

Cohen's Kappa

In addition to percent agreement, Cohen's Kappa is another way to evaluate inter-rater reliability. The reason that Cohen's Kappa is useful over and above the percent agreement measures is because it takes into account the possibility that two raters may arrive at the same score by chance alone. Cohen's Kappa is calculated using the following formula:

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

where Pr(a) is observed agreement and Pr(e) is the probability of chance agreement. Table 9 shows the individual Kappa estimates by task and rubric dimension for each subject calculated from a weighted average of Kappa estimates across districts. Values can be interpreted in the following way: 0-.2 slight agreement, .21-.40 fair agreement, .41-.60 moderate agreement, .61-.80 substantial agreement, and 0.81-.1 represents almost perfect agreement. Across all districts, the Kappa estimates in ELA, math and science are between .41 and .84 with one exception (high school physical science rubric dimension 2). According to Cohen's rules of thumb, these results indicate moderate to substantial agreement.

Table 9. *Cohen's Kappa by Task and Rubric Dimension for All Districts*

	Rubric Dimension 1			Rubric Dimension 2			Rubric Dimension 3			Rubric Dimension 4			Rubric Dimension 5			Rubric Dimension 6		
	<i>K</i>	SE	Sig.															
<i>ELA</i>																		
4	0.71	0.04	0.00	0.56	0.05	0.00	0.71	0.04	0.00	0.57	0.05	0.00						
5	0.61	0.04	0.00	0.64	0.04	0.00	0.74	0.04	0.00	0.70	0.04	0.00						
6	0.58	0.05	0.00	0.59	0.05	0.00	0.64	0.05	0.00	0.62	0.05	0.00						
7	0.70	0.05	0.00	0.63	0.05	0.00	0.62	0.05	0.00	0.62	0.05	0.00						
9	0.63	0.05	0.00	0.62	0.05	0.00	0.65	0.05	0.00	0.66	0.05	0.00						
10	0.57	0.06	0.00	0.56	0.07	0.00	0.65	0.06	0.00	0.55	0.07	0.00						
<i>Math</i>																		
3	0.76	0.04	0.00	0.84	0.03	0.00	0.71	0.04	0.00	0.75	0.04	0.00	0.79	0.04	0.00			
5	0.76	0.04	0.00	0.79	0.04	0.00	0.77	0.04	0.00	0.68	0.04	0.00	0.68	0.04	0.00			
6	0.71	0.04	0.00	0.64	0.05	0.00	0.69	0.04	0.00									
7	0.74	0.05	0.00	0.64	0.05	0.00	0.63	0.05	0.00									
Alg	0.48	0.06	0.00	0.64	0.06	0.00	0.56	0.06	0.00	0.62	0.06	0.00	0.54	0.07	0.00	0.53	0.07	0.00
Geo	0.48	0.06	0.00	0.58	0.06	0.00	0.60	0.06	0.00	0.57	0.07	0.00	0.45	0.07	0.00	0.45	0.07	0.00
<i>Science</i>																		
4	0.66	0.05	0.00	0.69	0.05	0.00	0.67	0.05	0.00	0.73	0.04	0.00	0.67	0.05	0.00			
8	0.58	0.05	0.00	0.58	0.05	0.00	0.63	0.05	0.00	0.66	0.05	0.00	0.60	0.05	0.00			
Life	0.63	0.05	0.00	0.63	0.05	0.00	0.61	0.05	0.00	0.58	0.05	0.00	0.69	0.05	0.00			
Phys	0.64	0.06	0.00	0.36	0.06	0.00	0.63	0.05	0.00	0.57	0.06	0.00	0.53	0.06	0.00			
Chm	0.75	0.09	0.00	0.67	0.10	0.00	0.74	0.09	0.00	0.75	0.09	0.00	0.60	0.11	0.00			

Table 10 below shows the individual Kappa estimates by subject area and rubric dimension for each district (see Appendix B for Cohen’s Kappa by Task, District, and Rubric Dimension). The Kappa estimates for each subject area and rubric dimension are a weighted average of Kappa estimates across tasks in that subject area. Any Kappa estimate lower than moderate agreement (0.41) is highlighted in green.

Table 10. *Cohen’s Kappa by District, Subject Area, and Rubric Dimension*

	Rubric Dimension 1			Rubric Dimension 2			Rubric Dimension 3			Rubric Dimension 4			Rubric Dimension 5			Rubric Dimension 6		
	K	SE	Sig.															
CON																		
ELA	0.53	0.06	0.00	0.57	0.06	0.00	0.50	0.06	0.00	0.52	0.06	0.00						
Math	0.82	0.04	0.00	0.73	0.05	0.00	0.78	0.05	0.00	0.72	0.07	0.00	0.76	0.06	0.00	0.72	0.10	0.00
Sci	0.48	0.08	0.00	0.70	0.06	0.00	0.69	0.06	0.00	0.58	0.07	0.00	0.61	0.07	0.00			
EPP																		
ELA	0.60	0.06	0.00	0.67	0.06	0.00	0.75	0.05	0.00	0.79	0.05	0.00						
Math	0.75	0.05	0.00	0.74	0.05	0.00	0.64	0.06	0.00	0.69	0.06	0.00	0.55	0.07	0.00	0.40	0.11	0.00
Sci	0.55	0.08	0.00	0.56	0.08	0.00	0.65	0.07	0.00	0.59	0.07	0.00	0.67	0.07	0.00			
MON																		
ELA	0.35	0.12	0.00	0.21	0.12	0.03	0.28	0.11	0.01	0.30	0.12	0.00						
Math	0.55	0.11	0.00	0.29	0.11	0.00	0.24	0.10	0.01	0.32	0.17	0.02	0.10	0.20	0.58			
Sci	0.19	0.29	0.37	0.00	0.05	1.00	-0.12	0.25	0.63	-0.05	0.23	0.80	0.36	0.19	0.04			
PIT																		
ELA	0.61	0.05	0.00	0.51	0.05	0.00	0.67	0.04	0.00	0.58	0.05	0.00						
Math	0.63	0.05	0.00	0.65	0.05	0.00	0.58	0.05	0.00	0.52	0.08	0.00	0.74	0.06	0.00			
Sci	0.66	0.05	0.00	0.61	0.05	0.00	0.73	0.04	0.00	0.78	0.04	0.00	0.64	0.04	0.00			

<i>ROC</i>																		
ELA	0.75	0.04	0.00	0.66	0.05	0.00	0.80	0.04	0.00	0.71	0.05	0.00	0.75	0.05	0.00	0.51	0.11	0.00
Math	0.81	0.04	0.00	0.85	0.03	0.00	0.87	0.03	0.00	0.83	0.04	0.00	0.74	0.05	0.00			
Sci	0.89	0.04	0.00	0.81	0.05	0.00	0.78	0.05	0.00	0.79	0.05	0.00						
<i>SAN</i>																		
ELA	0.77	0.05	0.00	0.78	0.05	0.00	0.79	0.05	0.00	0.76	0.05	0.00						
Math	0.69	0.06	0.00	0.64	0.06	0.00	0.69	0.06	0.00	0.71	0.07	0.00	0.68	0.07	0.00	0.38	0.12	0.00
Sci	0.60	0.08	0.00	0.55	0.08	0.00	0.50	0.08	0.00	0.53	0.08	0.00	0.55	0.08	0.00			
<i>SAU35</i>																		
ELA	0.54	0.08	0.00	0.38	0.09	0.00	0.45	0.09	0.00	0.46	0.09	0.00						
Math	0.61	0.08	0.00	0.75	0.07	0.00	0.60	0.08	0.00	0.65	0.09	0.00	0.59	0.09	0.00	0.77	0.12	0.00
Sci	0.54	0.10	0.00	0.22	0.11	0.02	0.40	0.10	0.00	0.62	0.09	0.00	0.75	0.07	0.00			
<i>SEA</i>																		
ELA	0.88	0.04	0.00	0.87	0.04	0.00	0.81	0.05	0.00	0.76	0.06	0.00						
Math	0.68	0.06	0.00	0.84	0.04	0.00	0.71	0.05	0.00	0.70	0.08	0.00	0.77	0.07	0.00			
Sci	0.67	0.15	0.00	0.78	0.12	0.00	0.65	0.14	0.00	0.78	0.11	0.00	0.26	0.15	0.03			
<i>SOU</i>																		
ELA	0.29	0.11	0.00	0.28	0.11	0.00	0.42	0.12	0.00	0.17	0.11	0.07						
Math	0.06	0.09	0.52	0.40	0.11	0.00	0.32	0.11	0.00	0.44	0.10	0.00	0.32	0.09	0.00	0.31	0.11	0.00
Sci	0.47	0.11	0.00	0.17	0.11	0.08	0.45	0.10	0.00	0.20	0.10	0.03	0.36	0.10	0.00			

Table 10 Cont'd

This analysis reveals that most of the inter-rater reliability estimates show at least moderate agreement (and many substantial agreement) on all rubric dimensions except for two districts. The level of agreement demonstrated in Souhegan and Monroe is systematically lower in all three subject areas. In some cases, the Kappa estimate is not significantly different than zero. This lack of statistical significance, however, is likely due in part to lack of power from the reduced sample size. Souhegan only participated at the high school level and Monroe only has a small number of students per grade and subject area so they are not able to submit the requested number of student work samples. That said, further conversations about strengthening inter-rater reliability in Souhegan and Monroe are warranted based on these analyses as are conversations with the districts where a couple teachers appeared to have consensus scored instead of individually scored.

The results of both analyses provide overwhelming support for the degree of inter-rater consistency in the scoring of the common performance tasks. This evidence suggests that teachers within districts are able to successfully conduct calibration sessions and comparably evaluate student work. Both analyses point to a potential problem with the consistency of scoring in the Souhegan and Monroe school districts. The Center for Assessment is working closely with these districts to better understand the possible sources for reduced inter-rater reliability in these districts, and to find ways to improve the scoring practices. For example, the Center for Assessment provided a follow-up inter-rater reliability analysis report to Souhegan High School and discussed inter-rater reliability results with school administrators and instructional coaches alongside suggestions for improving calibration processes within the school.

Appendix C: Generalizability Analysis in 2016 and 2017

Generalizability Analysis 2016

In New Hampshire's PACE innovative assessment and accountability system there could be upwards of seventy local assessments contributing to students' overall achievement estimates. One of the technical challenges of estimating student achievement based on a limited set of classroom assessment evidence is the generalizability of such estimates. For example, would students likely demonstrate similar levels of achievement had they been given a different set of assessment tasks? And how many classroom assessments are needed to provide a stable measure of student achievement? These questions can be evaluated using generalizability theory.

In generalizability theory, a distinction is made between generalizability (G) studies and decision (D) studies. The purpose of a G-study is to provide as much information as possible about the sources of variation in the measurement due to persons and tasks, for example; whereas, a D-study uses the information provided by a G-study to design the best possible application of the measurement for a particular purpose. The purpose of this analysis is to (1) examine the reliability of generalization from a collection of classroom assessments intended to measure student achievement to the universe of all possible assessments and (2) determine an efficient number of classroom assessments necessary to ensure high reliability of estimates of student achievement made in the NH PACE pilot project.

Using electronic grade book data provided by one of the eight districts implementing NH's PACE pilot in 2015-2016, we examined the generalizability of the individual scores that go into achievement estimates (e.g., summative tests, quizzes, projects, performance tasks) in six subject/grade combinations: English language arts (grade 5 & 7), math (grade 3 & 6), and science (grade 4 & 8)—see Table 28 for the number of students and assessment tasks.

Table 28.

Number of persons and tasks by subject and grade

Subject	Grade	Persons	Tasks
ELA	5	18	72
	7	74	20
Math	3	22	69
	6	54	21
Science	4	12	6
	8	77	12

The variance of assessment (task) scores can be partitioned into independent sources of variation due to differences between persons, tasks, and the residual. This is called a one-facet crossed design. In this analysis, both persons and tasks are regarded as random samples from the universe of tasks and population of persons that could have been included. As a result, a random effects ANOVA can be used to estimate the four sources of variability in competency score data: systematic differences among persons (p), systematic differences among tasks (t), person-by-task interaction (p x t), and random error. Random error is confounded with the p x t interaction. Variance component estimates and generalizability coefficients were calculated for both relative

decisions (rank ordering) and absolute decisions (level of performance) because the generalizability of a measure depends on how the data will be used.

Table 29 shows the estimated variance components and percent of total variance, both of which reflect the magnitude of error in generalizing from a student's score on a single assessment task to his or her universe score. For example, in all grade/subject combinations, one assessment (task) does not account for a large percent of the variance in individual student achievement (only 8-15%). The largest variance component in all grade/subject combinations is the residual (between 38-73%). Large residual variance suggests a few things: (1) a large $p \times t$ interaction; (2) sources of error variability in the competency score measurement that the one-facet $p \times t$ design has not captured, or (3) both. A large variance component for the $p \times t$ interaction indicates that the relative standing (or rank order) of students differs from assessment to assessment, which is not surprising. We would expect that not all people would find the same tasks easy or difficult.

Table 29.

Variance component estimates for the person x task G study by subject and grade

Grade/ Subject	Source of Variance	Df	Sum of Squares	Mean Squares	Variance Components	% of Total Variance
5ELA	<i>p</i>	17	185.905	10.936	0.147	24.31%
	<i>t</i>	71	139.897	1.970	0.089	14.74%
	<i>p x t</i>	1156	424.941	0.368	0.368	60.95%
7ELA	<i>p</i>	72	398.349	5.533	0.257	35.77%
	<i>t</i>	19	99.136	5.218	0.065	9.08%
	<i>p x t</i>	1320	522.290	0.396	0.396	55.15%
3MATH	<i>p</i>	21	119.697	5.700	0.080	24.12%
	<i>t</i>	68	68.521	1.008	0.036	10.95%
	<i>p x t</i>	1256	268.825	0.214	0.214	64.93%
6MATH	<i>p</i>	53	589.409	11.121	0.512	52.73%
	<i>t</i>	20	94.646	4.732	0.081	8.31%
	<i>p x t</i>	1042	394.006	0.378	0.378	38.96%
4SCI	<i>p</i>	11	3.935	0.358	0.035	16.77%
	<i>t</i>	5	1.970	0.394	0.020	9.84%
	<i>p x t</i>	52	7.863	0.151	0.151	73.39%
8SCI	<i>p</i>	76	471.644	6.206	0.485	47.72%
	<i>t</i>	11	123.715	11.247	0.141	13.88%
	<i>p x t</i>	808	315.140	0.390	0.390	38.40%

Note. VAR COMPS procedure in SPSS was used to estimate sum of squares and mean squares.

Generalizability theory also provides a reliability coefficient called a generalizability (G) coefficient. This G coefficient shows how accurate the generalization is from a student's observed score, based on a sample of the student's work, to his or her universe score. Applied to this analysis, the G coefficient represents the proportion of variability in observed assessment scores attributable to systematic differences in students' competency. Table 30 provides the variance component estimates and generalizability coefficients for both relative decisions (rank ordering) and absolute decisions (level of performance) because in G theory how generalizable a measure is depends on how the data will be used in the D study. For example, relative decisions use the data to rank order students (or schools), whereas absolute decisions use the data to determine student proficiency in a given content domain.

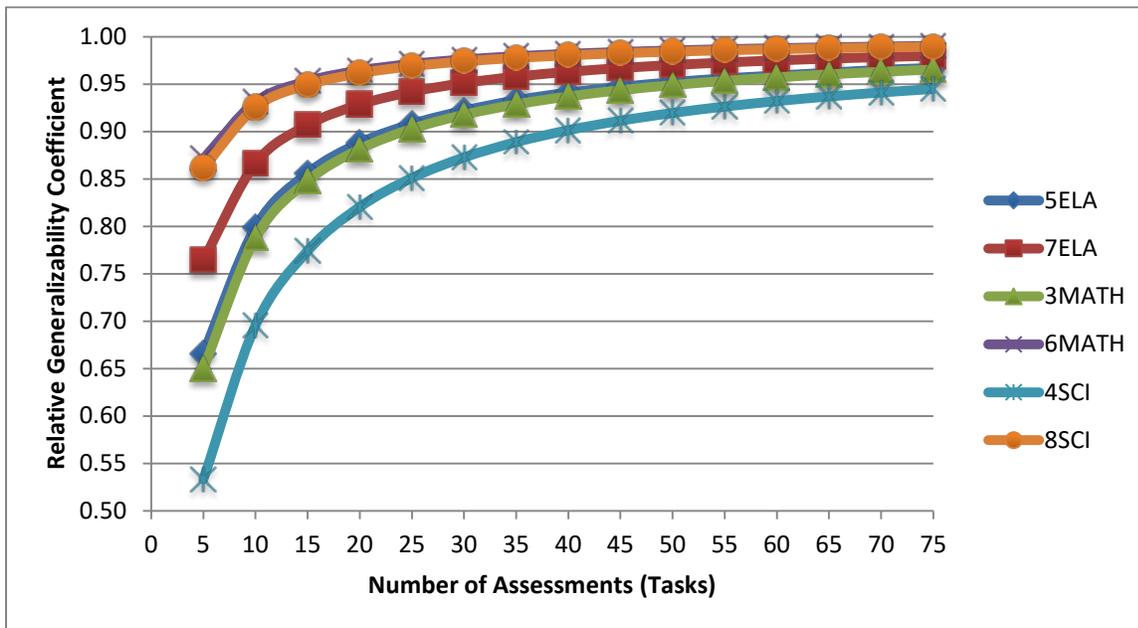
Other than grade 4 science, where only 6 assessments were used to calculate a student's overall district-level competency scores—there are high G coefficients for both absolute and relative decisions. This means that the collection of classroom assessments provide for stable estimates of student achievement in a given content domain.

Table 30.

Variance component estimates and generalizability coefficients for relative and absolute error D study by subject and grade

Grade/ Subject	Relative error variance	Absolute error variance	Relative error generalizability coefficient $E\rho^2$	Absolute error generalizability coefficient ϕ
5ELA	0.005	0.006	0.966	0.958
7ELA	0.019	0.023	0.928	0.917
3MATH	0.003	0.003	0.962	0.956
6MATH	0.018	0.021	0.966	0.959
4SCI	0.025	0.028	0.578	0.547
8SCI	0.032	0.044	0.937	0.916

In the D study, we show how increasing the number of assessments included in achievement estimates results in diminishing returns beyond approximately 20 assessments. Figure 20 shows sample plots showing estimated relative and absolute error generalizability coefficients as a function of the number of assessments by grade and subject.



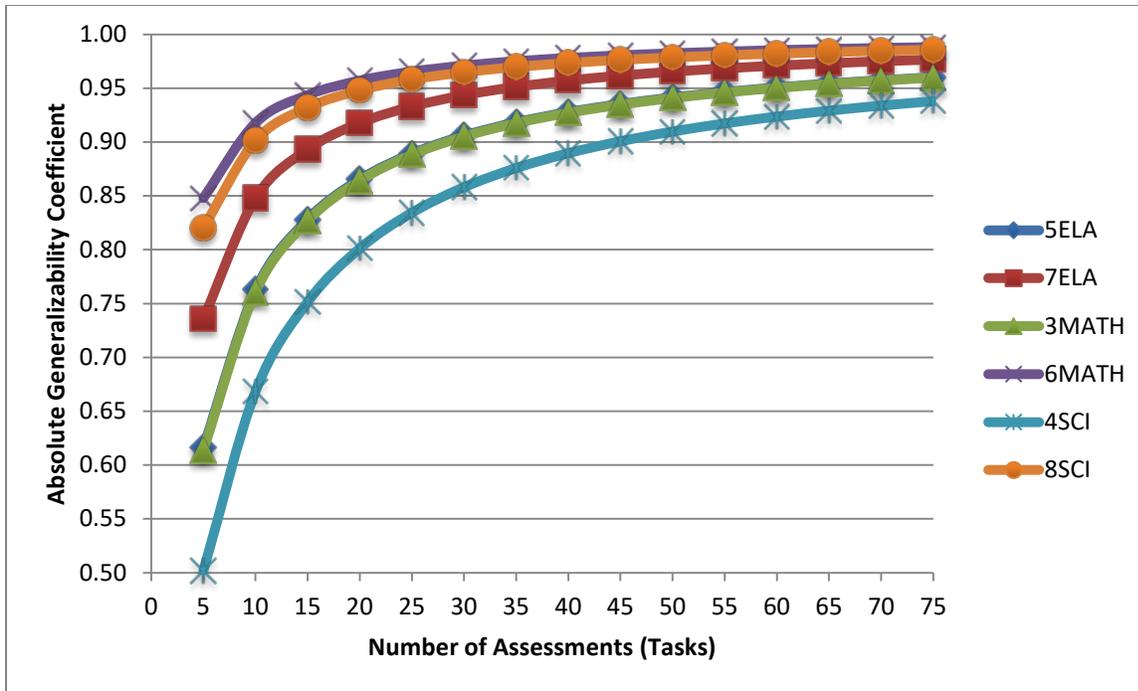


Figure 20. Sample plots showing estimated $E\rho^2$ scores (top) and ϕ coefficient (bottom) as a function of the number of assessments by grade and subject

Averaging across the grades and subjects by the number of assessments (tasks), there is a high degree of relative and absolute stability estimates (around 0.90) of student achievement between 15-20 classroom assessments—see Table 31.

Table 31.

Average estimated $E\rho^2$ scores (relative generalizability coefficient) and ϕ coefficient (absolute generalizability coefficient) as a function of the number of assessments across subjects and grades

Number of Assessments (Tasks)	$\overline{E\rho^2}$	$\overline{\Phi}$
5	0.72	0.69
10	0.83	0.81
15	0.88	0.86
20	0.91	0.89
25	0.92	0.91
30	0.94	0.92
35	0.94	0.93
40	0.95	0.94
45	0.96	0.95
50	0.96	0.95
55	0.96	0.96
60	0.97	0.96
65	0.97	0.96
70	0.97	0.97
75	0.97	0.97

These results suggest that classroom assessments can provide for reliable estimates of student achievement for use in a school accountability context like the NH PACE pilot project. Approximately 15-20 assessments per year provide for an efficient trade-off while still ensuring a high degree of relative and absolute decision reliability and all of our participating districts employ approximately this number of assessments in each PACE grade and subject.

Generalizability Analysis 2017

There could be upwards of seventy local assessments contributing to students' overall achievement estimates in New Hampshire's PACE innovative assessment and accountability system. One of the technical challenges of estimating student achievement based on a limited set of classroom assessment evidence is the generalizability of such estimates. For example, would students likely demonstrate similar levels of achievement had they been given a different set of assessment tasks? And how many classroom assessments are needed to provide a stable measure of student achievement? These questions can be evaluated using generalizability theory.

In generalizability theory, a distinction is made between generalizability (G) studies and decision (D) studies. The purpose of a G-study is to provide as much information as possible about the sources of variation in the measurement due to persons and tasks, for example; whereas, a D-study uses the information provided by a G-study to design the best possible application of the measurement for a particular purpose. The purpose of this analysis is to (1) examine the generalizability of inferences from a collection of classroom assessments intended to measure student achievement to the universe of all possible assessments and (2) determine an efficient number of classroom assessments necessary to ensure high reliability of estimates of student achievement made in the NH PACE pilot project.

Results from the NH PACE 2016 Generalizability Analysis Report suggested that classroom assessments can provide for reliable estimates of student achievement for use in a school accountability context like the NH PACE pilot project. The 2016 analysis used electronic grade book data from one school district (N=257) and found that approximately 15-20 assessments per year provided for an efficient trade-off while still ensuring a high degree of relative and absolute decision reliability.

The 2017 Generalizability Analysis Report uses electronic grade book data from three of the nine districts implementing the PACE pilot in 2016-17 with a total of 3,348 students. Table 50 shows the number of students by grade, subject, and district included in the analyses. As before, we examined the generalizability of the individual scores that go into achievement estimates (e.g., summative tests, quizzes, projects, performance tasks).

Grade	N	Subject	N	District	N
3	800	ELA	1365	Concord	2403
4	1143	Math	1307	Rochester	250
5	844	Science	676	Sanborn	695
6	151	Total	3348	Total	3348
7	199				
8	211				
Total	3348				

Table 50. *Number of students in the analyses by grade, subject, and district (N=3348)*

Only a sample of grades and subject areas were requested. Table 51 shows the number of students (persons) and assessment tasks (tasks) for the grade and subject combinations where the generalizability analyses was performed by district.

District	Subject	Grade	Person	Tasks
Concord	ELA	3	338	19
		4	339	19
		5	355	15
	Math	3	338	23
		4	339	21
		5	355	20
	Science	4	339	6
Rochester	ELA	5	19	96
		7	66	29
	Math	3	18	39
		6	53	19
	Science	4	20	5
		8	74	16
Sanborn	ELA	5	115	41
		7	133	17
	Math	3	106	25
		6	98	7
	Science	4	106	12
		8	137	14

Table 51. *Number of persons and tasks by district, subject, and grade*

The variance of assessment (task) scores can be partitioned into independent sources of variation due to differences between persons, tasks, and the residual. This is called a one-facet crossed design. In this analysis, both persons and tasks are regarded as random samples from the universe of tasks and population of persons that could have been included. As a result, a random effects ANOVA can be used to estimate the four sources of variability in competency score data: systematic differences among persons (p), systematic differences among tasks (t), person-by-task interaction ($p \times t$), and random error. Random error is confounded with the $p \times t$ interaction. Variance component estimates and generalizability coefficients were calculated for both relative decisions (rank ordering) and absolute decisions (level of performance) because inferences of generalizability depend on how the data will be used.

Table 52 shows the estimated variance components and percent of total variance, both of which reflect the magnitude of error in generalizing from a student's score on a single assessment task to his or her universe score. For example, in all grade/subject combinations, one assessment (task) does not account for a large percent of the variance in individual student achievement (only 1-25%). The largest variance component tends to be the residual except in Concord where there is a larger sample size and the largest variance component tends to be students (or persons, p). Large residual variance suggests a few things: (1) a large $p \times t$ interaction; (2) sources of error variability in the competency score measurement that the one-facet $p \times t$ design has not captured, or (3) both. A large variance component for the $p \times t$ interaction indicates that the relative standing (or rank order) of students differs from assessment to assessment, which is well known in the measurement literature. We would expect that not all students would find the same tasks easy or difficult.

District	Grade/ Subject	Source of Variance	df	Sum of Squares	Mean Squares	Variance Components	% of Total Variance
CON	3ELA	<i>p</i>	337	1843.412	5.47	0.279	57.28%
		<i>t</i>	18	196.498	10.916	0.032	6.53%
		<i>p*t</i>	5983	1055.255	0.176	0.176	36.18%
	4ELA	<i>p</i>	338	1799.298	5.323	0.272	58.40%
		<i>t</i>	18	207.646	11.535	0.034	7.21%
		<i>p*t</i>	6049	972.976	0.16	0.160	34.39%
	5ELA	<i>p</i>	354	2303.548	34.126	2.236	79.12%
		<i>t</i>	14	77.674	2.73	0.006	0.21%
		<i>p*t</i>	4794	834.65	0.584	0.584	20.66%
	3MATH	<i>p</i>	334	2216.995	6.637	0.281	58.06%
		<i>t</i>	22	157.948	7.179	0.021	4.28%
		<i>p*t</i>	7073	1288.409	0.182	0.182	37.65%
	4MATH	<i>p</i>	338	2350.501	6.954	0.322	60.71%
		<i>t</i>	20	93.9	4.695	0.013	2.50%
		<i>p*t</i>	6495	1270.436	0.195	0.195	36.78%
	5MATH	<i>p</i>	352	2511.308	7.134	0.342	51.53%
		<i>t</i>	19	122.006	6.421	0.017	2.60%
		<i>p*t</i>	6356	1937.892	0.304	0.304	45.87%
	4SCI	<i>p</i>	36	461.025	1.372	0.206	58.80%
		<i>t</i>	5	18.632	3.726	0.011	3.02%
		<i>p*t</i>	1623	216.735	0.134	0.134	38.18%

Table 52. Variance component estimates for the person \times task G study by district, grade, and subject

District	Grade/ Subject	Source of Variance	df	Sum of Squares	Mean Squares	Variance Components	% of Total Variance
ROC	5ELA	<i>p</i>	18	614.261	34.126	0.349	33.39%
		<i>t</i>	95	259.367	2.73	0.113	10.79%
			168				
	7ELA	<i>p*t</i>	5	983.443	0.584	0.584	55.81%
		<i>p</i>	65	623.52	9.593	0.318	38.25%
		<i>t</i>	28	246.484	8.803	0.128	15.37%
			175				
	3MATH	<i>p*t</i>	4	674.476	0.385	0.385	46.38%
		<i>p</i>	17	24.946	1.467	0.033	14.61%
		<i>t</i>	38	26.907	0.708	0.030	13.20%
			584	96.34	0.165	0.165	72.19%
	6MATH	<i>p*t</i>	52	615.766	11.842	0.605	61.97%
		<i>p</i>	18	33.289	1.849	0.028	2.91%
		<i>t</i>	931	319.324	0.343	0.343	35.12%
			109				
	4SCI	<i>p*t</i>	19	15.486	0.815	0.122	29.81%
		<i>p</i>	4	7.167	1.792	0.079	19.43%
<i>t</i>		73	15.099	0.207	0.207	50.75%	
		73	339.19	4.646	0.267	31.33%	
8SCI	<i>p</i>	15	236.445	15.763	0.208	24.41%	
	<i>t</i>	109					
	<i>p*t</i>	5	412.742	0.377	0.377	44.26%	

Table 52 Cont'd

District	Grade/ Subject	Source of Variance	df	Sum of Squares	Mean Squares	Variance Components	% of Total Variance
SAN	5ELA	<i>p</i>	113	864.354	7.649	0.182	45.54%
		<i>t</i>	40	59.165	1.479	0.011	2.78%
		<i>p*t</i>	3766	779.463	0.206	0.206	51.68%
	7ELA	<i>p</i>	125	619.939	4.96	0.276	45.01%
		<i>t</i>	16	156.163	9.76	0.071	11.64%
		<i>p*t</i>	1950	517.776	0.266	0.266	43.36%
	3MATH	<i>p</i>	105	257.598	2.453	0.088	24.41%
		<i>t</i>	24	44.487	1.854	0.015	4.19%
		<i>p*t</i>	2141	551.075	0.257	0.257	71.41%
	6MATH	<i>p</i>	94	338.085	3.597	0.460	50.90%
		<i>t</i>	6	43.451	7.242	0.070	7.75%
		<i>p*t</i>	485	181.311	0.374	0.374	41.35%
	4SCI	<i>p</i>	103	139.849	1.358	0.104	47.39%
		<i>t</i>	11	5.073	0.461	0.003	1.50%
		<i>p*t</i>	735	82.189	0.112	0.112	51.11%
8SCI	<i>p</i>	132	1021.478	7.738	0.519	50.99%	
	<i>t</i>	13	52.129	4.012	0.026	2.54%	
	<i>p*t</i>	1562	738.725	0.473	0.473	46.47%	

Note. VARCOMP procedure in SAS was used to estimate sum of squares and mean squares.
Table 52 Cont'd

Generalizability theory also provides a reliability coefficient called a generalizability (G) coefficient. This G coefficient shows how accurate the generalization is from a student's observed score, based on a sample of the student's work, to his or her universe score. Applied to this analysis, the G coefficient represents the proportion of variability in observed assessment scores attributable to systematic differences in students' competency. Table 53 provides the variance component estimates and generalizability coefficients for both relative decisions (rank ordering) and absolute decisions (level of performance).

Other than grade 4 science in Rochester—where only 5 assessments were used to calculate a student's overall district-level competency scores—there are high G coefficients for both absolute and relative decisions. This means that the collection of classroom assessments provide stable estimates of student achievement in a given content domain.

District	Grade/ Subject	Relative error variance	Absolute error variance	Relative error generalizability coefficient $E\rho^2$	Absolute error generalizability coefficient ϕ
Concord	3ELA	0.009	0.011	0.968	0.962
	4ELA	0.008	0.010	0.970	0.964
	5ELA	0.039	0.039	0.983	0.983
	3MATH	0.008	0.009	0.973	0.970
	4MATH	0.009	0.010	0.972	0.970
	5MATH	0.015	0.016	0.957	0.955
	4SCI	0.022	0.024	0.902	0.895

Table 53. Variance component estimates and generalizability coefficients for relative and absolute error D study by district, grade, and subject

District	Grade/ Subject	Relative error variance	Absolute error variance	Relative error generalizability coefficient $E\rho^2$	Absolute error generalizability coefficient ϕ
Rochester	5ELA	0.006	0.007	0.983	0.980
	7ELA	0.013	0.018	0.960	0.947
	3MATH	0.004	0.005	0.888	0.870
	6MATH	0.018	0.020	0.971	0.969
	4SCI	0.041	0.057	0.746	0.680
	8SCI	0.024	0.037	0.919	0.879
Sanborn	5ELA	0.005	0.005	0.973	0.972
	7ELA	0.016	0.020	0.946	0.933
	3MATH	0.010	0.011	0.895	0.890
	6MATH	0.053	0.063	0.896	0.879
	4SCI	0.009	0.010	0.918	0.915
	8SCI	0.034	0.036	0.939	0.936

Table 53 Cont'd

In the D study, we show how increasing the number of assessments included in achievement estimates results in diminishing returns in terms of increasing reliability/generalizability beyond approximately 15-20 assessments. Figures 29-30 show sample plots using data from all districts showing estimated relative and absolute error generalizability coefficients as a function of the number of assessments by district, grade, and subject. It is important to note that the Grade 3 Math sample from Rochester contained only 18 students and therefore the lower line plot is most likely reflective of more error in the model.

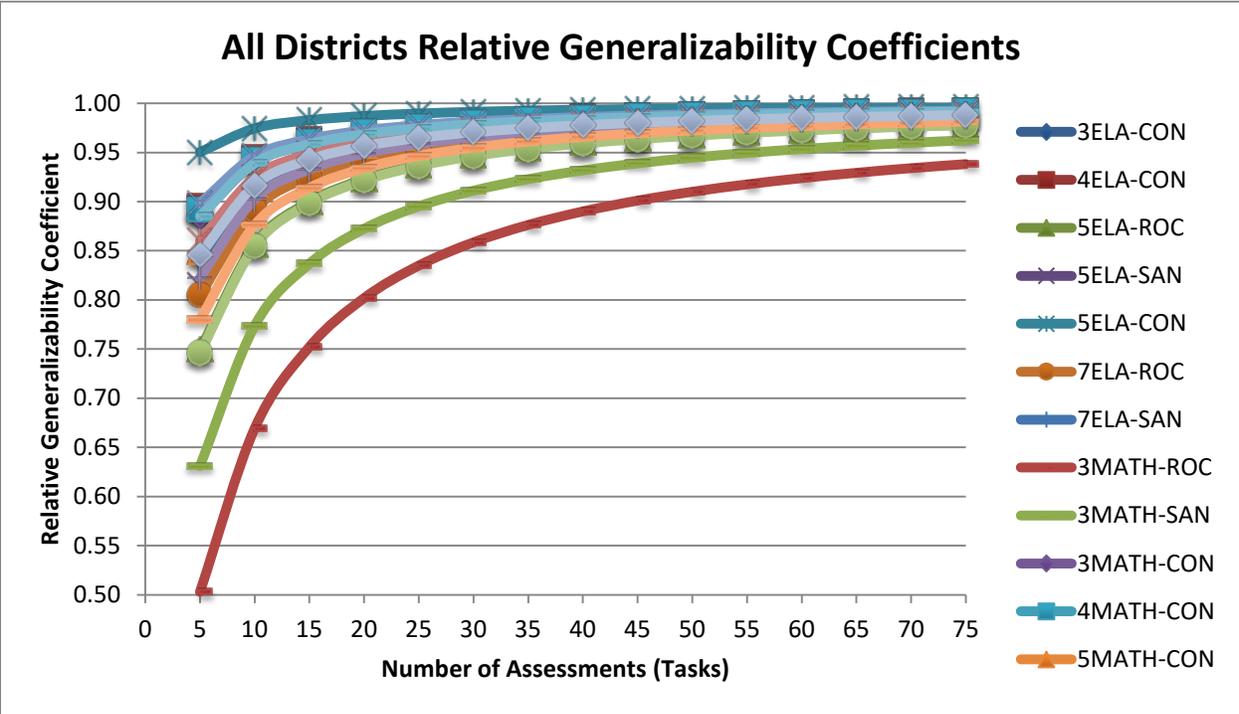


Figure 29. Sample plot showing estimated $E\rho^2$ scores as a function of the number of assessments by district, grade and subject

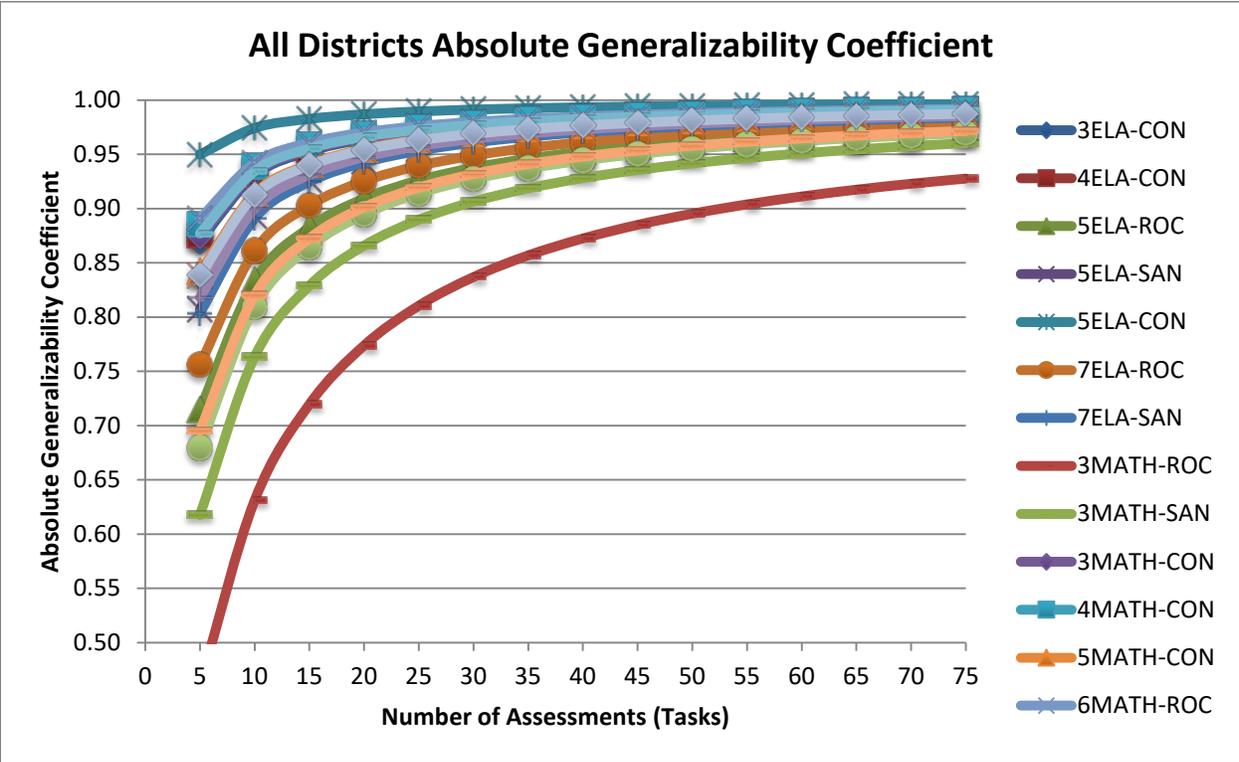


Figure 30. Sample plot showing estimated ϕ coefficient as a function of the number of assessments by district, grade and subject

Averaging across the grades and subjects by the number of assessments (tasks), there is a high degree of relative and absolute stability estimates (around 0.90) of student achievement between 15-20 classroom assessments—see Table 54.

Number of Assessments (Tasks)	$\overline{E\rho^2}$	$\overline{\Phi}$
5	0.82	0.79
10	0.90	0.88
15	0.93	0.92
20	0.94	0.93
25	0.95	0.95
30	0.96	0.95
35	0.97	0.96
40	0.97	0.97
45	0.97	0.97
50	0.98	0.97
55	0.98	0.97
60	0.98	0.98
65	0.98	0.98
70	0.98	0.98
75	0.98	0.98

Table 54. Average estimated $E\rho^2$ scores (relative generalizability coefficient) and Φ coefficient (absolute generalizability coefficient) as a function of the number of assessments across districts, subjects and grades

Based on our limited analyses thus far, the results suggest that classroom assessments can provide reliable estimates of student achievement for use in a school accountability context like the NH PACE pilot project. Approximately 10-15 assessments per year provide for an efficient trade-off while still ensuring a high degree of relative and absolute decision reliability. Given the concerns noted in Grade 4 Science in Rochester last year as well as this year, follow-up conversations and targeted supported is needed to help teachers in Rochester Grade 4 Science understand the impact of a low number of grade book assessments on generalizable estimates of student proficiency.

Appendix D: Calibration Analysis in 2015, 2016, and 2017

Calibration Analysis 2015

Annual determinations are currently being reported as a result of the PACE Pilot. Therefore, claims about comparability must hold at the level of the annual determination. This means, that proficient in one district should mean the same or at least very similar things as proficient in another district. It is with this in mind, a full-day calibration audit was run to evaluate the comparability claims made in the PACE system. The calibration audit is intended to collect evidence about the degree of difference in scoring across districts that can be used to support decision making about adjustments to cut scores. On August 10, 2015, 78² teachers and leaders from seven of the eight PACE districts participated in the calibration audit. To start the day, participants were given an introduction to the purpose of the meeting by Scott Marion and Paul Leather, and were then welcomed by New Hampshire Commissioner of Education, Virginia Barry. High school science teachers and leaders were then separated from the rest of the group to participate in a modified training and calibration process due to the uniqueness of the high school science situation to be detailed later in this report. The full agenda for calibration workshop is included in Appendix AA.

Overview of the Consensus Scoring Method

The consensus scoring method involves pairing teachers together, each representing different districts, to score student work samples. The student work samples were gathered from each PACE common performance tasks from of the four districts participating in the 2014-2015 PACE accountability pilot: Epping, Rochester, Sanborn, and Souhegan. Both judges within each pair were asked to individually score their assigned samples of student work. Working through the work samples one at a time, the judges would discuss their individual scores and then come to an agreement on a “consensus score”. In the very few cases where consensus could not be reached, an expert scorer (who did not have affiliation with any particular district) would decide on the appropriate consensus score. Each pair was asked to score six random samples of student work (from across all districts) in the morning. In the afternoon, pairs were shuffled so that teachers were still working with colleagues representing different districts, but working with a new partner to score another six student work samples. The purpose of collecting consensus score data is to get the best estimate of the “true score” to be used as a “calibration weight.” These consensus scores are then used in follow-up analyses to detect any systematic, cross-district differences in the stringency of standards used for scoring. This method is based on the cross-moderation method used in the UK to evaluate the comparability of the General Certificate of Secondary Education. The history and a full description of the method can be found in the book, *Techniques for monitoring the comparability of examination standards* (Adams, 2007).

Judge Training

The judge training was led by Carla Evans and was conducted in a large group setting where teachers from all subjects and grade levels, with the exception of high school science, were trained together on the consensus scoring process. During the training, the judges were equipped with guidelines for establishing consensus scores, and a demonstration of how the process should work. The logistical details were also covered to help ensure a high level of data entry accuracy. The training materials can be found in Appendix BB.

² This is the number of registered teachers, actual participation still under verification.

Analysis & Results

Teachers and leaders were able to come to agreement on consensus scores for 460 student work samples. The consensus and original, teacher-given rubric scores used in this analysis are the result of averaging across all scored rubric dimensions. The mean score is used because often, the teacher-given scores and consensus scores were determined using different versions of the scoring rubric. The different versions sometimes entailed disordinal rubric dimensions, or even differing numbers of scorable dimensions all together. Therefore, the individual analytical rubric “sub scores” were not analyzed. Teacher-given scores on rubric dimensions assessing the NH Work-Study Practice Competencies were withheld from the mean score calculation. In the case of double scoring, the double scores were averaged to produce a single teacher-given score and a single consensus score for each student work sample. Once the consensus scores had been compiled from their respective flash drives and excel files, the only unique identifier for each piece of student work was the student ID (SASID); subject area was not recorded. Unfortunately, the SASIDs were not unique to each piece of work meaning that in some cases, schools had submitted more than one work sample from the same student (e.g., one from math, one from ELA). In order to ensure the consensus and teacher scores were being matched correctly, any SASID that appeared more than once in the file was discarded. This brought the sample size of consensus scores down to 400, which is likely to have a negligible influence on the estimates resulting from this analysis. Bias due to this type of listwise deletion is not likely an issue since there is no reason to assume that the missing data is not missing completely at random. The duplicate cases were submitted relatively evenly across all districts, 11 from Epping, 4 from Rochester, 3 from Sanborn, and 9 from Souhegan³. Due to data entry issues by both the judges recording consensus scores and the schools submitting teacher scores, of the 400 student work samples eligible for matching with teacher-given scores, only 353 were successfully matched⁴. Table 1 reports the frequency of scores across grade levels⁵, subject areas, and district.

Table 1. Number of SASIDs by Grade, Subject, and District

Grade	Frequency	Subject	Frequency	District	Frequency
3	26	ELA	135	Epping	106
4	47	Math	156	Rochester	108
5	52	Science	62	Sanborn	117
6	41	Total	353	Souhegan	22
7	58			Total	353
8	29				
9	50				
10	50				
Total	353				

³ These numbers do not match the numbers of duplicates removed in the consensus score file due to non-matching SASIDs across the district submitted and consensus score files.

⁴ Data entry errors were not minimized by the fact that the SASIDs are ten-digits long. In future iterations of this process, a better way to identify student work samples should be devised.

⁵ Often grade level was not reported for Geometry and Algebra. When this data was missing, it was assumed that Geometry was taken in 10th grade and Algebra in 9th.

To detect any systematic discrepancies in the relatively leniency and stringency of district scoring, we calculated a mean deviance index. This index is the mean difference between the consensus score and teacher score across all student work samples for each district as calculated by the following, for District k :

$$Deviance_k = \frac{\sum_i^n (teacher_i - consensus_i)}{n_k}$$

Using this index, a negative mean deviance would indicate systematic underestimation of student scores by classroom teachers (i.e., district stringency), and positive mean deviance scores would indicate systematic overestimation of student scores by classroom teachers (i.e., district leniency). The values of the deviance metric are on the scale of the rubric points. Table 2 below shows the average observed deviance by district. As an example, the interpretation of the mean deviance for Epping is that, on average, teachers in Epping scored their student work on the common performance tasks .235 points higher than the same work was scored by the consensus raters.

Table 2. Average Deviance by District

District	N	Deviance
Epping	106	.2349
Rochester	108	.4367
Sanborn	117	.2915
Souhegan	22	.3939

Across all districts, the consensus scoring yielded scores that were a bit lower than the teacher-given scores. This finding is not necessarily problematic from a comparability perspective, as long as the relative leniency of the teacher-given scores is even across districts. A factorial analysis of variance was run in order to investigate the variance in the discrepancy index that can be attributed to differences in districts including the interaction effects of district and grade level, and district and subjects. The results of this ANOVA are shown in Table 3 below.

Table 3. Deviance by District, Grade and Subject - ANOVA

	df	F	Sig.	Effect Size $\eta^2_{partial}$
District	3	3.108	.027	.029
Grade	7	1.969	.059	.043
Subject	2	.332	.717	.002
District * Grade	14	2.201	.008	.091
District * Subject	4	2.698	.031	.034
Grade * Subject	4	.655	.623	.008
District * Grade * Sub	8	2.078	.038	.051

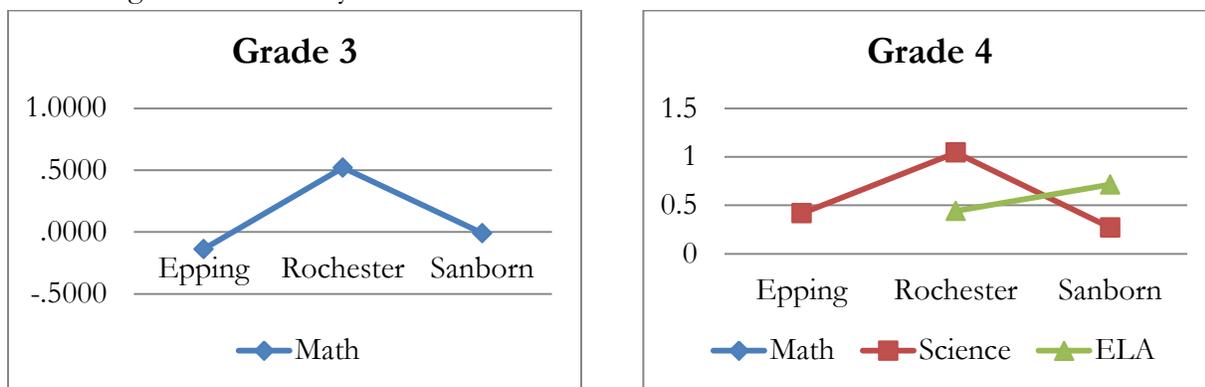
The results show that the variation in the deviance index across districts is statistically significant at $\alpha = .05$. However, interpretation of this finding is limited given the statistical significance of the interaction effects. The significant three-way interaction effect indicates that the relationship between district and subject area changes by grade level. This means that average deviance varies depending on the unique district, grade-level, and subject area combination. Pairwise post-hoc analyses (as shown in Table 4) reveal that there are not significant differences in mean deviance between any two districts, rather, the significance in the district main effect is driven by the interaction effect (i.e., differences in the unique district, grade, and subject units).

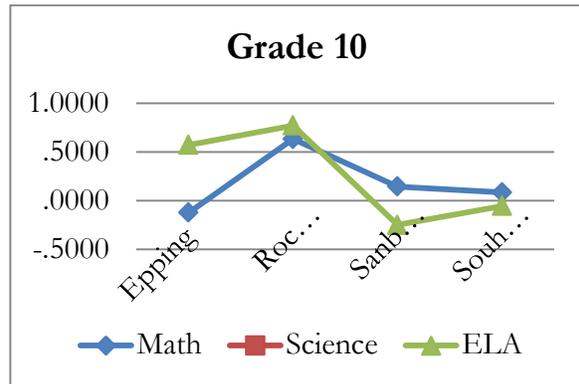
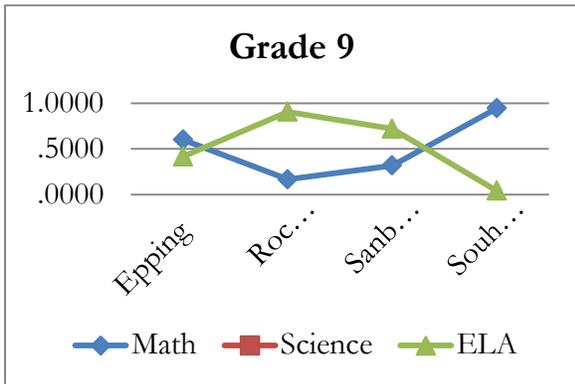
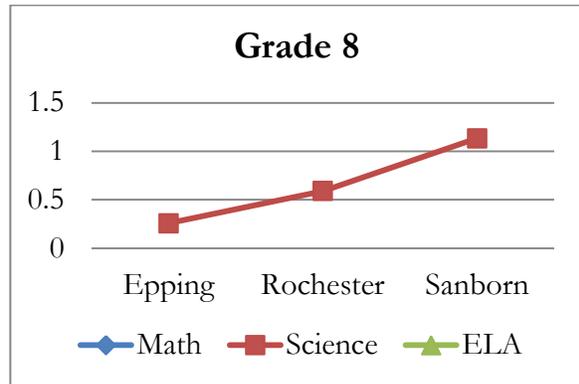
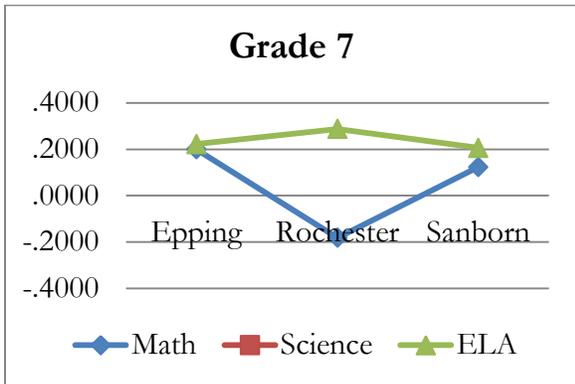
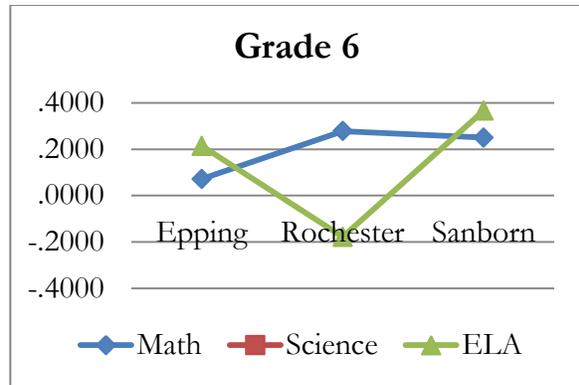
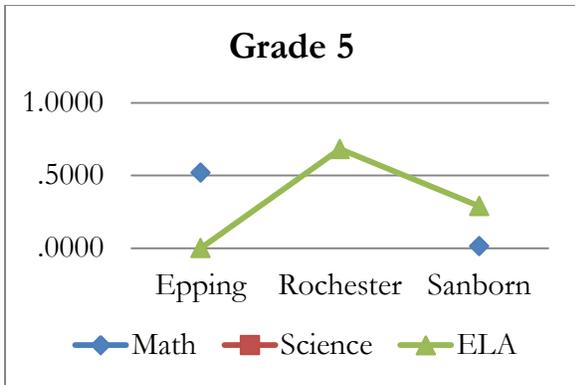
Table 4. Follow-up Pairwise Comparisons

District 1	District 2	Mean Difference	Standard Error	Sig.
Epping	Rochester	-.2018	.08670	.124
Epping	Sanborn	-.0565	.08503	1.000
Epping	Souhegan	-.1590	.14857	1.000
Rochester	Sanborn	.1452	.08462	.523
Rochester	Souhegan	.0428	.14833	1.000
Sanborn	Souhegan	-.1024	.14736	1.000

Figure 1, on the next page, illustrates the interaction effect by demonstrating the relationship between subject and district changes by grade level. Despite the statistical significance of the interaction effect overall, interpretation of any single data point is cautioned due the limited sample size of any particular district, grade-level, subject unit.

Figure 1. Three-way interaction.





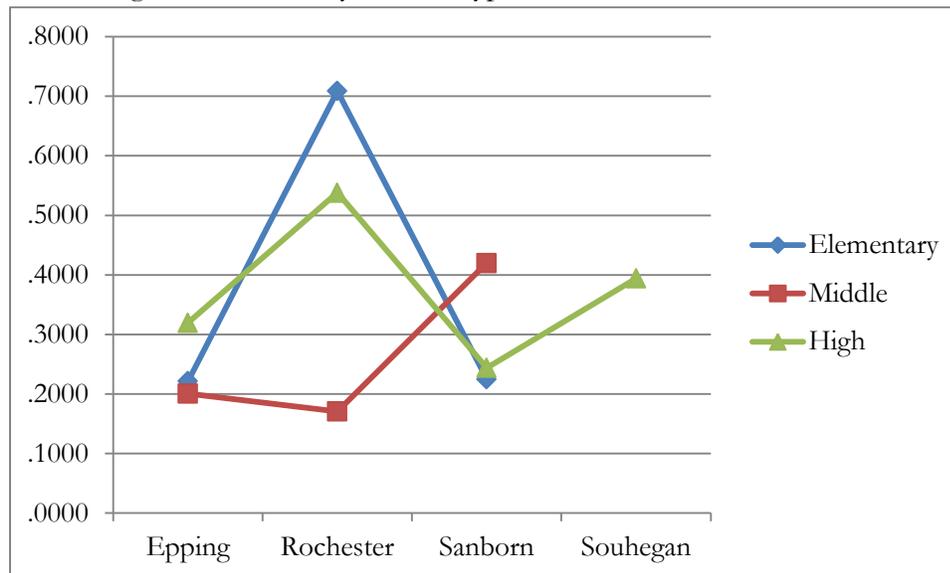
One interesting pattern that emerges from examining the figure above is that Rochester seems to have a higher deviance than the other districts in many of the subject areas and grade levels. With the exception of the middle school grade levels, where the deviances in math and ELA are counter-balanced, Rochester seems to have a more lenient standard in scoring than the other districts. To test this hypothesis, we ran a follow-up ANOVA to test for a district by school type (e.g., elementary, middle, and high school) effect. The results of this ANOVA are presented in Table 5.

Table 5. Deviance by District and School Type - ANOVA

	df	F	Sig.	Effect Size $\eta^2_{partial}$
District	3	2.112	.098	.018
School Type	2	1.149	.318	.007
District * School Type	4	3.527	.008	.040

In this model, district and school type alone are not sufficient for explaining variation in the deviance index, however, the district by school type interaction effect is statistically significant at the $\alpha = .05$ level. Figure 2 illustrates this interaction effect.

Figure 2. District by School Type interaction



While these findings do not support making unilateral adjustments to the district cut scores, they do suggest a possible need for increased cross-school and cross-district calibration.

Ranking Audit: High School Science

High school science presented a unique challenge in calibrating the cross-district scores because there were no common tasks across districts; each district assigned completely unique tasks for the three subject areas—Earth science, physical science, life science. Typically, score calibration procedures require one of two conditions to be met: 1) common persons across tasks, or 2) common tasks across persons. Because neither of these conditions was satisfied in the 2014-2015 implementation of PACE performance tasks for high school science, we looked to alternate methods of score calibration. Thus, the high school life science calibration process for PACE was modeled after a judgmental, ranking cross-moderation method used to maintain comparability of written examination standards across multiple awarding bodies for England’s National Curriculum Test (Bramley, 2005).

Overview of Method

Unlike the consensus scoring methodology used for the other PACE subjects and grade levels, the calibration method used for high school science involved an individual rank-ordering process. Because we only had eight judges, we limited the calibration process to the life science tasks with the assumption that due within-district calibration, any systematic scoring differences discovered in life science, would likely apply to the remaining two disciplines. The eight participating judges were given packets of student work that had been grouped by average rubric score, and asked to rank order the student work based on quality. Each packet contained 10 student work samples and student work from all four districts was represented in each packet. The order of the papers placed in the packet was arbitrary. Each teacher was asked to rank 4 packets (1,3,5,7; 2,4,6,8; or 9,10,11,12), this ensured that every teacher saw 40 of the 45 work samples. See Appendix CC for the packet assignment of for the student work samples. Two student work samples (highlighted in yellow in Appendix CC) were not ranked due to poor copy quality, this brought the total number of student work samples to 12, 10, 10, and 11 from Epping, Rochester, Sanborn, and Souhegan, respectively.

For a complete description of the method on which this PACE calibration is based, please see “A Rank-Ordering Method for Equating Tests by Expert Judgment” by Tom Bramley in the 2005 volume of the *Journal of Applied Measurement*.

Judge Training

Before beginning the ranking exercise, judges were first asked to familiarize themselves with each of the different tasks. In order to do so, the judges read through blank copies of the tasks and the associated Tool 8. Then, for each task, a district representative was asked to briefly provide an overview of the performance task, including any parts that were particularly useful for discriminating among students and items or parts that were particularly difficult or did not run smoothly (as this was often the first year the tasks has been implemented). Judges then took the opportunity to ask clarifying and follow-up questions to the district representative. There were no high school science teachers present from Sanborn, so in order to familiarize themselves with the task from that district, judges discussed their impressions in pairs. One pair was asked to report out as if they were the district representatives and a professional large-group discussion around the task ensued.

Once the judges felt comfortable with the four tasks, judges were trained on the ranking process. The instructions for the judges were based also on similar studies completed in England. As Bramley (2007) explains, “The need for the whole exercise in first place arises from the fact that the different boards [i.e., districts in the case of PACE] have different specifications and question papers. The judges are really therefore being asked to judge which performance is better, taking into account any differences in the perceived demands of the questions (and specifications)” (p. 265). The judges involved in the PACE calibration for high school science were likewise instructed to rank papers based on merit, evidence of student understanding, demonstrated competence, and student knowledge of nature of science, which are all different ways of saying “better,” as Bramley (2007) succinctly puts it. In order to minimize construct irrelevant variance, judges were also explicitly told to not rank on task specific entities, handwriting, grammar (when not relevant to construct), length, and the quality of the copy job (e.g., dark or light markings). For the training materials used, see Appendix CC.

Analysis & Results

The rank-order datasets resulting from the teacher and leaders work on August 10, were re-organized to represent dependent⁶ pairwise comparisons. The Thurstone model for paired comparison data was used to fit a unidimensional scale representing quality, on which each student work sample will be located:

$$\ln\left[\frac{P_{ij}}{1 - P_{ij}}\right] = B_i - B_j$$

where P_{ij} is the probability that work i beats work j , and B_i and B_j are their respective measures on a unidimensional scale. The results of this analysis indicate placement of student work along an interval-level scale representing quality. Similar to the discrepancy analysis completed for the consensus score results, the Thurstone scores can be compared to the original teacher scores with a deviance index. However, unlike the teacher-given and consensus scores, the Thurstone scores and the teacher-given scores are not on the same scale. To account for the differences in the scales, both sets of scores were first transformed into standard scores before calculating the deviance index.

$$Deviance_k = \frac{\sum_i^n (zscore_{teacher} - zscore_{Thurstone})}{n_k}$$

⁶ Though the comparisons are dependent in that they are self-consistent, treatment of the comparisons as independent should produce measurements that are close to linearly related to the measures produced had the dependence been accounted for (Smith & Smith, 2007). Due to the increased computational load produced when dependent rankings are long, the pairs are treated as independent. The range of the measures will likely be overestimated and the standard errors underestimated; therefore, the results of the analysis will be treated as an upper bound on the amount of discrepancy in scoring across districts.

This metric can be interpreted similarly to the deviance calculated from the consensus scores, where positive deviances indicate district leniency, however, the units are not on the scale of the rubric scores, but rather represent standard deviation units.

The Thurstone model ran successfully and has good data-model fit indices. There were no student work samples or raters with infit greater than 2.0, and only one paper with outfit greater than 2.0. Additionally, the separation reliability is .96 indicating that the rank ordering procedure resulted in a strong differentiation in quality for student work. Figure 3 shows the respective score distributions for the rubric scores and the Thurstone measures. Both distributions of scores are approximately normal.

Figure 3. Score distributions

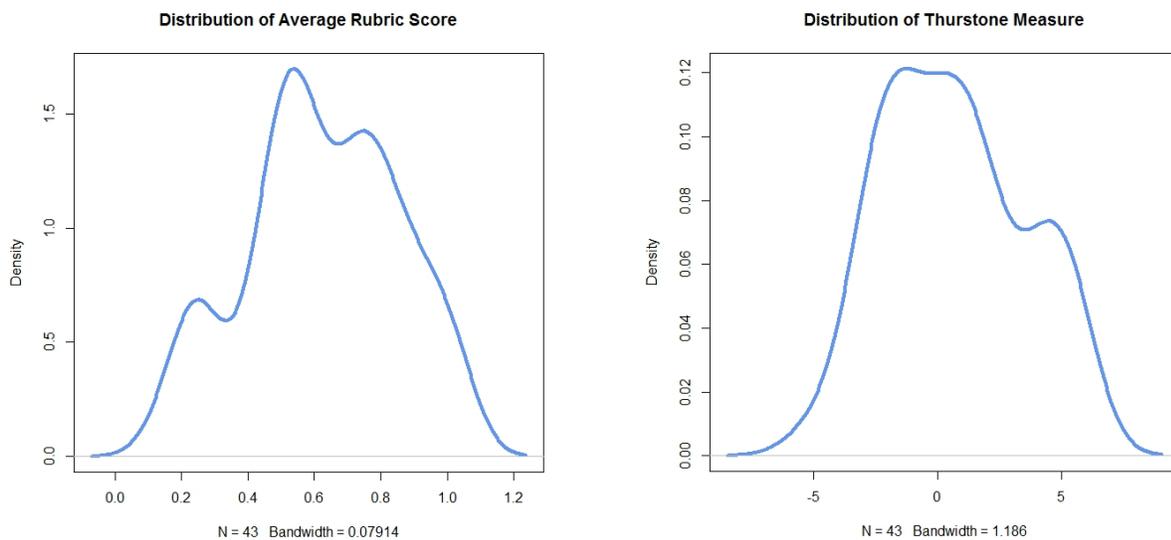
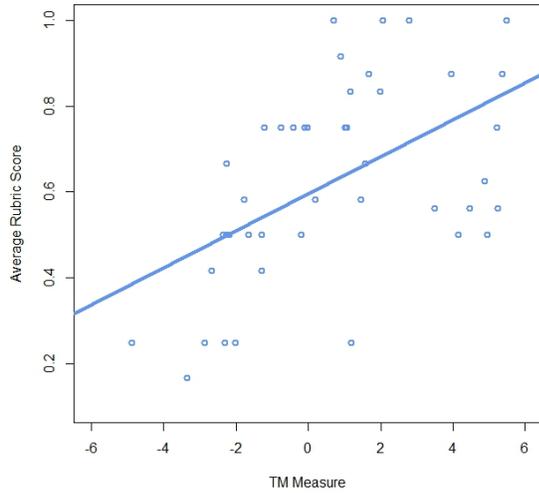


Figure 4 shows that there is a linear relationship between the Thurstone measure and the rubric scores. The existence of a linear relationship means the scores on the two scales can be meaningfully compared. The correlation between the average rubric scores and the Thurstone measure is strong, ($r = .526, p < .001$). Because a significant portion of variance in these two measures is shared, we can infer that the judgmental ranking analysis yielded reasonable estimates of student work quality.

Figure 4. Linear Relationship between the two scales



In order to determine if any of the districts is scoring systematically more leniently or stringently than the other districts, a deviance analysis was run. Table 6 shows the average deviance score for each district.

Table 6. Mean Deviance Scores by District

District	N	Mean Deviance
Epping	12	.258
Rochester	10	.165
Sanborn	10	.710
Souhegan	11	-1.07

These mean differences indicate systematic differences in the location of district work in the rubric score and Thurstone score distributions. Though district designation is conflated with distinctions on many other factors including task and student population, because all tasks were judged by the same people and placed on the same scale, this discrepancy score represents only differences in *scoring*, rather than differences in tasks or student populations. Unlike the results for the consensus scoring, these results cannot be directly interpreted in the scale of the rubric scores. Rather, these reflect relative differences in leniency or stringency across districts. Because judges were only asked to rank student work rather than score student work, the results of this analysis can only reveal *relative* differences in district leniency; the mean deviance metric is a “zero sum game.”

The results do reveal scoring differences across districts, most notably in Souhegan. On average, Souhegan teachers are scoring their student work a full standard deviation below where the judges placed the same student work in the sample. To a lesser extent, the opposite is the case for Sanborn, where the rubric scores tend to be systematically higher than they were judged during the calibration.

Rater Bias

Because Souhegan fared so favorably in the rank ordering exercise, we decided it would be worth checking into the possible effects of any kind of rater bias. Table 7 below shows that judges participating in the rank ordering were not evenly distributed across districts, in fact, there were three judges representing Souhegan, while all other districts had no more than one representative.

Table 7. Number of Judges by District

District	Number of Participants
Concord	1
Epping	1
Rochester	1
Sanborn	0
Souhegan	3
State	1

One possible reason why Souhegan seemed to fare especially well in the rank order exercise might be that judges, likely subconsciously, tend to favor the task and student work coming from their own district. If present, this kind of bias can be detected by looking at the relative rank ordering of districts (across packets) by judges. Because the packets were grouped roughly by average rubric score, the quality of the work is naturally controlled for when examining the median rank of each district across packets. Table 7 presents the results of these analyses.

Table 7. Mean District Rank -- by Judge

<u>Judge</u>	<u>Judge District</u>	<u>Median rank by district</u>	<u>Rank order of districts</u>
1	Souhegan	3.0	Souhegan
		4.0	Rochester
		6.5	Sanborn
		7.0	Epping
2	Concord	3.0	Epping
		3.5	Souhegan
		6.0	Rochester
		7.0	Sanborn
3	Rochester	2.0	Souhegan
		5.0	Rochester
		6.0	Sanborn
		8.0	Epping
4	State	2.0	Souhegan
		3.5	Epping
		7.0	Rochester
		8.5	Sanborn
5	Epping	3.0	Souhegan
		5.0	Rochester
		6.0	Sanborn
		7.0	Epping
6	Souhegan	2.5	Souhegan
		6.0	Epping
		6.5	Sanborn
		7.5	Rochester
7	Souhegan	2.0	Souhegan
		5.0	Epping
		7.0	Rochester
		7.5	Sanborn

Luckily, we did not find any evidence to suggest that judges tend to rank work from their own district more favorably than work from other districts. Rather, we see that the student work from Souhegan was consistently ranked highly across all judges. Interestingly, the median rank orders have a high degree of spread, which indicates that the rank ordering of work within packets was strongly predicted by district. This is one more piece of evidence to suggest that there are cross-district differences in stringency of scoring for high school science.

Discussion & Recommendations

Taking all of the evidence into account, we do not recommend that any adjustments be made to the district-level cut scores. Evaluating the results of the high school science calibration in conjunction with the consensus scoring analysis suggests that the effect detected in Souhegan for high school life science is just more evidence of the three-way district, by grade-level, by subject area interaction. Due to this significance of this interaction effect, it is clear that further efforts to strengthen the comparability of cross-district scoring are needed. Part of this can be accomplished through higher quality task and rubric design, an effort already underway, as well as cross-district scorer training. We do not find the district differences meaningful enough to warrant changes to the district-level cut scores. Such comparability challenges are not unexpected during the first year of this complex pilot and the results of this study point out areas where improvement is necessary.

References

- Adams, R. (2007). Cross-moderation methods. *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.
- Baird, J.-A. (2007). Alternative conceptions of comparability. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 124–165). London: Qualifications and Curriculum Authority. Retrieved from <http://hdl.handle.net/1983/1004>
- Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, 6(2), 202-223.
- Bramley, T. (2007). Paired comparison methods. In P. E. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). London: Qualifications and Curriculum Authority.
- Firth, D., & Turner, H. L. (2012). Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software*, 48(9).
- Smith, E. V., & Smith, R. M. (2007). *Rasch measurement: Advanced and specialized applications*. JAM Press.

Calibration Analyses 2016

The PACE innovative assessment system uses common performance tasks across districts to evaluate the degree of comparability in local scoring. These analyses rest on the assumption that patterns in scoring for the common tasks is representative of district relative stringency and leniency in scoring of the local performance tasks and assessments.). The calibration audit is intended to uncover differences in scoring between districts that can be used to support decision-making about any adjustments to cut scores that may be need to be considered due systematic cross-district differences. The scores of student work on PACE performance tasks that result from this audit serves as the “calibration weights” so that more generalized inferences about relative leniency or stringency of district scoring practices can be made. On July 25th, 2016, teachers and leaders from the eight PACE districts participated in the calibration audit.

The moderation audit in 2016 was closely modeled on the same process conducted in the summer of 2015 with incremental improvements based on lessons learned (e.g., the evaluation of student work and scoring will all occur online rather than paper-based). This audit is heavily based on methods that have been successful in Queensland, Australia for decades. The consensus scoring method involves pairing teachers together, each representing different districts, to score student work samples. The student work samples were gathered from each PACE common performance tasks from of the eight districts participating in the 2015-2016 PACE pilot. Both judges within each pair were asked to individually score their assigned samples of student work. Working through the work samples on at a time, the judges would discuss their individual scores and then come to an agreement on a “consensus score”. In the very few cases where consensus could not be reached, an expert scorer (who did not have affiliation with any particular district) would decide on the appropriate consensus score. The purpose of collecting consensus score data is to get the best estimate of the “true score” to be used as a “calibration weight.” These consensus scores are then used in follow-up analyses to detect any systematic, cross-district differences in the stringency of standards used for scoring.

Students with scores for any rubric dimension that were out of range were removed listwise. Consensus scores were matched with the local, teacher-given task scores on Student ID, district, grade, and subject. This matching resulted in 1,417 total students with both consensus scores and local scores for the common task work. The distribution of these students across grades, subjects, and district is provided in Table 11.

Table 11.

Number of Matched Students by Grade, Subject, and District

Grade	Subject	Concord	Epping	Monroe	Pittsfield	Rochester	Sanborn	Seacoast	Souhegan
3	Math	18	18	7	18	0	16	2	
4	ELA	16	18	9	18	18	6	14	
	Sci	15	17	6	16	14	11	12	
5	ELA	17	18	11	18	18	14	12	
	Math	16	17	9	17	17	17	12	
6	ELA	18	18	11	17	18	17	10	
	Math	17	14	7	18	17	19	15	
7	ELA	17	18	5	18	17	6	17	
	Math	17	16	2	14	17	17	15	
8	Sci	14	12	5	14	16	13	9	
9	ELA	13	18		18	16	15		14
	Math	11	15		17	14	13		8
	Sci	0	7		0	16	7		9
10	ELA	18	18		16	14	14		6
	Math	11	15		13	17	12		9
	Sci	16	9		13	12	14		2
Total		234	248	72	245	241	211	118	48

To detect any systematic discrepancies in the relatively leniency and stringency of district scoring, we calculated a mean discrepancy index. This index is the mean difference between the consensus score and teacher score across all student work samples for each district as calculated by the following, for District k:

$$Discrepancy_k = \frac{\sum_i^n (teacher_i - consensus_i)}{n_k}$$

A negative mean discrepancy would indicate systematic underestimation of student scores by classroom teachers (i.e., district stringency), and positive mean discrepancy scores would indicate systematic overestimation of student scores by classroom teachers (i.e., district leniency). The values of the discrepancy metric are on the scale of the rubric points. Table 12 shows the average observed discrepancy by district.

Table 12.
Average Discrepancy by District

District	Discrepancy	N	Std. Deviation
Concord	0.259	234	0.55
Epping	0.281	248	0.68
Monroe	0.547	72	0.65
Pittsfield	0.342	245	0.65
Rochester	0.341	241	0.61
Sanborn	0.227	211	0.66
Seacoast	0.194	118	0.68
Souhegan	0.335	48	0.55

The observed positive discrepancies indicate a systematic overestimation of common task scores by the classroom teachers. Positive discrepancy scores are not necessarily problematic from a comparability perspective; we mainly interested in looking for differences among the districts in average discrepancy. Monroe’s average discrepancy score stands out as being particularly high. Post-hoc analyses with a Bonferroni correction revealed that the district marginal deviances are not significantly different from one another except for Monroe, where the deviance is significantly higher than Concord, Epping, Sanborn, and Seacoast.

A three-factor analysis of variance reveals a significant 3-way interaction for district, by grade, by subject combinations. This means we cannot justify any unilateral adjustments to any one districts’ cut scores across the board. Instead, more nuanced decisions must be made based on follow-up analyses.

Table 13.
ANOVA – District by Grade by Subject

Source	df	F	Partial Eta Squared	Sig.
District	7	4.121	.021	.000
Grade	7	5.095	.026	.000
Subject	2	4.399	.007	.012
District * Grade	38	4.371	.112	.000
District * Subject	14	5.211	.053	.000
Grade * Subject	6	2.021	.009	.060
District * Grade * Subject	28	2.296	.047	.000

R Squared = .236 (Adjusted R Squared = .177)

The plots generated by this analysis of variances are provided for each subject area in Figures 2, 3, and 4 on the next pages.

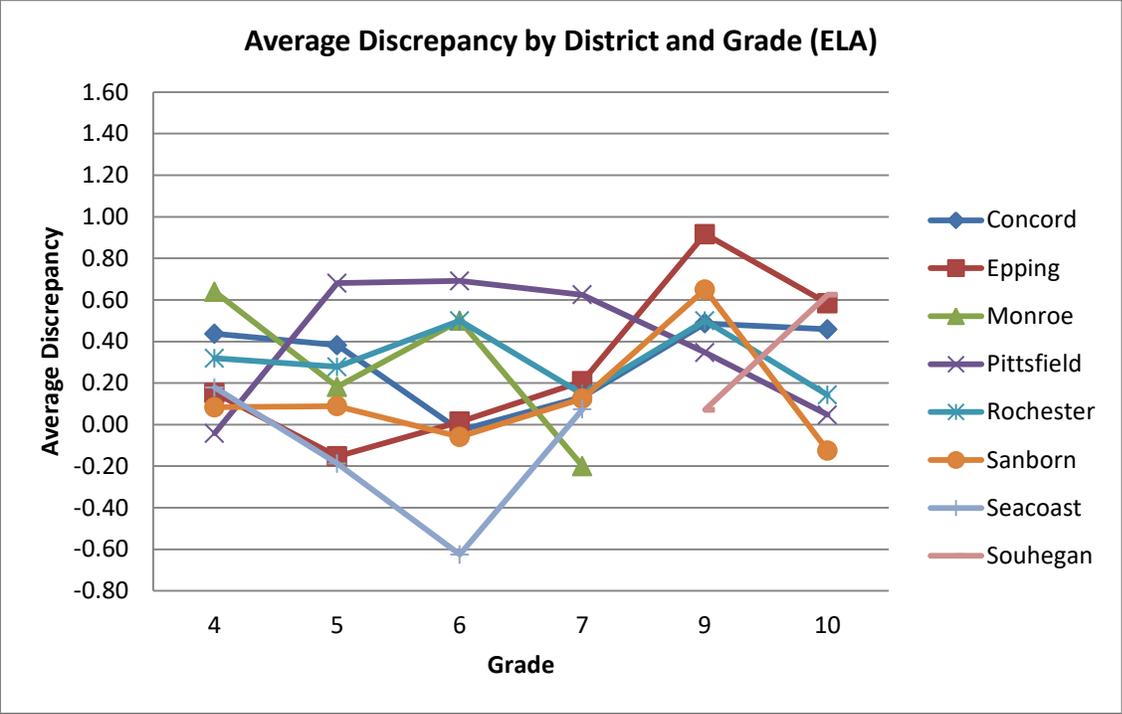


Figure 2. Marginal Means for ELA

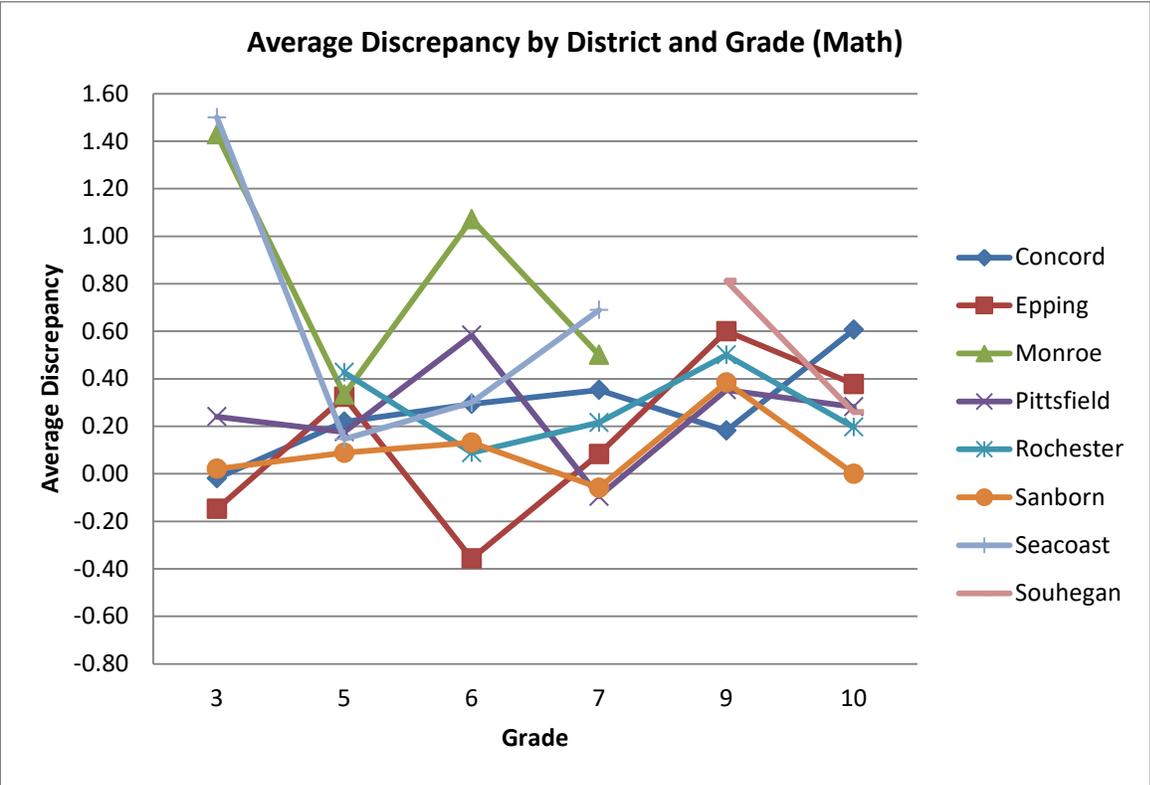


Figure 3. Marginal Means for Math

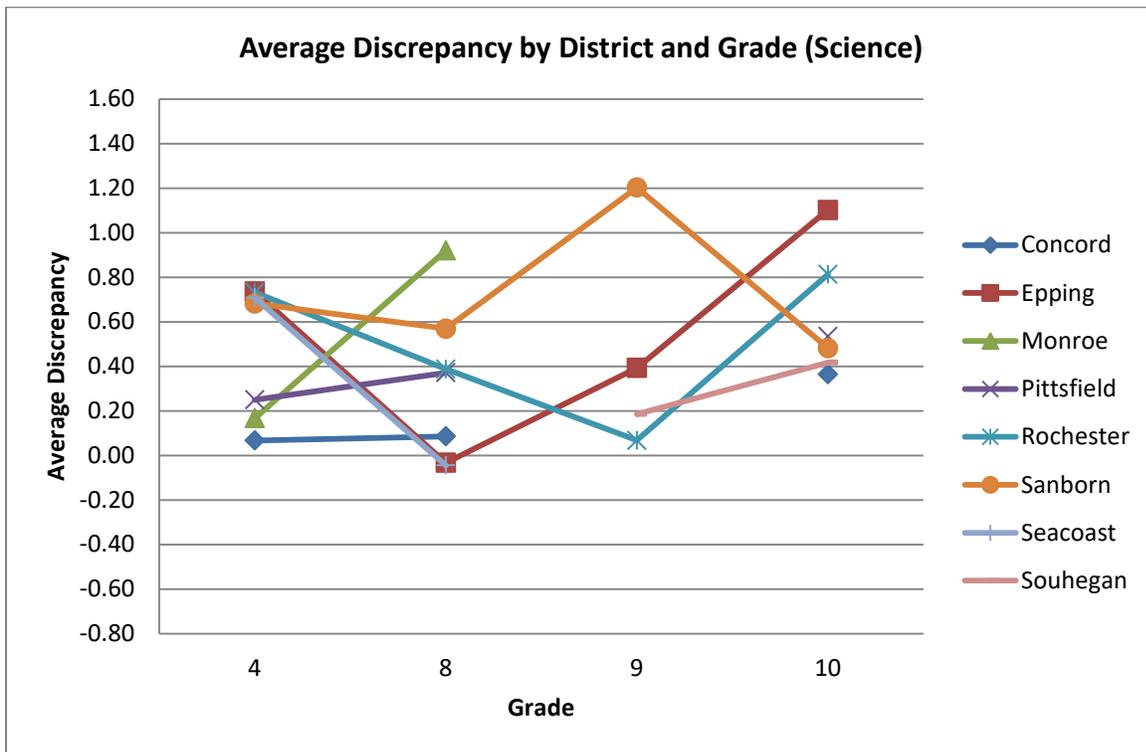


Figure 4. Marginal Means for Science

Overall, it seems that the ELA teachers and consensus scorers are more consistent than the teachers and scorers in math and science. The one exception seems to be Seacoast Grade 6 ELA which stands apart from the rest as having a strong, negative discrepancy score. This may indicate stringent scoring on the part of the Grade 6 ELA teacher in Seacoast. However, it may be that the high fluctuation in Seacoast is more of a function of the particularly small sample size for this public charter school. To follow-up further, Seacoast Grade 6 ELA is flagged for additional review.

To more deeply investigate the earlier findings with Monroe, we looked at the grade level and subject combinations where Monroe's discrepancy is significantly different than the other districts'. Using complex contrast post-hoc analyses, with no type-1 error correction, we analyzed the mean differences in discrepancy for Monroe as compared with all other districts for each subject and grade. The equality of variance assumption was met for all combinations except fifth grade math for which the appropriate *t* value correction was made.

Table 14.
Follow-up comparisons for Monroe

Subject	Grade	Mean Difference	<i>t</i>	<i>df</i>	Sig.
ELA	4	-0.442	-1.907	97	.059
	5	0.024	.140	106	.889
	6	-0.365	-1.589	107	.115
	7	0.434	1.693	96	.094
Math	3	-1.364	-6.746	77	.000
	5	-0.099	-.520	9.608	.615
	6	-0.881	-3.956	105	.000
	7	-0.302	-.605	96	.546
Sci	4	0.348	1.241	89	.218
	8	-0.674	-2.760	81	.007

For Monroe, the following grades and subjects show evidence of significant overestimation of scores, Grade 3 Math, Grade 5 Math, and Grade 8 Science, which have the following discrepancy averages respectively, 1.43, 1.07, and .92. These discrepancy scores can provide benchmarks within each of the math and science subject areas to flag high discrepancy averages. Using these scores as the flagging criteria for identifying other high scores, the following district by grade by subject combinations are identified for further review: Seacoast Grade 3 Math, Sanborn Grade 9 Science, and Epping Grade 10 Science.

Table 15.
Flagged Discrepancy Scores with Cut Scores

District	Grade	Subject	Average Rubric Discrepancy	Competency Score Scale	Cut Scores		
					Level 2	Level 3	Level 4
Epping	10	Sci	1.102	0-100	62.26	71.87	93.30
Monroe	3	Math	1.429	1.00-4.00	2.26	2.66	3.05
Monroe	6	Math	1.071	1.00-4.00	2.02	2.25	3.00
Sanborn	9	Sci	1.202	1.00-4.00	1.60	2.79	3.63
Seacoast	3	Math	1.500	1.00-4.00	1.88	2.75	3.76
Seacoast	6	ELA	-0.625	1.00-4.00	1.60	2.81	3.35

With each of the flagged courses, we followed-up by examining the impact data associated with the preliminary cut scores generated from the contrasting groups standard setting methodology. These distributions are shown in the following figures.

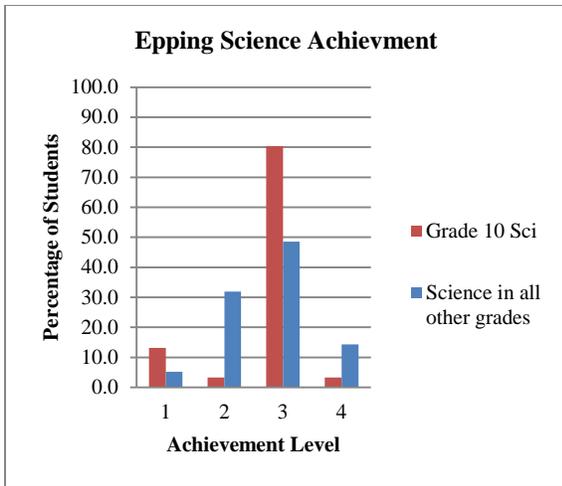


Figure 5. Epping G10 Science Comparison

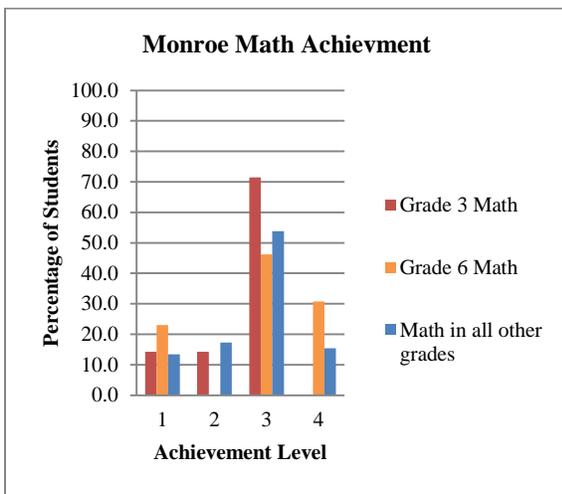


Figure 6. Monroe G3 & G6 Math Comparisons

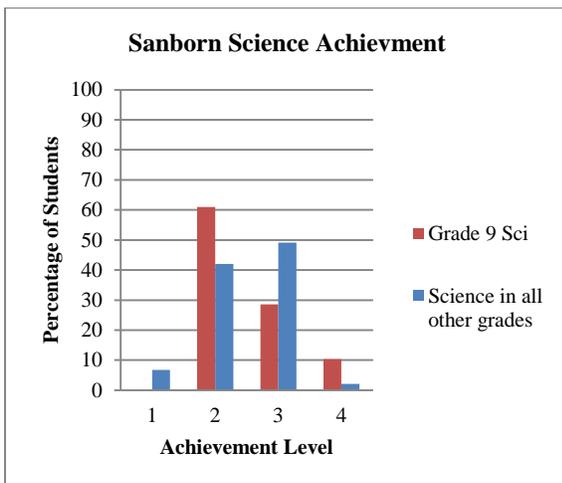


Figure 7. Sanborn G9 Science Comparison

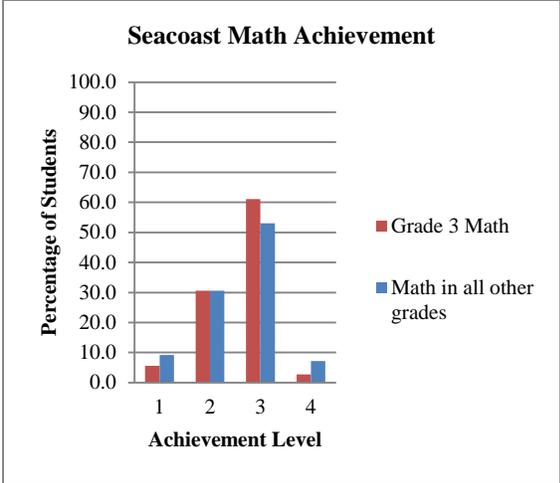


Figure 8. Seacoast G3 Math Comparison

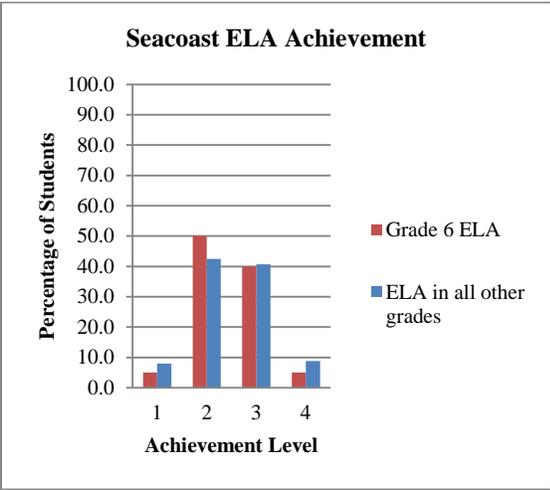


Figure 9. Seacoast G6 ELA Achievement

To better understand the differences in patterns of achievement we tested whether the percentage of students proficient in the grade level and subject of interest, is significantly different than the percentage of students who are proficient in that subject area in the other grades in that district.

Table 16.
Independent Samples t-tests for %Proficient

Course	Difference in %Prof	t	df	Sig.
Epping Grade 10 Science	12.70%	-3.665	239.131	.000
Monroe Grade 3 Math	-5.80%	0.288	37.000	.775
Monroe Grade 6 Math	-11.30%	.730	43	.469
Sanborn Grade 9 Science	0.23%	-.084	3414	.933
Seacoast Grade 3 Math	7.80%	-1.558	99.777	.122
Seacoast Grade 3 ELA	-1.57%	.331	8068	.741

Of all the tests, only the test for Epping grade 10 Science was statistically significant and in the expected direction. Combined with the information generated from the consensus scoring analysis, this evidence suggests that the teachers in Grade 10 Science for Epping scored systematically more leniently than the consensus scorers and their science teacher colleagues in other grade levels in Epping. Therefore, a cut score adjustment to the level 3 cut was made using an equipercenile standard setting technique using Grade 9 science achievement at the reference distribution. The Table 17 and Figure 10 show the cut score adjustments and resulting achievement level distribution for Grade 10 Science in Epping. No other cut score adjustments were made since Epping Grade 10 Science was the only course with multiple sources of evidence pointing to incomparability (i.e., flagged discrepancy and significantly different distribution of achievement).

Table 17.
Epping Grade 10 Science Cut Score Adjustments

	Level 2	Level 3	Level 4
Original Cut Scores	62.26	71.87	93.30
Adjusted Cut Scores	62.26	81.80	93.30

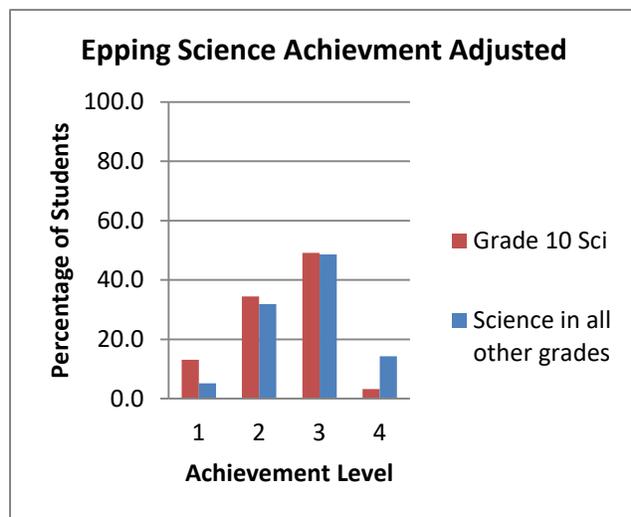


Figure 10. Resulting G10 Science Distribution Comparison

As we have noted throughout this report and other communications, PACE is built on a reciprocal accountability framework. As such, instead of adjusting district performance standards in isolation, PACE leadership works with district leadership to implement improved practices based on observed results. As an example, the Rochester School District scoring was generally more lenient than other districts last year, particularly at the elementary school level. Rochester used these analyses to focus professional development on improved scoring processes, which contributed to much better results for Rochester this year.

OK, I understand how this analysis evaluates the effect on the common task, but I am not sure I understand how this will affect the locally developed tasks which also may not have a consistent scoring.

Calibration Analyses 2017

The PACE innovative assessment system uses common performance tasks across districts to evaluate the degree of comparability in local scoring. These analyses rest on the assumption that patterns in scoring for the common tasks are representative of district's relative stringency and leniency in scoring of the local performance tasks and assessments. The calibration audit is intended to uncover differences in scoring among districts that can be used to support decision-making about any adjustments to cut scores that may be need due systematic cross-district differences. The scores of student work samples on PACE performance tasks that result from this audit serves as the "calibration weights" so that more generalized inferences about relative leniency or stringency of district scoring practices can be made. On August 14, 2017, teachers and leaders from the nine PACE Tier 1 districts and five PACE Tier 2 districts participated in the calibration audit.

The consensus scoring method involves pairing teachers together, each representing different districts, to score student work samples. The student work samples were gathered for each of the PACE common performance tasks from the nine districts participating in the 2016-17 PACE pilot. Both judges within each pair were asked to individually score their assigned samples of student work. Working through the samples one at a time, the judges discussed their individual scores and then agreed on a "consensus score". In the one case where consensus could not be reached, an expert scorer (who did not have affiliation with any particular district) decided on the appropriate consensus score. The purpose of collecting consensus score data is to get the best estimate of the "true score" for use as a "calibration weight." These consensus scores are then used in follow-up analyses to detect any systematic, cross-district differences in the stringency of standards used for scoring.

Students with scores that were out of range were removed from the cut score analyses for that grade and subject (N=38). Consensus scores were matched with the local, teacher-given task scores on Student ID, district, grade, subject, and assessment name. This matching resulted in 1,622 total students with both consensus scores and local scores for the common task work. The distribution of these students across grades, subjects, and district is provided in Table 11 on the next page. One of the first things to observe in the table is the number of cells with very few students, due in large part that many of these schools and districts were very small, rural districts. This causes challenges for our ability to evaluate comparability and to establish cut scores with any degree of precision.

Table 11. Number of Matched Students by Grade, Subject, and District

Grade	Subject	Bethlehem	Concord	Epping	Lafayette	Landaff	Lisbon	Monroe	Pittsfield	Profile	Rochester	Sanborn	Seacoast	Souhegan	Total
3	Math	4	18	12	7	3	6	2	18		17	17	14		118
4	ELA	6	20	18				5			17	11	18		95
	Sci		17	6				2	19		18	10			72
5	ELA	9	18	18	6		6	10			18	18	16		119
	Math	6	17	15	6		2	10	29		16	15	18		134
6	ELA	5	20	8	6		5		16		19	17	11		107
	Math		18	18	5		3	9	14		17	17			101
7	ELA		19	17			9	12	18	6	18	12	20		131
	Math		8	16			8	11	6	9	13	19	18		108
8	Sci		13	13			6	6	13	9	16	16	18		110
9	ELA		15	18			9			9	16	14		18	99
	Math		19	18			9		7	5	12	10			80
	Sci			12			4			8	9	9		17	59
10	ELA		17	17			9			7	12	18		18	98
	Math		17	18			9		16	4	12	2		15	93
	Sci			18					18	10	18	16		18	98
Total		30	236	242	30	3	85	67	174	67	248	221	133	86	1622

To detect any systematic discrepancies in the relatively leniency and stringency of district scoring, we calculated a mean deviation index. This index is the mean difference between the consensus score and teacher score across all student work samples for each district as calculated by the following, for District k:

$$Deviation_k = \frac{\sum_i^n (teacher_i - consensus_i)}{n_k}$$

Using this index, a negative mean deviation would indicate systematic underestimation of student scores by classroom teachers (i.e., district stringency), and positive mean deviation scores would indicate systematic overestimation of student scores by classroom teachers (i.e., district leniency). The values of the deviation metric are on the scale of the rubric points. Table 12 below shows the average observed deviation by district.

Table 12. Average Deviation by District			
District	Deviation	N	SD
Bethlehem	0.4783	30	0.50
Concord	0.2726	236	0.59
Epping	0.3402	242	0.66
Lafayette	0.3517	30	0.69
Landaff	-0.1333	3	0.23
Lisbon	0.4935	85	0.64
Monroe	0.4435	67	0.62
Pittsfield	0.3071	174	0.69
Profile	0.6565	67	0.68
Rochester	0.3442	248	0.64
Sanborn	0.2875	221	0.61
Seacoast	0.1193	133	0.57
Souhegan	-0.0314	86	0.60

Positive scores indicate a systematic overestimation of common task scores by the classroom teachers. If they are all high it is not necessarily problematic from a comparability perspective, we are just looking for differences among the districts in average deviation. Profile’s average deviation score stands out as being particularly high. Post-hoc analyses with a Bonferroni correction revealed that Profile’s marginal deviations are significantly different from seven other districts (Concord, Epping, Pittsfield, Rochester, Sanborn, Seacoast, and Souhegan). Based on this, Profile will receive further review.

A three-factor analysis of variance reveals a significant 3-way interaction for district, by grade, by subject combinations. This means we cannot justify any unilateral adjustments to any one districts’ cut scores across the board. Instead, more nuanced decisions must be made based on follow-up analyses.

Table 13. ANOVA – District by Grade by Subject				
Source	df	F	Sig.	Partial Eta Squared
District	12	7.943	0.000	0.060
Grade	7	3.995	0.000	0.018
Subject	2	14.170	0.000	0.019
District * Grade	52	3.717	0.000	0.115
District * Subject	19	3.164	0.000	0.039
Grade * Subject	6	3.735	0.001	0.015
District * Grade * Subject	28	4.088	0.000	0.071

R Squared = .306 (Adjusted R Squared = .248)

The plots generated by this analysis of variances are provided for each subject area in Figures 1-3 on the next pages.

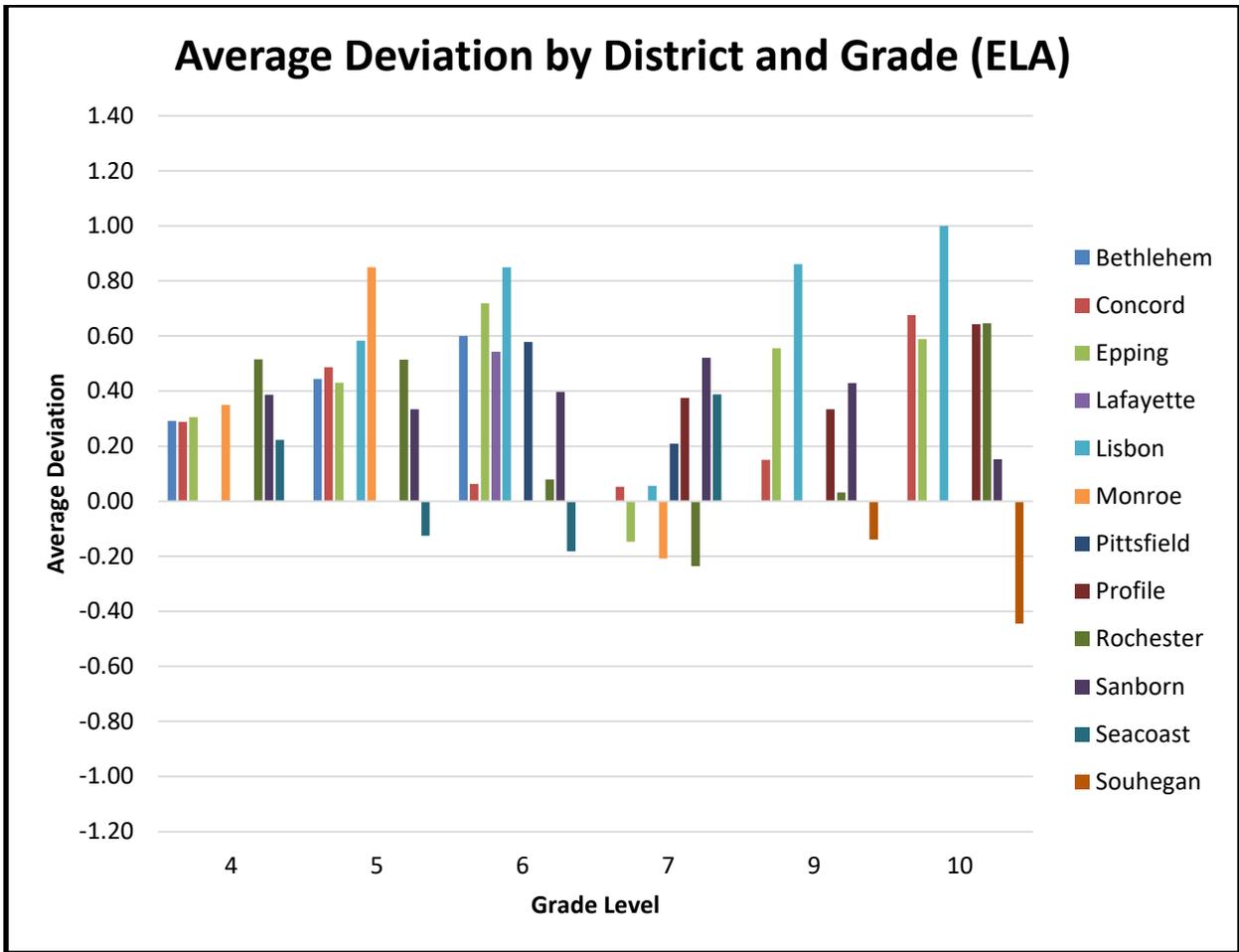


Figure 1. Marginal Mean Deviations by District and Grade for ELA

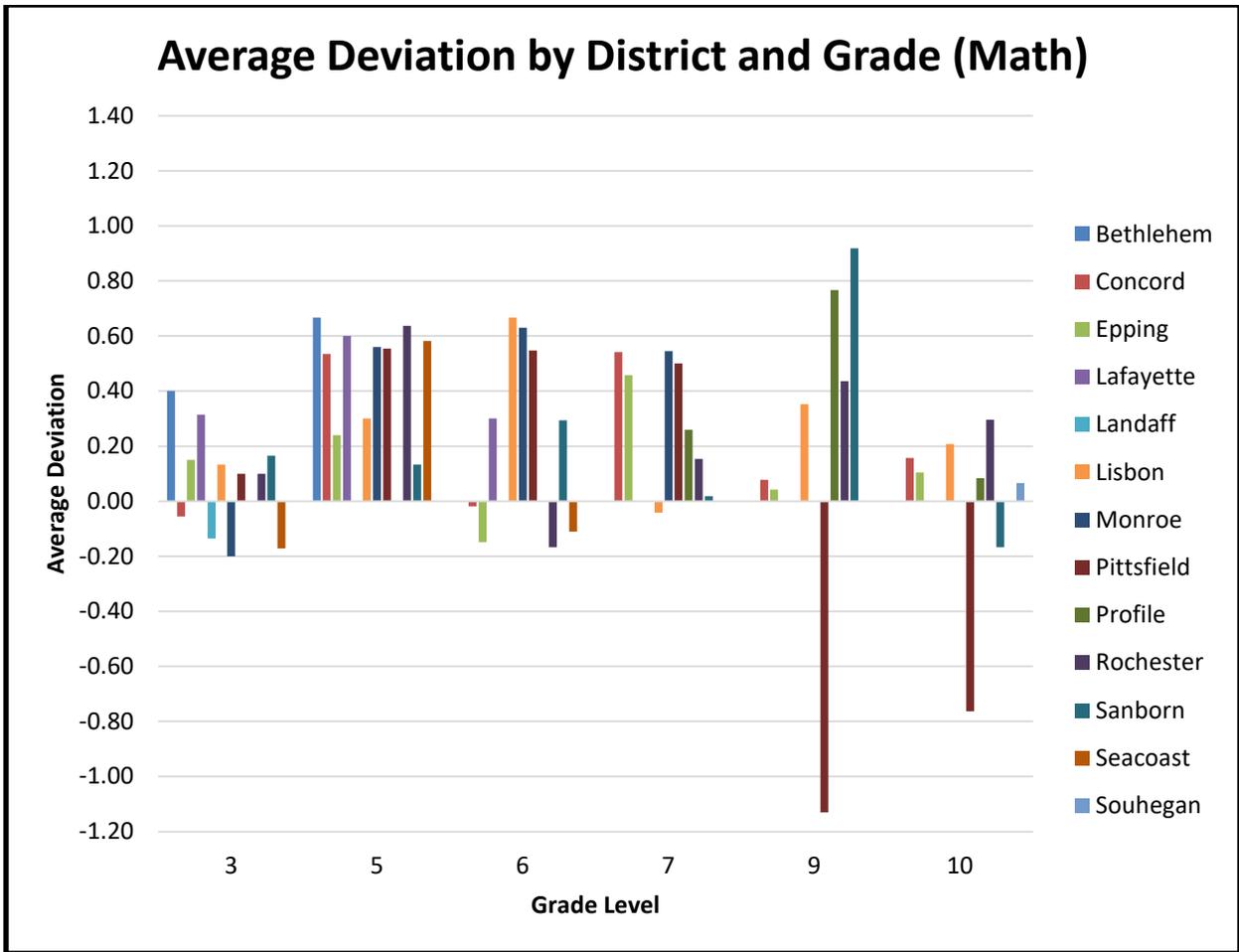


Figure 2. Marginal Mean Deviations by District and Grade for Math

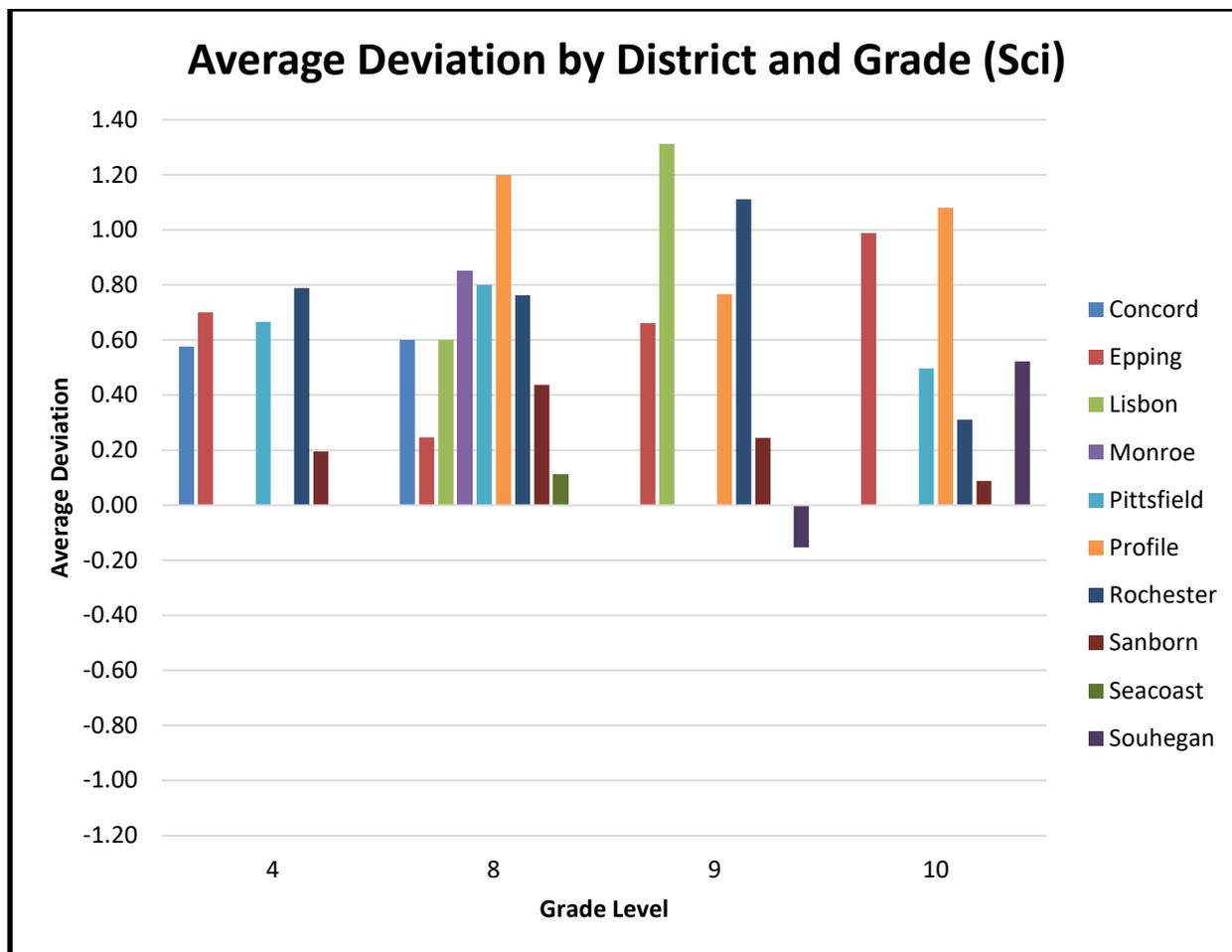


Figure 3. Marginal Mean Deviations by District and Grade for Science

Overall, it seems that the math teachers and consensus scorers are more consistent than the teachers and scorers in ELA and science with two exceptions. The two exceptions are Pittsfield Grade 9 and 10 math both of which stand apart from the rest as having strong, negative deviations (-1.13 and -0.76, respectively). This may indicate stringent scoring on the part of the Grade 9 and 10 Math teachers in Pittsfield. There were no Teacher Judgment Surveys submitted for Grade 10 Math from Pittsfield so annual determinations will not be reported in that grade. To follow-up further, Grade 9 Math in Pittsfield will receive additional review.

Souhegan Grade 10 ELA stands apart from the rest as having a negative deviation score (-0.44). This may indicate stringent scoring on the part of the Grade 10 ELA teacher in Souhegan. To follow-up further, Souhegan Grade 10 ELA will receive additional review.

Profile Grade 8 and Lisbon Grade 9 science stand apart from the rest as having strong, positive deviation scores (1.20 and 1.31, respectively). This may indicate lenient scoring on the part of Profile's Grade 8 and Lisbon's Grade 9 science teachers. However, the high fluctuation in Profile and Lisbon in these grades/subject areas may be an artifact of the particularly small sample size for these schools (N=9 and N=4, respectively). To follow-up further, these two grades (Profile Grade 8 and Lisbon Grade 9 Science) will receive additional review.

To further investigate the earlier findings with Profile we looked at the grade level and subject combinations where Profile’s deviation is significantly different than the other districts’. Using complex contrast post-hoc analyses, with no type-1 error correction, we analyzed the mean differences in deviation for Profile as compared with all other districts for each subject and grade. The equality of variance assumption was met for all combinations except Grade 10 math and Grade 10 science for which the appropriate *t*-value corrections were made.

Table 14. Follow-Up Comparisons for Profile					
Subject	Grade	Mean Difference	t	df	Sig.
ELA	7	0.29500	1.322	129	0.189
	9	0.06667	0.297	97	0.767
	10	0.28022	0.935	96	0.352
Math	7	0.05051	0.234	106	0.815
	9	0.60822	1.844	78	0.069
	10	0.10112	0.137	3.048	0.899
Sci	8	0.69208	3.654	108	0.000
	9	0.31993	1.106	57	0.273
	10	0.58966	6.120	32.848	0.000

Table 14 shows that Grade 8 and Grade 10 Science in Profile show evidence of significant overestimation of scores ($p < .05$), which have the following deviation averages respectively, 1.20 and 1.08. These deviation scores can provide benchmarks within science to flag high deviation averages. Using these scores as the flagging criteria for identifying other high scores, the following district by grade by subject combinations are identified for further review: Lisbon and Rochester Grade 9 Science.

Table 15 below includes all the district, grade, and subject areas noted for further review in this document. It is organized alphabetically by district name then grade level and subject area. Cells highlighted in orange were adjusted because they were either out of range or not estimated.

Table 15. Flagged Deviation Scores with Cut Scores							
District	Grade	Subject	Average Rubric Deviation	Competency Score Scale Range	Cut Scores		
					Level 2	Level 3	Level 4
Lisbon	9	Science	1.31	1.00-4.00	2.06	2.78	3.32
Pittsfield	9	Math	-1.13	1.00-4.00	2.73	3.58	4.00
Profile	8	Science	1.20	0-100	68.65	85.21	95.02
Profile	10	Science	1.08	1.00-4.00	1.50	2.97	3.51
Rochester	9	Science	1.11	1.00-4.00	2.34	3.63	4.00
Souhegan	10	ELA	-0.44	1.00-4.00	1.78	2.78	3.61

With each of the flagged courses in Table 15, we followed-up by examining the impact data associated with the preliminary cut scores generated from the contrasting groups standard setting methodology. These distributions are shown in Figures 4-8.

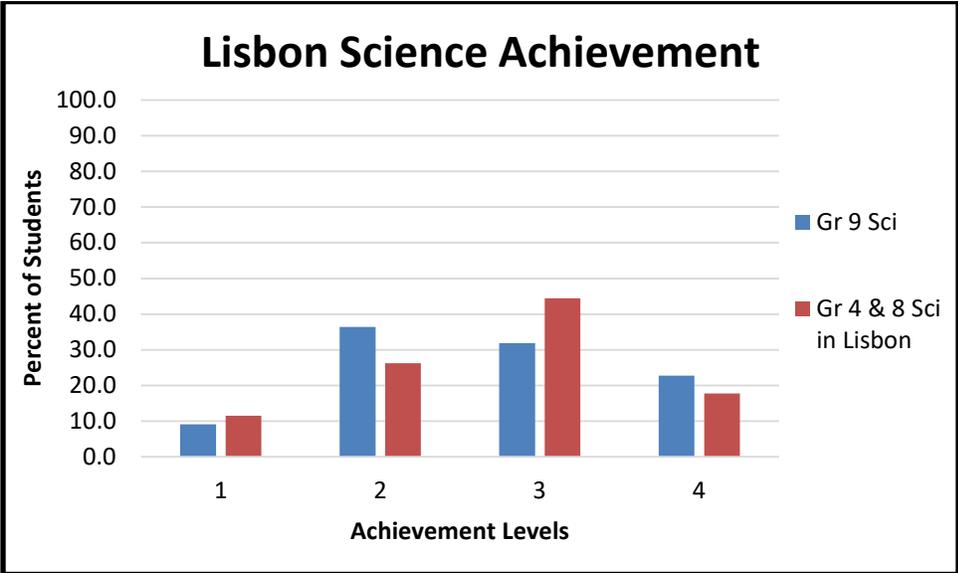


Figure 4. Lisbon Grade 9 Science Comparison

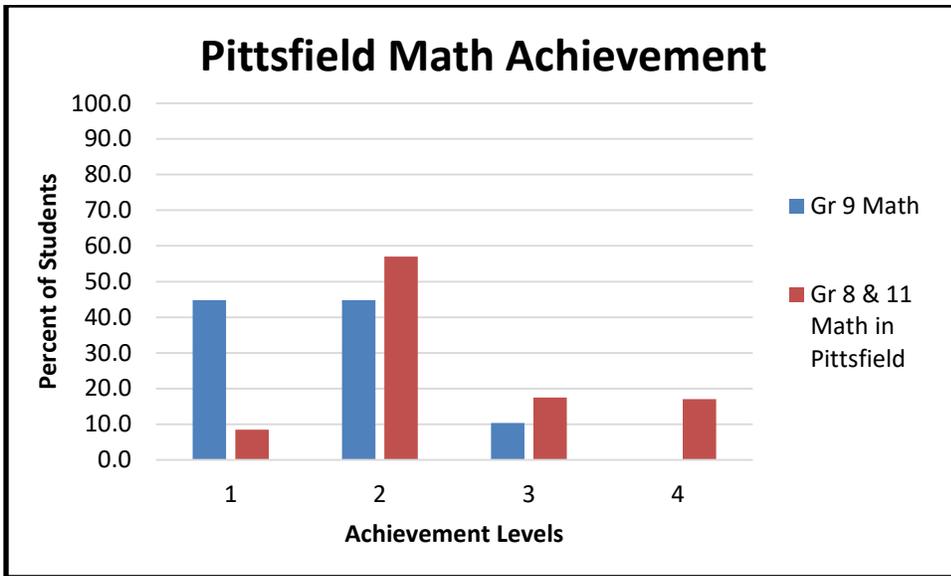


Figure 5. Pittsfield Grade 9 Math Comparison

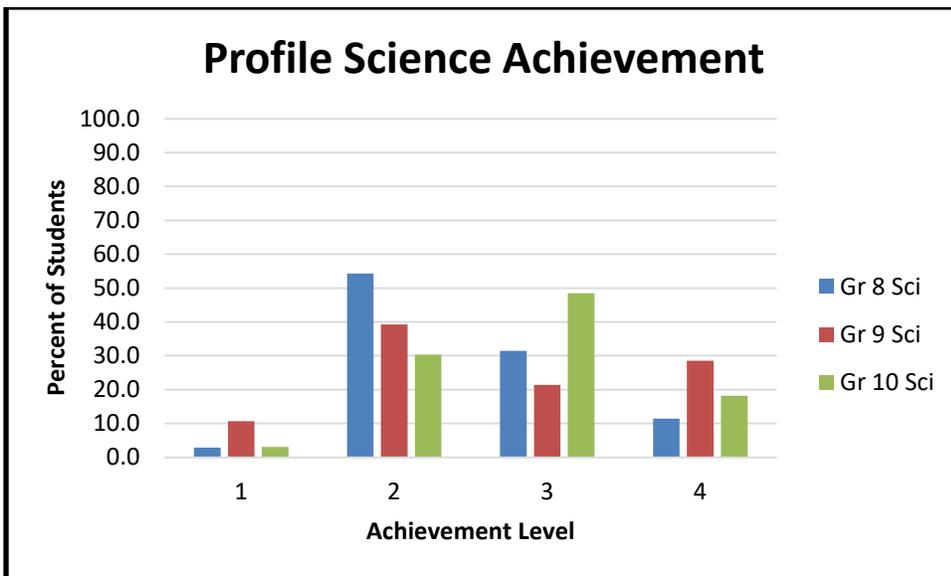


Figure 6. Profile Grade 8 & 10 Science Comparison

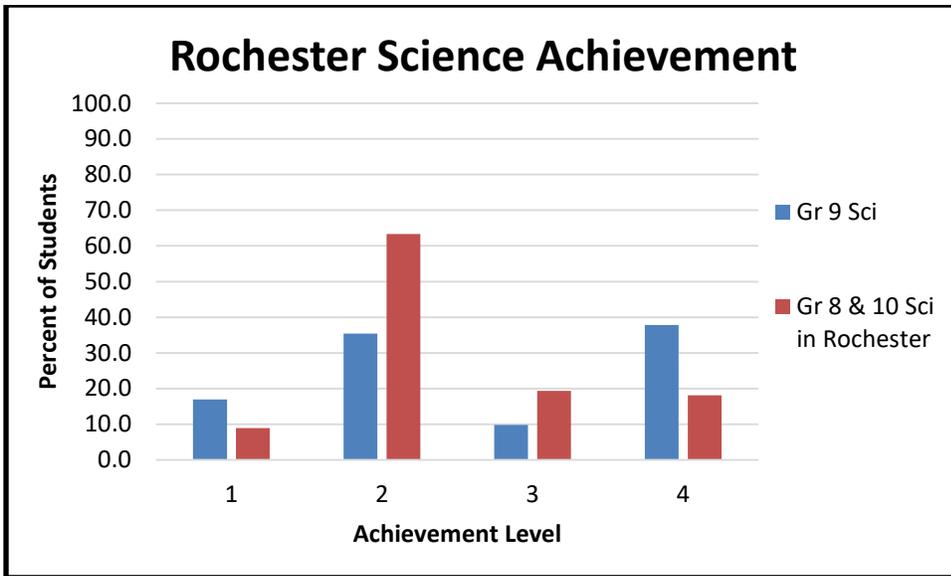


Figure 7. Rochester Grade 9 Science Comparison

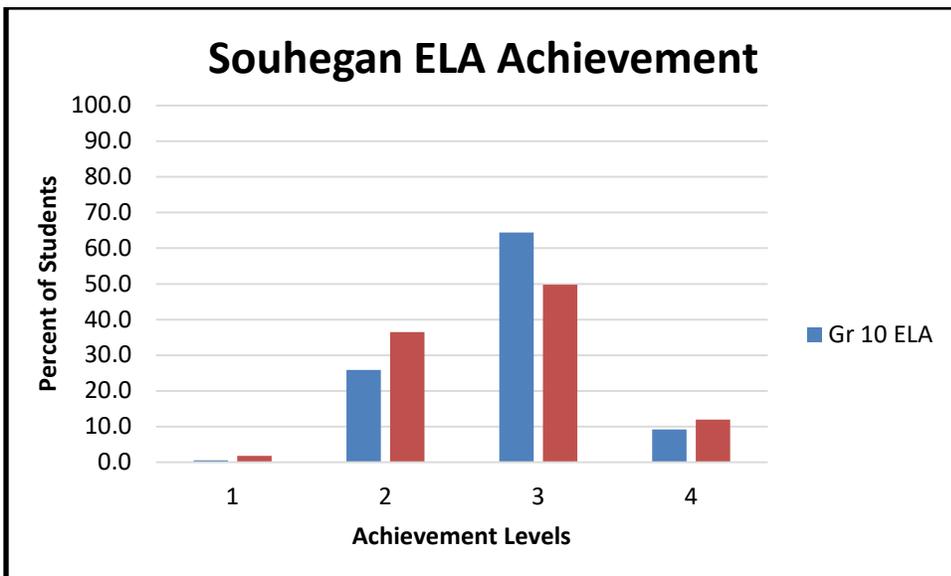


Figure 8. Souhegan Grade 10 ELA Comparison

To better understand the differences in patterns of achievement we tested whether the percentage of students proficient or above in the grade level and subject of interest, is significantly different than the percentage of students who are proficient or above in that subject area in the other grades in that district.

Reference Group	Comparison Group	Difference in %Prof or above	t	df	Sig.
Lisbon Grade 9 Science	Lisbon Grade 4 & 8 Science	-7.75%	-0.631	81	0.530
Pittsfield Grade 9 Math	Pittsfield Gr 8 & 11 Math	-23.94%	-2.635	50.916	0.011
Profile Grade 8 Science	Profile Grade 9 Science	-7.14%	-0.558	61	0.579
Profile Grade 10 Science	Profile Grade 9 Science	16.67%	1.316	59	0.193
Rochester Grade 9 Science	Rochester Grade 8 & 10 Science	19.64%	5.921	766.224	0.000
Souhegan Grade 10 ELA	Souhegan Grade 9 & 11 ELA	10.97%	2.672	352.828	0.008

Four *t*-tests in Table 16 were statistically significant. A few districts did not have differences in percent proficient or above that were in the expected direction. For example, based on the consensus scoring analysis where Lisbon Grade 9 Science and Profile Grade 8 Science had positive deviation scores, we would expect there to be a higher percent proficient or above since the teachers scored more leniently than the consensus scorers, on average. However, Table 6 shows a lower percent proficient in both Lisbon Grade 9 Science and Profile Grade 8 Science. Similarly, based upon the negative deviation score from the consensus scoring analysis for Souhegan Grade 10 ELA, we would expect a lower percent proficient in this grade and subject area. Instead, Souhegan Grade 10 ELA has about 11% more students proficient or above than Grade 9 and 11 ELA, on average. No cut score adjustments will be made in these grade and subject areas because there is not enough evidence to suggest a systematic pattern as the results of the consensus scoring analysis and the percent proficient or above analysis are not in the same direction.

This leaves Pittsfield Grade 9 Math and Rochester Grade 9 Science with *t*-test results that are significant in Table 16 and results in the expected direction.

Pittsfield Grade 9 Math had lower percent proficient or above than in comparison to the average of Grades 8 & 11 Math if Pittsfield. Combined with the information generated from the consensus scoring analysis, this evidence suggests that the teachers in Pittsfield Grade 9 Math scored systematically more stringently than the consensus scorers and their math teacher colleagues in other grade levels in Pittsfield. Therefore, cut score adjustments to the Level 2, 3, and 4 Math Grade 9 cuts were made using an equipercentile standard setting technique with the average of Pittsfield Grades 8 & 11 as the reference distributions. Table 17 and Figure 9 show the cut score adjustments and resulting achievement level distribution for Grade 9 Math in Pittsfield.

	Level 2	Level 3	Level 4
Original Cut Scores	2.73	3.58	5.16
Adjusted Cut Scores	1.25	3.00	3.50

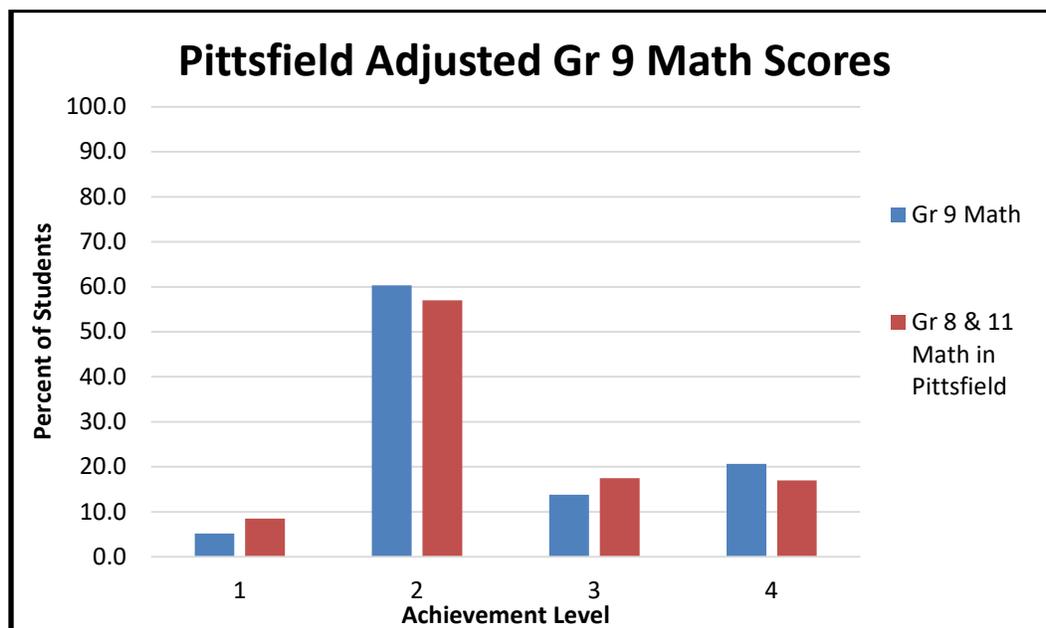


Figure 9. Resulting Grade 9 Pittsfield Math Distribution Comparison

Rochester Grade 9 Science had higher percent proficient or above than in the other grades in science in Rochester. Combined with the information generated from the consensus scoring analysis, this evidence suggests that the teachers in Rochester Grade 9 Science scored systematically more leniently than the consensus scorers and their science teacher colleagues in other grade levels in Rochester. However, when an equipercentile standard setting procedure was used based on an average of Rochester Grade 8 & 10 Science as the reference distribution, the estimated cut scores were unusual. The Level 2 cut would have been 1.33 and the Level 3 and 4 cut would have been 4.0. This resulted since a majority (53.4%) of the students in Rochester Grade 9 Science having end of year competency scores of 4. Therefore, no cut score adjustments in Rochester Grade 9 Science were made except the Level 4 cut score was adjusted to the highest obtainable scale score (HOSS). No other cut score adjustments were made since Pittsfield Grade 9 Math was the only course with multiple sources of evidence pointing to incomparability (i.e., flagged discrepancy and significantly different distribution of achievement).

As we have noted throughout this report and other communications, PACE is built on a reciprocal accountability framework. As such, instead of adjusting district performance standards in isolation, PACE leadership works with district leadership to implement improved practices based on observed results.

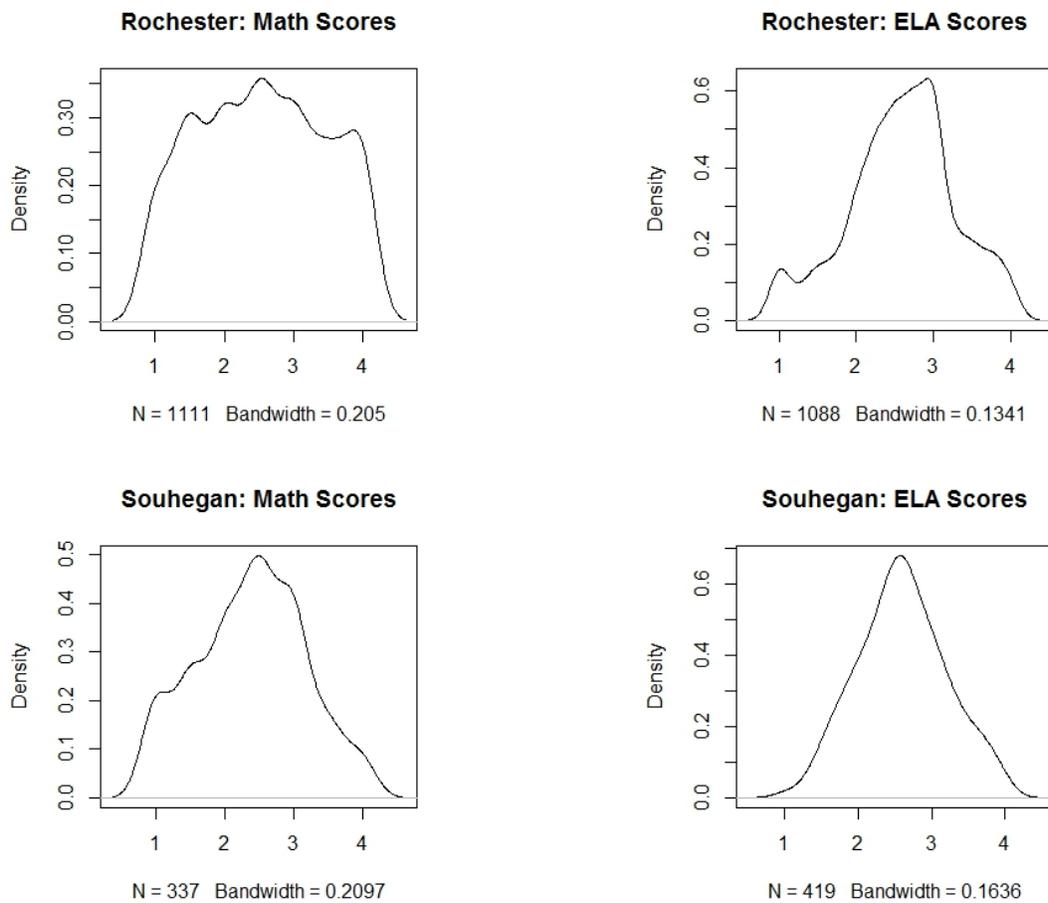
Appendix E: Standard Setting Reports in 2015, 2016, and 2017

NH Pace: Standard Setting Report 2015

Susan Lyons, Ph.D.
Center for Assessment
10-30-15

For the participating PACE districts, student scores in the PACE subject areas and grade levels were calculated by averaging the rubric scores from the submitted competency scores from throughout the course of the year. In any given subject students had a range of between 2 and 16 rubric scores contributing to their PACE average score with an average of 3 scores in math, 6 scores in ELA, and 4 scores in science. Figure 1 below shows the PACE score distributions of average rubric score for each subject area in two PACE districts: Rochester and Souhegan.⁷

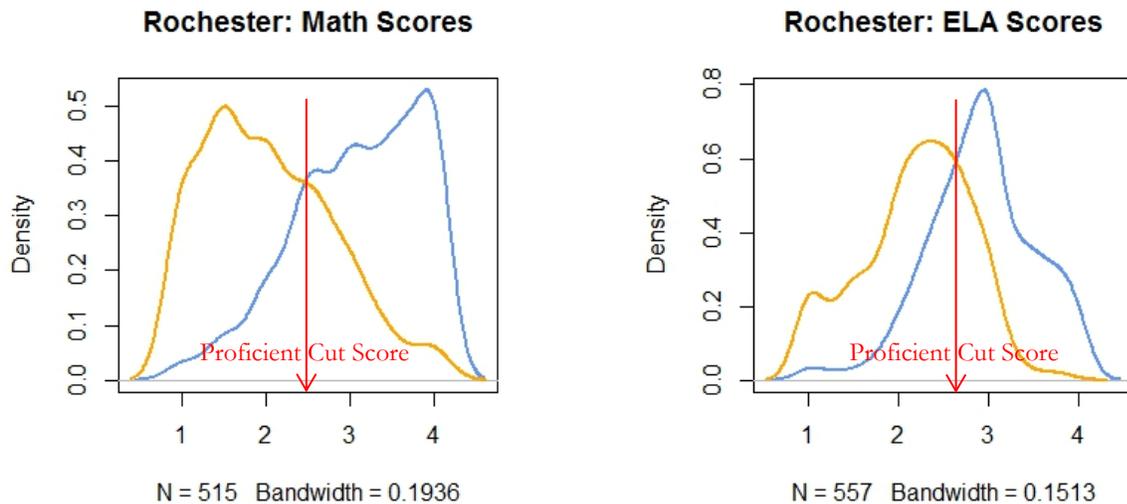
Figure 1
PACE Score Distributions



⁷ Rochester and Souhegan are the two districts analyzed throughout this report because Epping and Sanborn are still in the process of submitting their standard setting data. This report will be updated once those data are complete.

The purpose of the standard setting is to determine where in the score distributions the appropriate “cut points” lie for establishing achievement levels. To establish cut points we used an examinee-centered judgmental method called contrasting groups. This standard setting method involves judgments from panelists about the qualifications of the examinees based on prior knowledge of the examinee. To implement this method for the PACE pilot, we asked PACE teachers to make judgments about which achievement level best described each of their students from the previous year. This process relied on the achievement level descriptors (ALDs) that were written by teachers on August 11, 2015.⁸ The subject and grade level specific ALDs were entered into an online survey where teachers could easily read the descriptions and match their former students to the appropriate achievement level. 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. For example, Figure 2 plots the PACE score distribution for Rochester students classified as proficient by their teachers (i.e., Level 3 and Level 4 shown in blue) along with the PACE score distribution for students classified as non-proficient (i.e., Levels 1 and 2 shown in yellow).

Figure 2
Contrasting Groups Score Distributions



The point on the PACE score continuum where the two distributions intersect is our best estimate of proficiency cut score. There are a few different methods for determining this mid-point, but the most common way is using logistic regression (Cizek, 2007). 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). This kind of logistic regression analysis was run separately for each cut point, Level 2, Level 3, and Level 4, in each district, content area, and grade level. Table 1 on the next page shows

⁸ See Appendix K for an example of the Achievement Level Descriptors used for standard setting

the number of teacher judgments for each district, subject, and grade level that were used to estimate the cut scores. Note: Smarter Balanced (SBAC) was administered in Grade 3 and 8 for ELA, Grade 4 and 8 for math, and grades 3 and 5-7 for science; therefore, no ALD classification judgments were requested for the SBAC grades.

Table 1
Number of judgements for cut score estimation

	Grade	ELA	Math	Science
Rochester	3	0	330	0
	4	234	0	235
	5	280	241	0
	6	221	220	0
	7	309	237	0
	8	0	0	320
Souhegan	9	224	210	134
	10	211	127	212

The results of the contrasting groups standard setting analyses are shown in Tables 2-4 below. Those cells highlighted in orange were generated using a modified methodology for estimating the cut scores. Across the board, the standard setting resulted in Level 4 cut scores that were very stringent and often times the estimated cut scores were above to high obtainable PACE score. Because we fundamentally believe that there are students in the PACE districts that have attained achievement as Level 4, we adjusted the Level 4 cut scores to be the mid-way point between the Level 3 cut score and a PACE score of 4.0. Similarly, due to restriction of range issues, Level 2 cut scores were not estimable for grade 9 math and science in Souhegan and grade 4 science in Rochester. A similar procedure was used where the estimated cut score became the midpoint between 1.0 and the Level 3 cut.

Table 2
Math Cut scores

	Grade	Level 2	Level 3	Level 4
Rochester	3	1.68	2.95	3.48
	5	1.46	2.85	3.42
	6	1.07	2.35	3.18
	7	1.89	3.00	3.50
Souhegan	9	1.28	1.57	2.78
	10	1.62	2.61	3.31

Table 3
ELA Cut scores

	Grade	Level 2	Level 3	Level 4
Rochester	4	1.66	2.78	3.39
	5	1.88	2.98	3.49
	6	1.65	2.56	3.28
	7	1.23	2.51	3.25
Souhegan	9	1.68	2.57	3.29
	10	1.07	2.44	3.22

Table 4
Science Cut scores

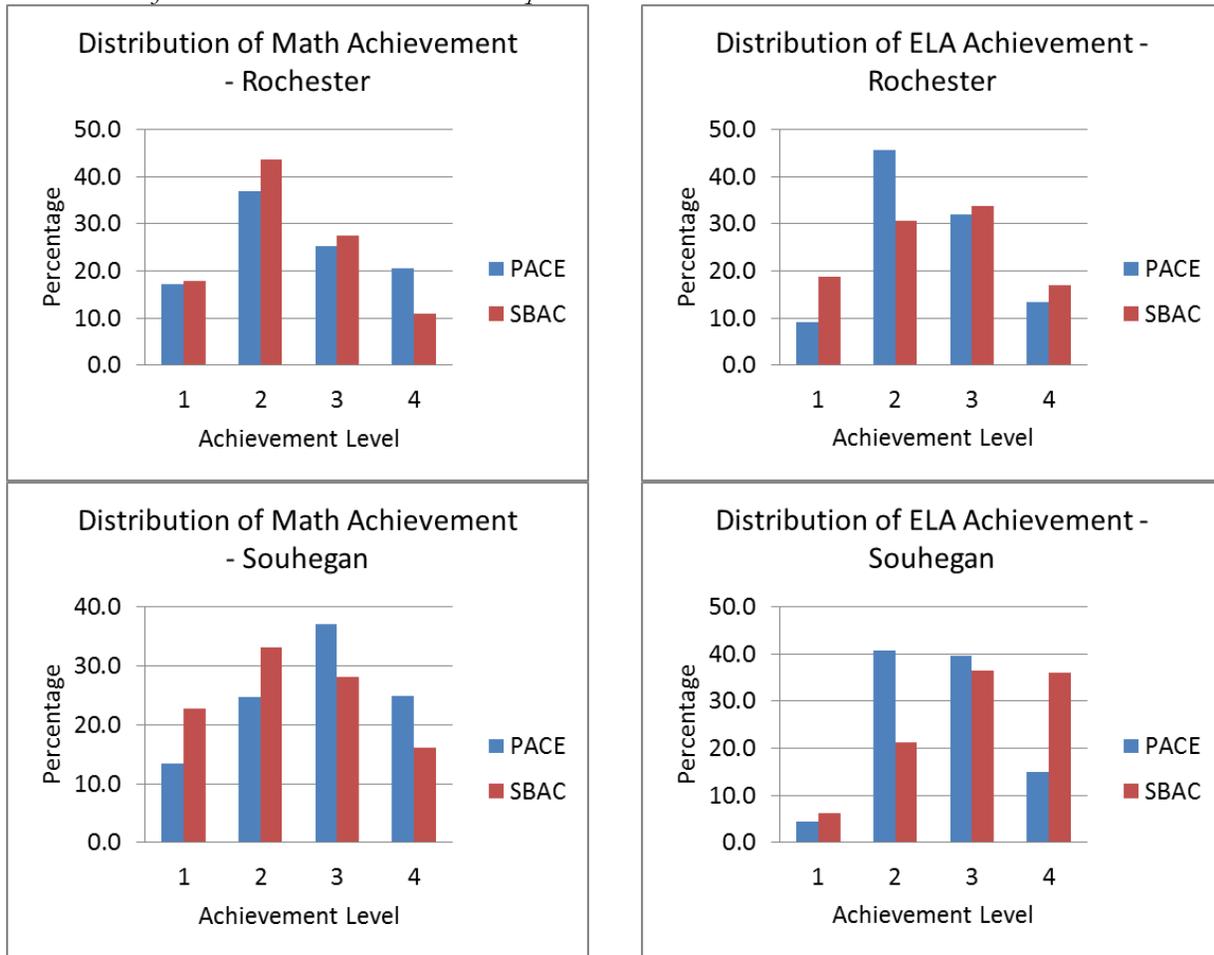
	Grade	Level 2	Level 3	Level 4
Rochester	4	1.53	2.06	3.03
	8	1.62	2.97	3.49
Souhegan	9	1.40	1.81	2.90
	10	1.09	2.04	3.02

As a cross-validation exercise for the resulting standards, the frequency distributions of for the PACE achievement levels in each district were compared with the district Smarter Balanced results. Table 5 below shows the percentage of students meeting proficiency and Figure 3 shows the distribution of achievement across all four levels. Examination of the results reveals that the PACE cut scores maintain the proportional relationship of percent proficient across districts (i.e., a higher percentage of students reach proficiency in Souhegan in than in Rochester). Additionally, the results show that the PACE standards may be more lenient in math than SBAC, but more stringent than SBAC in ELA.

Table 5
Proficiency Rates across Assessment Types

	%Proficient Math		%Proficient ELA		%Proficient Science
	PACE	SBAC	PACE	SBAC	PACE
Rochester	45.7	38.3	45.2	50.8	45.6
Souhegan	62.0	44.2	54.6	72.5	70.1

Figure 3
Distribution of PACE Achievement Levels Compared to SBAC



Examination of Figure 3 shows that in general, the distribution of PACE achievement levels does not deviate greatly from the distribution of Smarter Balanced achievement levels. Only for ELA in Souhegan do we see a contrast in the shapes of the distributions. A markedly high percentage of students in Souhegan achieved Level 4 in ELA for Smarter Balanced. This high achievement is not reflected in the PACE Achievement level determinations; instead, the shape of the PACE achievement level distribution for ELA in Souhegan more closely mirrors state averages on SBAC.⁹

References

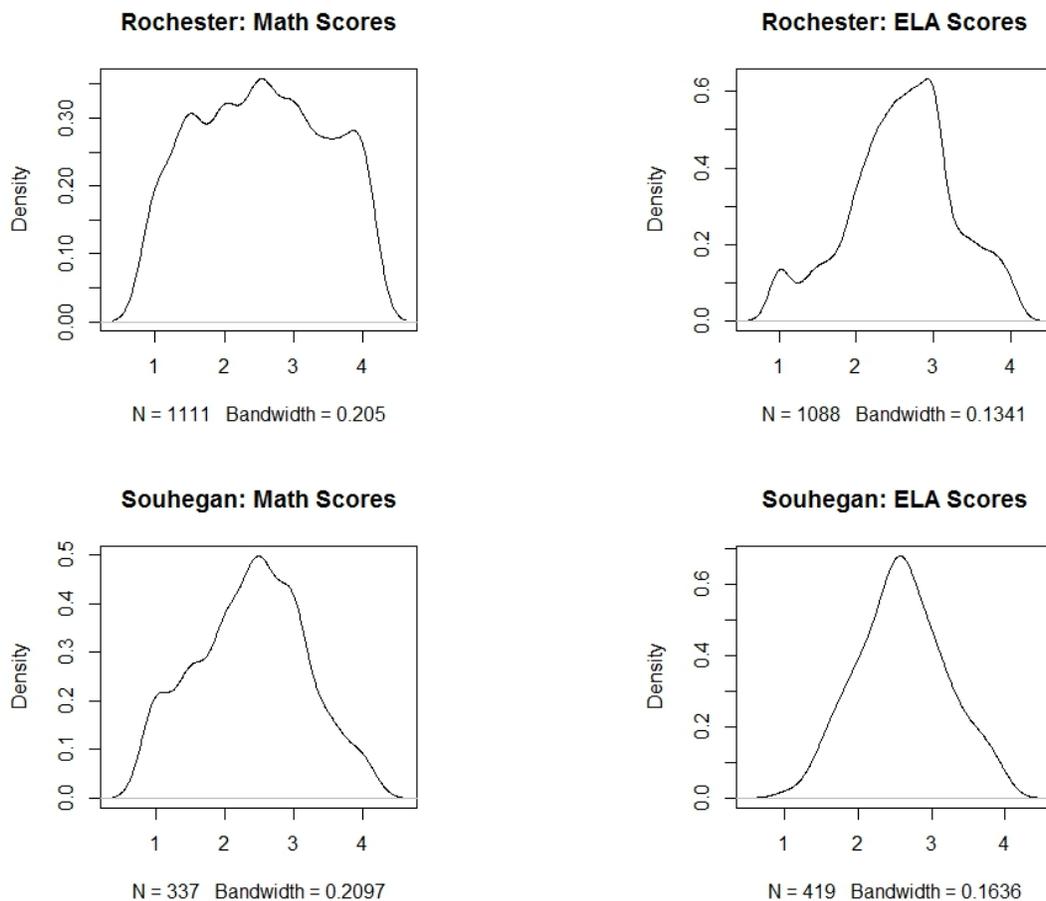
Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications Ltd.

⁹ See the results from the Cross-district Comparability Report.

Standard Setting Report 2015

For the participating PACE districts, student scores in the PACE subject areas and grade levels were calculated by averaging the rubric scores from the submitted competency scores from throughout the course of the year. In any given subject students had a range of between 2 and 16 rubric scores contributing to their PACE average score with an average of 3 scores in math, 6 scores in ELA, and 4 scores in science. Figure 1 below shows the PACE score distributions of average rubric score for each subject area in two PACE districts: Rochester and Souhegan.¹⁰

Figure 1
PACE Score Distributions

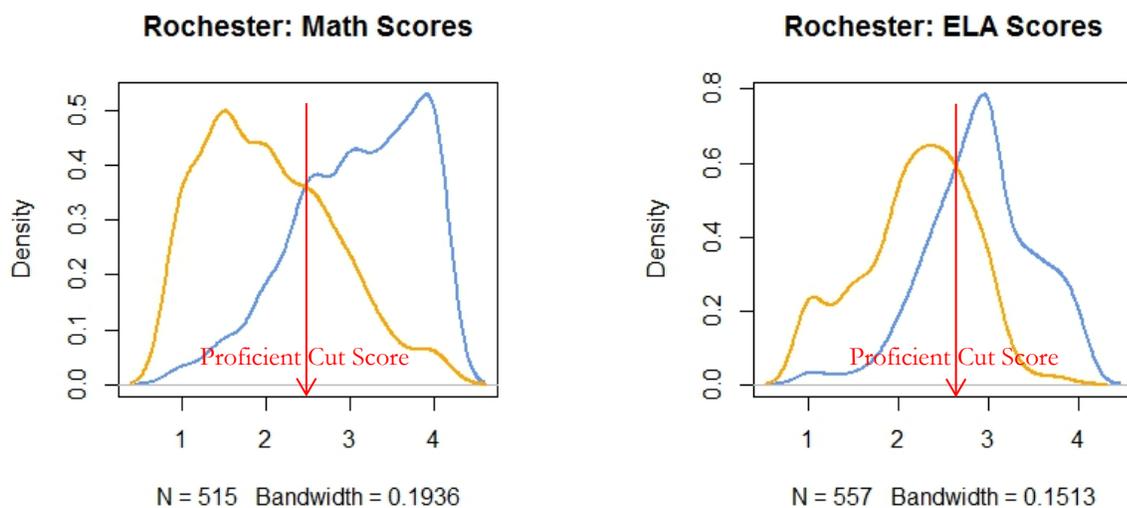


The purpose of the standard setting is to determine where in the score distributions the appropriate “cut points” lie for establishing achievement levels. To establish cut points we used an examinee-centered judgmental method called contrasting groups. This standard setting method involves judgments from panelists about the qualifications of the examinees based on prior knowledge of the examinee. To implement this method for the PACE pilot, we asked PACE teachers to make

¹⁰ Rochester and Souhegan are the two districts analyzed throughout this report because Epping and Sanborn are still in the process of submitting their standard setting data. This report will be updated once those data are complete.

judgments about which achievement level best described each of their students from the previous year. This process relied on the achievement level descriptors (ALDs) that were written by teachers on August 11, 2015.¹¹ The subject and grade level specific ALDs were entered into an online survey where teachers could easily read the descriptions and match their former students to the appropriate achievement level. 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. For example, Figure 2 plots the PACE score distribution for Rochester students classified as proficient by their teachers (i.e., Level 3 and Level 4 shown in blue) along with the PACE score distribution for students classified as non-proficient (i.e., Levels 1 and 2 shown in yellow).

Figure 2
Contrasting Groups Score Distributions



The point on the PACE score continuum where the two distributions intersect is our best estimate of proficiency cut score. There are a few different methods for determining this mid-point, but the most common way is using logistic regression (Cizek, 2007). 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). This kind of logistic regression analysis was run separately for each cut point, Level 2, Level 3, and Level 4, in each district, content area, and grade level. Table 1 on the next page shows the number of teacher judgments for each district, subject, and grade level that were used to estimate the cut scores. Note: Smarter Balanced (SBAC) was administered in Grade 3 and 8 for ELA, Grade 4 and 8 for math, and grades 3 and 5-7 for science; therefore, no ALD classification judgments were requested for the SBAC grades.

Table 1

¹¹ See Appendix K for an example of the Achievement Level Descriptors used for standard setting

Number of judgements for cut score estimation

	Grade	ELA	Math	Science
Rochester	3	0	330	0
	4	234	0	235
	5	280	241	0
	6	221	220	0
	7	309	237	0
	8	0	0	320
Souhegan	9	224	210	134
	10	211	127	212

The results of the contrasting groups standard setting analyses are shown in Tables 2-4 below. Those cells highlighted in orange were generated using a modified methodology for estimating the cut scores. Across the board, the standard setting resulted in Level 4 cut scores that were very stringent and often times the estimated cut scores were above to high obtainable PACE score. Because we fundamentally believe that there are students in the PACE districts that have attained achievement as Level 4, we adjusted the Level 4 cut scores to be the mid-way point between the Level 3 cut score and a PACE score of 4.0. Similarly, due to restriction of range issues, Level 2 cut scores were not estimable for grade 9 math and science in Souhegan and grade 4 science in Rochester. A similar procedure was used where the estimated cut score became the midpoint between 1.0 and the Level 3 cut.

Table 2
Math Cut scores

	Grade	Level 2	Level 3	Level 4
Rochester	3	1.68	2.95	3.48
	5	1.46	2.85	3.42
	6	1.07	2.35	3.18
	7	1.89	3.00	3.50
Souhegan	9	1.28	1.57	2.78
	10	1.62	2.61	3.31

Table 3
ELA Cut scores

	Grade	Level 2	Level 3	Level 4
Rochester	4	1.66	2.78	3.39
	5	1.88	2.98	3.49
	6	1.65	2.56	3.28
	7	1.23	2.51	3.25
Souhegan	9	1.68	2.57	3.29
	10	1.07	2.44	3.22

Table 4
Science Cut scores

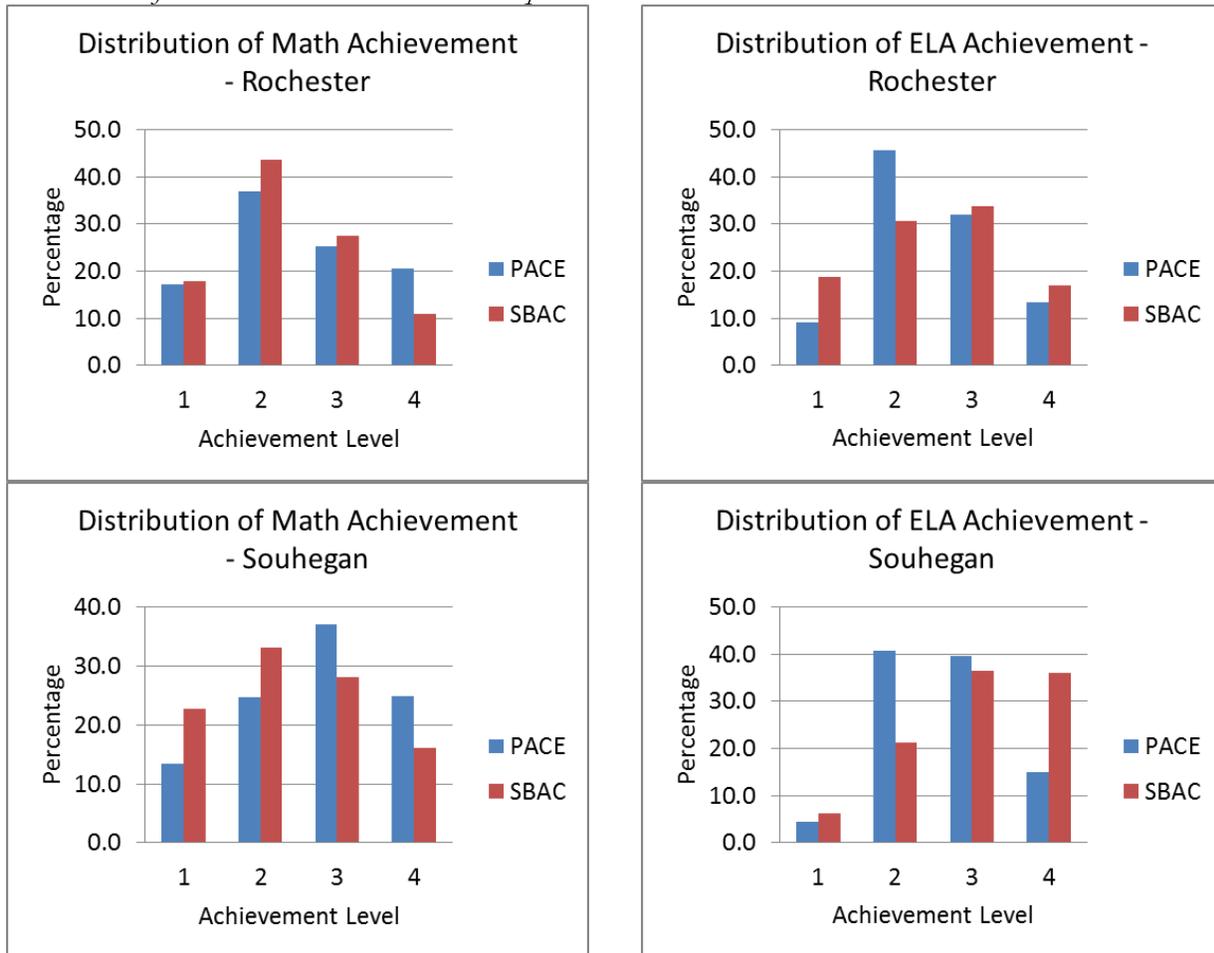
	Grade	Level 2	Level 3	Level 4
Rochester	4	1.53	2.06	3.03
	8	1.62	2.97	3.49
Souhegan	9	1.40	1.81	2.90
	10	1.09	2.04	3.02

As a cross-validation exercise for the resulting standards, the frequency distributions of for the PACE achievement levels in each district were compared with the district Smarter Balanced results. Table 5 below shows the percentage of students meeting proficiency and Figure 3 shows the distribution of achievement across all four levels. Examination of the results reveals that the PACE cut scores maintain the proportional relationship of percent proficient across districts (i.e., a higher percentage of students reach proficiency in Souhegan in than in Rochester). Additionally, the results show that the PACE standards may be more lenient in math than SBAC, but more stringent than SBAC in ELA.

Table 5
Proficiency Rates across Assessment Types

	%Proficient Math		%Proficient ELA		%Proficient Science
	PACE	SBAC	PACE	SBAC	PACE
Rochester	45.7	38.3	45.2	50.8	45.6
Souhegan	62.0	44.2	54.6	72.5	70.1

Figure 3
Distribution of PACE Achievement Levels Compared to SBAC



Examination of Figure 3 shows that in general, the distribution of PACE achievement levels does not deviate greatly from the distribution of Smarter Balanced achievement levels. Only for ELA in Souhegan do we see a contrast in the shapes of the distributions. A markedly high percentage of students in Souhegan achieved Level 4 in ELA for Smarter Balanced. This high achievement is not reflected in the PACE Achievement level determinations; instead, the shape of the PACE achievement level distribution for ELA in Souhegan more closely mirrors state averages on SBAC.¹²

References

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications Ltd.

¹² See the results from the Cross-district Comparability Report.

Standard Setting Report 2016

The Scale

The purpose of the standard setting is to determine where in the competency scales the appropriate cut points lie for establishing achievement levels. For the participating PACE districts, student scores in the PACE subject areas and grade levels were calculated by averaging the competency scores uploaded into Performance Plus by the participating districts. Because the competencies differ across districts and the sample of students within any given district is small, a weighted factor score cannot be computed. For the standard setting dataset, students who had competency scores that fell out of range (e.g., 0.75 on a 1.00-4.00 scale) for a given subject area were removed from that subject area.

Standard Setting Method

To establish cut points we used an examinee-centered judgmental method called contrasting groups. This standard setting method involves using judgments from panelists about the qualifications of the examinees based on prior knowledge of the examinee. To implement this method for the PACE pilot, we asked teachers at the end of the school year to make judgments about which achievement level best described 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 were entered into an online survey where teachers could easily read the descriptions and match their students to the appropriate achievement level. 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 was run separately for each cut point—Level 2, Level 3, and Level 4—in each district, content area, and grade level. The results of the contrasting groups standard setting analyses are shown in in the figure on the next two pages. Those cells highlighted in orange were modified based on flagging and adjusting protocols (described in more detail after the figures).

Concord				
		Level 2	Level 3	Level 4
ELA	4	1.48	2.47	3.46
	5	1.48	2.25	3.37
	6	74.35	86.76	94.50
	7	75.79	88.21	95.76
	9	70.43	85.78	97.27
	10	66.25	86.34	95.82
Math	3	1.85	2.67	3.82
	5	1.60	2.51	3.66
	6	67.45	75.71	85.63
	7	69.82	85.24	96.75
	9	58.86	76.84	92.00
	10	62.04	79.65	92.92
Sci	4	1.19	2.68	3.60
	8	68.72	85.87	100.00
	9	63.59	80.58	94.72
	10	did not participate		

Epping				
		Level 2	Level 3	Level 4
ELA	4	2.34	2.78	3.30
	5	1.95	2.81	3.30
	6	68.04	86.19	93.06
	7	62.46	82.70	94.82
	9	81.46	90.44	97.62
	10	63.81	74.84	100.00
Math	3	1.89	2.55	3.23
	5	1.88	2.80	3.51
	6	76.33	84.41	93.85
	7	56.54	83.61	94.27
	9	64.20	80.96	94.27
	10	65.23	68.56	85.96
Sci	4	1.95	2.71	3.22
	8	71.08	89.78	104.12
	9	71.14	85.13	92.07
	10	62.26	81.80	93.30

Monroe				
		Level 2	Level 3	Level 4
ELA	4	1.76	2.47	2.90
	5	2.26	2.78	3.01
	6	2.19	2.76	3.13
	7	2.25	2.49	3.24
	9			
	10			
Math	3	2.26	2.66	3.05
	5	2.24	2.74	3.13
	6	2.02	2.25	3.00
	7	1.01	2.14	2.88
	9			
	10			
Sci	4	2.28	2.74	3.00
	8	2.00	3.06	3.53
	9			
	10			

Pittsfield				
		Level 2	Level 3	Level 4
ELA	4	2.34	3.11	3.58
	5	1.19	2.99	3.81
	6	2.20	2.87	3.41
	7	2.34	3.06	3.60
	9	1.98	2.97	3.59
	10	1.86	2.71	3.33
Math	3	1.88	2.51	3.13
	5	2.22	2.79	3.50
	6	1.32	2.89	3.58
	7	2.49	3.01	3.84
	9	2.16	3.17	3.85
	10	2.66	3.19	3.67
Sci	4	2.53	3.01	3.61
	8	1.80	2.90	3.78
	9	2.51	3.16	3.58
	10	no ALD judgments		

Rochester				
		Level 2	Level 3	Level 4
ELA	4	2.43	3.22	4.00
	5	2.42	3.19	4.00
	6	2.72	3.69	4.00
	7	2.42	3.49	4.00
	9	2.25	3.61	4.00
	10	2.20	3.44	4.00
Math	3	2.26	2.83	3.59
	5	2.33	3.19	4.00
	6	2.85	3.63	4.00
	7	2.91	3.57	4.00
	9	2.21	3.33	4.00
	10	2.57	3.50	3.75
Sci	4	1.91	3.21	4.00
	8	2.18	3.50	3.99
	9	2.26	3.13	4.00
	10	2.46	3.60	4.00

Sanborn				
		Level 2	Level 3	Level 4
ELA	4	2.11	2.92	3.29
	5	2.37	2.95	3.47
	6	1.49	2.64	3.45
	7	2.14	2.92	3.66
	9	1.40	2.57	3.44
	10	1.66	2.79	3.52
Math	3	1.83	2.86	3.32
	5	1.98	2.89	3.34
	6	1.45	2.61	3.42
	7	1.31	2.96	3.68
	9	1.28	2.84	3.72
	10	1.32	2.71	3.95
Sci	4	1.52	2.70	3.58
	8	1.55	2.35	3.82
	9	1.60	2.79	3.63
	10	2.12	2.92	3.79

Seacoast				
		Level 2	Level 3	Level 4
ELA	4	2.18	2.92	3.93
	5	1.84	2.75	3.32
	6	1.60	2.81	3.35
	7	1.64	2.64	3.82
	9			
	10			
Math	3	1.88	2.75	3.76
	5	1.73	2.71	3.40
	6	2.13	2.73	3.38
	7	2.00	3.00	3.51
	9			
	10			
Sci	4	2.40	2.97	3.64
	8	2.00	2.54	4.00
	9			
	10			

Souhegan				
		Level 2	Level 3	Level 4
ELA	4			
	5			
	6			
	7			
	9	1.53	3.09	3.54
	10	1.40	2.80	3.80
Math	3			
	5			
	6			
	7			
	9	1.20	2.26	3.34
	10	1.53	2.59	3.33
Sci	4			
	8			
	9	1.01	2.66	4.00
	10	2.13	2.97	3.62

Figure 1. 2015-2016 Final Performance Standards

Flagging Rules

Cut scores were flagged for potential adjustment for four reasons.

1. **Non-significant.** In some cases, while the logistic regression was able to generate estimates, the model itself was not able to explain a statistically significant amount of variance in the dependent variable.
2. **Out of range.** In some cases (see many Level 4 cuts in Rochester), teachers tended to rate their students more harshly on the ALD judgment surveys than the competency scale scores reflected. In these cases, the estimated cut score for the highest achievement level would often fall outside the obtainable competency score range.
3. **Not estimated.** In some cases there was insufficient data for the logistic regression model to converge. For example, this would happen if within a given course, the teachers awarded very few Level 1's or Level 4's.
4. **Evidence of Incomparability in Local Scoring.** In one case, there were multiple sources of evidence indicating an issue of incomparability in local scoring. The flagging rules and adjustment procedure are detailed more in the section entitled "Cross-District Comparability Analyses."

Adjustment Protocols

The following adjustment protocols describe the cut score modifications that were made in reaction to the flagged cut scores. These cut score adjustment procedures are sequential in that they were followed in order, if the first modification was not suitable, the second was attempted, if not suitable, the third, and so on.

1. **No adjustment.** In the case of non-significant model estimation, the cut score estimate was within reasonable expectation and remained the most justifiable best guess for where the cut score should be given the data. In those cases, the cut score was left unaltered.
2. **Adjustment to HOSS.** When the cut score fell above of the obtainable competency score range, the cut score for Level 4 was adjusted to the highest obtainable scale score.
3. **Midpoint.** When the cut score was estimated, and fell between two estimated cut scores, the cut score was determined to be the midpoint between the two estimated cut scores.
4. **Equipercntile.** When there are no estimated cut scores on either side of the flagged cut score (e.g., Level 2 or Level 4 cuts), an equipercntile equating procedure was used to estimate the cut score that would closely replicate the distributions of achievement across the performance levels in the same district and subject for the other grade levels with unadjusted cut scores. In the few cases where there were no other grade levels with unadjusted cut scores, the same grade level was used in the other content areas to approximate the distribution of achievement.
5. **Midpoint.** In the few cases where the equipercntile cut score was not estimable (due to small sample sizes or low variability), the midpoint between the LOSS and the Level 3 cut was used to estimate the Level 2 cut, and the midpoint between the HOSS and the Level 3 cut was used to estimate the Level 4 cut.

Cross-District Comparability Analyses

****This section is contained in the body of the Report****

Sending Cut Scores to New Hampshire DOE

The final cut scores were sent to the NH DOE on September 23, 2016. The cut scores were submitted along with directions for calculating the NH PACE Reported Annual Determinations (see Appendix A). Additionally, with the submission of the performance standards to the state, the Center for Assessment included a recommendation that the results of PACE are reported to schools and parents along with a caution that the annual determinations are based on an innovative assessment system for which the validity evidence is still accumulating. We recommended that the score reports additionally specify that there is not yet enough evidence to support the use of these scores for high stakes uses (e.g., school accountability). When making educational decisions, these scores should just be considered as one data point among many including teacher and counselor reports and observations.

Calculating NH PACE Reported Annual Determinations
Business Rules

September 22, 2016

1. Clean the data

- a. It should be first checked that there is at least one score submitted for each student in all PACE subject areas (as determined by the table below). The exception to this rule will be Grade 10 Science in Concord and Pittsfield as we do not have cut scores for those courses. This means that students in Grade 10 Science in Concord and Pittsfield will not have reportable scores.

PACE Administration Chart

	ELA	Math	Science
Grade 3		PACE	
Grade 4	PACE		PACE
Grade 5	PACE	PACE	
Grade 6	PACE	PACE	
Grade 7	PACE	PACE	
Grade 8			PACE
Grade 9	PACE	PACE	PACE
Grade 10	PACE	PACE	PACE

- b. Secondly, ensure that all scores to be included in the score calculation fall within the intended range. If any scores submitted for any student fall outside the range (e.g., 0.75 on a 1.00-4.00 scale, 102 on a 100-point scale) they should be reconciled (e.g., follow up with the district or school to correct the data entry or scoring error). The one exception to this rule is Epping Grade 8 Science. There are a number of students with scores >100.00 on a 100-point scale. These scores are legitimate and should be maintained. These high scores were factored into the standard setting process.
- c. Students with no competency scores are considered non-participants.

2. Calculate mean scores by subject area

- a. All submitted competency scores for each student in each subject area need to be averaged. The resulting student-by-subject averages are henceforth referred to as the student average competency scores.
- b. Note: **For high school, the average competency scores should be computed according the grade level associated with the assessment** rather than the grade level associated with the student. We have a course-based competency system for

high school and therefore the grade level of the student may not reflect the correct competency framework.

- c. Round the average competency scores to two decimal places.

3. Determine the reportable achievement level of each student

- a. The average competency scores that result from step 2 need to be classified into achievement levels using the provided cut scores (attached). The cut score indicates the score that students must meet or exceed to be classified into the corresponding achievement level.
- b. Though the occurrence is rare, some average competency scores will fall outside the expected score range, even with follow-up reconciliation with districts. This is most commonly due to the awarding of zero's for achievement that is so low that the student work consistently does not meet the expectations for scoring a level 1 on a 4-point rubric. Students falling below the expected score range (e.g., .75 on a 1.00-4.00 scale) should be awarded the lowest possible achievement level—Level 1. Students scoring above the expected range should be awarded the highest possible achievement level—Level 4.

Standard Setting Report 2017

Introduction

The purpose of standard setting is to determine where along a continuum of scores different levels of achievement or performance are located. Standard setting studies are used to identify the specific scores, or cut scores, that separate one performance level from another. The “score continuum” for participating PACE districts is the average competency student score for each of the PACE subject areas and grade levels. The average competency score is used instead of something more complex such as a weighted factor score because of the small numbers of students in many of the districts (i.e., factor analysis requires fairly large sample sizes to produce stable results). Students who had competency scores that fell out of range (e.g., 0.75 on a 1.00-4.00 scale) for a given grade level and subject area were removed from the standard setting dataset for that grade level and subject area.

Standard Setting Method

An examinee-centered judgmental method called contrasting groups was used to establish the cutscores. This standard setting method involves asking panelists to make judgments about the qualifications of the examinees based on prior knowledge of the examinee and rich narrative descriptions of various achievement levels called Achievement Level Descriptors (ALD). To implement this method for the PACE pilot, we asked teachers at the end of the school year to make judgments about which achievement level best described each of their students. This process relies heavily on a common understanding and interpretation of the ALDs. The subject and grade specific ALDs were entered into an online survey where teachers could easily read the descriptions and match their students to the appropriate achievement level. The contrasting groups standard setting methodology then involves comparing the average PACE competency 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 with a score of X has a 50% or greater probability of being classified in Level 3 or higher). A logistic regression analysis was run separately for each cut point—Level 2, Level 3, and Level 4—in each district, subject area, and grade level. The results of the contrasting groups standard setting analyses are shown in the figure on the next three pages. Those cells highlighted in orange were modified based on flagging and adjusting protocols (described in more detail after the figures).

Concord				
		Level 2	Level 3	Level 4
ELA	4	1.94	2.74	3.71
	5	1.81	2.60	3.47
	6	72.51	85.84	95.21
	7	70.95	85.77	96.29
	9	51.03	85.07	96.39
	10	67.01	85.93	96.39
Math	3	2.10	2.84	3.72
	5	1.87	2.62	3.31
	6	73.05	85.99	93.56
	7	73.37	86.14	98.88
	9	57.92	79.33	95.06
	10	63.38	80.93	93.70
Sci	4	1.68	2.64	4.00
	8	59.20	82.68	95.64
	9	64.25	79.93	94.03
	10	no annual determinations reported		

Epping				
		Level 2	Level 3	Level 4
ELA	4	1.67	2.78	3.61
	5	1.64	2.50	3.41
	6	56.75	78.39	95.62
	7	50.26	85.24	96.95
	9	1.99	2.71	3.83
	10	1.70	2.50	3.77
Math	3	2.00	2.58	3.64
	5	1.38	2.56	3.35
	6	74.79	85.17	93.83
	7	57.07	81.82	97.90
	9	1.49	2.90	3.51
	10	1.51	2.26	3.74
Sci	4	2.25	2.59	3.54
	8	66.92	85.82	98.10
	9	1.00	2.64	3.58
	10	1.01	2.63	3.59

Monroe				
		Level 2	Level 3	Level 4
ELA	4	1.99	2.99	3.25
	5	2.00	3.00	3.50
	6	no annual determinations reported		
	7	2.04	2.63	3.32
	9	no annual determinations reported		
	10	no annual determinations reported		
Math	3	1.80	2.60	3.01
	5	1.88	2.76	3.50
	6	no annual determinations reported		
	7	2.01	2.72	3.06
	9	no annual determinations reported		
	10	no annual determinations reported		
Sci	4	2.00	3.00	3.50
	8	1.80	2.59	3.30
	9	no annual determinations reported		
	10	no annual determinations reported		

Pittsfield				
		Level 2	Level 3	Level 4
ELA	4	2.40	3.14	3.69
	5	1.75	2.80	3.43
	6	1.81	2.93	3.66
	7	2.84	2.95	3.70
	9	1.73	2.81	3.52
	10	no annual determinations reported		
Math	3	1.57	3.18	3.61
	5	1.59	3.47	4.00
	6	1.36	2.46	3.46
	7	2.50	2.86	3.09
	9	1.25	3.00	3.50
	10	no annual determinations reported		
Sci	4	1.43	2.59	3.67
	8	1.40	2.70	3.52
	9	1.84	3.31	3.85
	10	no annual determinations reported		

Rochester				
		Level 2	Level 3	Level 4
ELA	4	2.53	3.25	3.99
	5	2.53	3.17	4.00
	6	2.41	3.46	4.00
	7	2.77	3.69	4.00
	9	2.54	3.82	4.00
	10	2.36	3.67	4.00
Math	3	2.38	2.95	3.68
	5	2.40	3.00	3.91
	6	2.79	3.45	4.00
	7	2.52	3.32	3.96
	9	3.13	3.92	4.00
	10	2.44	3.57	4.00
Sci	4	1.95	3.08	4.00
	8	1.98	3.49	4.00
	9	2.34	3.63	4.00
	10	2.47	3.90	4.00

Sanborn				
		Level 2	Level 3	Level 4
ELA	4	2.03	2.81	3.09
	5	2.27	2.83	3.43
	6	1.15	2.35	3.75
	7	1.64	2.95	3.58
	9	1.22	2.50	3.56
	10	1.44	2.71	3.95
Math	3	1.97	2.73	3.57
	5	2.06	2.73	3.34
	6	1.34	2.60	3.69
	7	2.05	2.83	3.86
	9	2.05	3.02	3.94
	10	1.19	2.97	4.00
Sci	4	1.77	2.75	3.61
	8	0.33	2.62	3.48
	9	1.45	2.77	3.72
	10	1.92	2.90	3.72

Seacoast				
		Level 2	Level 3	Level 4
ELA	4	1.26	2.76	3.47
	5	1.81	2.67	3.49
	6	2.00	2.83	3.63
	7	1.43	2.83	4.00
	9			
	10			
Math	3	1.88	2.81	3.51
	5	1.54	2.63	3.44
	6	1.76	2.97	3.26
	7	1.76	2.74	3.52
	9			
	10			
Sci	4	1.68	2.39	4.00
	8	1.33	2.59	3.80
	9			
	10			

Souhegan				
		Level 2	Level 3	Level 4
ELA	4			
	5			
	6			
	7			
	9	1.44	2.52	3.80
	10	1.78	2.78	3.61
Math	3			
	5			
	6			
	7			
	9	2.60	2.73	3.53
	10	1.95	2.82	3.35
Sci	4			
	8			
	9	1.75	3.13	4.00
	10	2.07	3.02	3.66

Bethlehem				
		Level 2	Level 3	Level 4
ELA	4	2.38	3.17	4.00
	5	1.69	2.94	3.25
	6	2.26	3.00	4.00
	7			
	9			
	10			
Math	3	1.68	2.89	3.31
	5	2.17	2.89	3.14
	6	2.15	2.79	3.42
	7			
	9			
	10			
Sci	4	2.01	2.15	3.61
	8			
	9			
	10			

Lafayette				
		Level 2	Level 3	Level 4
ELA	4	1.51	2.51	3.52
	5	1.98	2.99	3.64
	6	1.51	2.59	4.00
	7			
	9			
	10			
Math	3	1.50	2.50	3.52
	5	2.00	2.36	3.05
	6	1.51	2.05	3.08
	7			
	9			
	10			
Sci	4	2.00	2.51	3.52
	8			
	9			
	10			

Landaff				
		Level 2	Level 3	Level 4
ELA	4			
	5			
	6			
	7			
	9			
	10			
Math	3	2.76	3.17	3.57
	5			
	6			
	7			
	9			
	10			
Sci	4			
	8			
	9			
	10			

Lisbon				
		Level 2	Level 3	Level 4
ELA	4	2.03	2.36	2.97
	5	1.85	2.80	3.50
	6	2.45	2.80	3.71
	7	1.57	2.51	3.57
	9	1.51	3.00	3.50
	10	1.64	2.28	3.02
Math	3	2.05	2.45	2.80
	5	2.18	3.00	3.83
	6	1.76	2.68	3.50
	7	1.76	2.59	3.75
	9	2.42	3.26	3.58
	10	2.51	3.00	3.90
Sci	4	2.16	2.37	2.77
	8	2.50	2.89	3.57
	9	2.06	2.78	3.32
	10	1.91	2.82	3.62

Profile				
		Level 2	Level 3	Level 4
ELA	4			
	5			
	6			
	7	66.40	90.24	99.35
	9	68.43	87.45	93.06
	10	74.82	82.80	95.28
Math	3			
	5			
	6			
	7	13.86	74.99	88.88
	9	81.12	84.51	95.00
	10	84.00	85.66	91.53
Sci	4			
	8	68.65	85.21	95.02
	9	57.00	80.52	87.98
	10	1.50	2.97	3.51

Flagging Rules

Cut scores were flagged for potential adjustment for four reasons.

5. **Non-significant.** In some cases, while the logistic regression was able to generate estimates, the model itself was not able to explain a statistically significant amount of variance in the dependent variable.
6. **Out of range.** In some cases (see many Level 4 cuts in Rochester), teachers tended to rate their students more harshly on the ALD judgment surveys than the competency scale scores reflected. In these cases, the estimated cut score for the highest achievement level would often fall outside the obtainable competency score range (e.g., scores greater than 4 on a 1-4 scale).
7. **Not estimated.** In some cases, there were insufficient data for the logistic regression model to converge. For example, this would happen if within a given course, the teachers awarded very few Level 1's or Level 4's.
8. **Evidence of Incomparability in Local Scoring.** In a few cases, there were multiple sources of evidence indicating an issue of incomparability in local scoring (e.g., Pittsfield Grade 5 and 9 Math). The flagging rules and adjustment procedure are detailed more in the section entitled "Cross-District Comparability Analyses."

Adjustment Protocols

The following adjustment protocols describe the cut score modifications that were made in reaction to the flagged cut scores. These cut score adjustment procedures are sequential in that they were followed in order, if the first modification was not suitable, the second was attempted, if not suitable, the third, and so on.

6. **No adjustment.** In the case of non-significant model estimation, the cut score estimate was within reasonable expectation and remained the most justifiable best guess for where the cut score should be given the data. In those cases, the cut score was left unaltered.
7. **Adjustment to HOSS.** When the cut score fell above of the obtainable competency score range, the cut score for Level 4 was adjusted to the highest obtainable scale score.
8. **Midpoint.** When the cut score was not estimated, and fell between two estimated cut scores, the cut score was determined to be the midpoint between the two estimated cut scores.
9. **Equipercntile.** When there are no estimated cut scores on either side of the flagged cut score (e.g., Level 2 or Level 4 cuts), an equipercntile equating procedure was used to estimate the cut score that would closely replicate the distributions of achievement across the performance levels in the same district and subject for the other grade levels with unadjusted cut scores. In the few cases where there were no other grade levels with unadjusted cut scores, the same grade level was used in the other content areas to approximate the distribution of achievement.

10. Midpoint. In the few cases where the equipercentile cut score was not estimable (due to small sample sizes or low variability), the midpoint between the LOSS and the Level 3 cut was used to estimate the Level 2 cut, and the midpoint between the HOSS and the Level 3 cut was used to estimate the Level 4 cut.

Cross-District Comparability Analyses

****This section is contained in the body of this report****

Sending Cut Scores to New Hampshire DOE

The cut scores were sent to the NH DOE on October 23, 2017. The cut scores were submitted along with directions for calculating the NH PACE Reported Annual Determinations (see Appendix A). Additionally, with the submission of the performance standards to the state, the Center for Assessment included a recommendation that the results of PACE are reported to schools and parents along with a caution that the annual determinations are based on an innovative assessment system for which the validity evidence is still accumulating. We recommended that the score reports additionally specify that there is not yet enough evidence to support the use of these scores for high stakes uses (e.g., school accountability). When making educational decisions, these scores should just be considered as one data point among many including teacher and counselor reports and observations.

Appendix A of Standard Setting Report:

Calculating NH PACE Reported Annual Determinations

Business Rules

October 23, 2017

4. Clean the data

- a. It should be first checked that there is at least one score submitted for each student in all PACE subject areas (as determined by the table below). The exception to this rule is Grade 10 Science in Concord, Grade 6 ELA and Math in Monroe and Grade 10 ELA, Math and Science in Pittsfield as we do not have cut scores for those courses. This means that students in those grades/courses will not have reportable scores.

PACE Administration Chart

	ELA	Math	Science
Grade 3		PACE	
Grade 4	PACE		PACE
Grade 5	PACE	PACE	
Grade 6	PACE	PACE	
Grade 7	PACE	PACE	
Grade 8			PACE
Grade 9	PACE	PACE	PACE
Grade 10	PACE	PACE	PACE

- b. Secondly, ensure that all scores to be included in the score calculation fall within the intended range. If any scores submitted for any student fall outside the range (e.g., 0.75 on a 1.00-4.00 scale, 102 on a 100-point scale) they should be reconciled (e.g., follow up with the district or school to correct the data entry or scoring error). The one exception to this rule is Epping Grade 8 Science. There are a number of students with scores >100.00 on a 100-point scale. These scores are legitimate and should be maintained. These high scores were factored into the standard setting process.
- c. Students with no competency scores are considered non-participants.

5. Calculate mean scores by subject area
 - a. All submitted competency scores for each student in each subject area need to be averaged. The resulting student-by-subject averages are henceforth referred to as the student average competency scores.
 - b. Note: **For high school, the average competency scores should be computed according to the grade level associated with the assessment** rather than the grade level associated with the student. We have a course-based competency system for high school and therefore the grade level of the student may not reflect the correct competency framework.
 - c. Round the average competency scores to two decimal places.
6. Determine the reportable achievement level of each student
 - a. The average competency scores that result from step 2 need to be classified into achievement levels using the provided cut scores (attached).
 - b. Though the occurrence is rare, some average competency scores will fall outside the expected score range, even with follow-up reconciliation with districts. This is most commonly due to the awarding of zero's for achievement that is so low that the student work consistently does not meet the expectations for scoring a level 1 on a 4-point rubric. Students falling below the expected score range (e.g., .75 on a 1.00-4.00 scale) should be awarded the lowest possible achievement level—Level 1. Students scoring above the expected range should be awarded the highest possible achievement level—Level 4.

Appendix F: Body of Work Standards Validation 2016 and 2017

Body of Work (BOW) Standards Validation 2016

As part of validating the annual determinations produced for the 2015-2016 school year, we have collected a “body of evidence” for a small sample of students from a sample of courses in each participating district. Throughout the academic year we have asked that each district choose a sample of nine students, representing the range of performance in that district, for one content area per grade level. Teachers are asked to collect samples of student work from those nine students for each of the competencies. In July 2016, teachers from across the eight PACE districts came together to review the portfolios of student work to and make judgments about student achievement relative to the Achievement Level Descriptors. Like the consensus scoring activity, teachers were paired in cross-district teams and reviewed bodies of work from students who do not attend either of their home districts. These teacher judgments regarding the student achievement levels were then reconciled with the reported annual determinations as an additional source of validity evidence to support the PACE innovative assessment system.

For the Body of Work analysis, the ratings were kept for only those portfolios upon which the cross-district pair of teachers showed agreement on a common rating. 94.2% percent of the student portfolios received a common rating across the two teachers. Those portfolios that received a score of 0, indicating the work was not scoreable (e.g., copy quality was poor, copy was incomplete), were also removed from the analyses. In all, 110 student portfolios were analyzed in ELA, 92 in Math, and 73 in Science.

Figure 21 graphs the distribution of “body of work” or portfolio ratings for all of the students falling into each annual determination achievement level. The dark green bar represents a match between the PACE annual determination and the body of work rating. Table 32 further parses this data by subject area and reports on the correlation between the two sets of scores, and the percent exact and adjacent agreement.

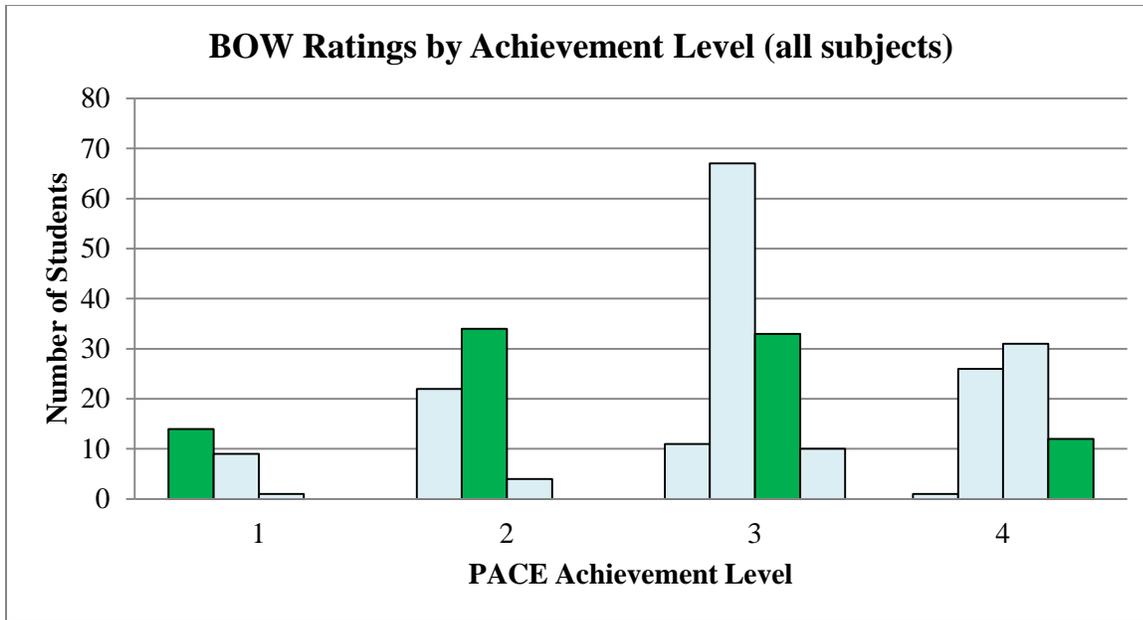


Figure 21. Distribution of BOW ratings by PACE Achievement Level

Table 32.

Agreement Rates by Subject

	Spearman Correlation	%Exact Agreement	%Adjacent Agreement	Exact or Adjacent (sum)
ELA	.629**	38.2%	53.6%	91.8%
Math	.580**	31.5%	51.1%	82.6%
Science	.378**	30.1%	50.7%	80.8%

**significant at the $\alpha = .01$ level.

In general, the agreement between the BOW ratings and the PACE annual determinations is not as strong as expected. Figure 21 shows evidence of systematic underestimation of the PACE Annual Determinations on the part of the teacher raters of the summer. This means that upon evaluating the evidence of student work, teacher raters were more likely to give the student a rating that was lower than the reported annual determination. Though this finding is unexpected and does not provide the intended validity evidence to support the PACE annual determinations, it does not necessarily provide evidence against score validity. Instead, many teachers reported that upon completion of this activity, they had a greater understanding of the purpose of collecting samples of student work throughout the year that are truly reflective of the students' achievement on the full range of competencies. Teachers found that the student work samples that had been selected to support this activity were generally of low level, and therefore, made it difficult to find evidence to support a high achievement level. Teacher reactions and logistical comments will be additionally provided in the HUMRRO independent evaluation report. Based on these reports, it is likely that the student work portfolios submitted for review for 2017 will be more representative of student achievement on the full range of competencies, and therefore we are likely to see greater degrees of agreement between ratings and the annual determinations. To support this effort, the Center for Assessment has provided additional training to educators on the purpose and nature of the bodies of evidence they should be collecting throughout the year (see Appendix E).

Body of Work (BOW) Standards Validation 2017

As part of validating the annual determinations produced for the 2016-17 school year, we have collected a “body of evidence” for a small sample of students from a sample of courses in each participating district. Throughout the academic year we have asked each district to choose a sample of nine students, representing the range of performance in that district, for one content area per grade level. Teachers were asked to collect samples of student work from those nine students for each of the competencies. In August 2017, teachers from across the PACE districts came together to review the portfolios of student work to make judgments about student achievement relative to the Achievement Level Descriptors. Teachers were randomly assigned to triads in cross-district teams and independently rated bodies of work from a mix of districts using the PACE grade level Achievement Level Descriptors. The independent ratings took place in two rounds with a discussion in between the rounds where triads discussed their independent rating with their assigned partners using evidence from the body of student work to support their rating. A crossed design was utilized to assign bodies of work such that there was one student body of work that was the same among triads in a step-wise fashion. The median value of the Round 2 BOW teacher judgments about the student’s achievement level were then reconciled with the within-district teacher report judgment on a Teacher Judgment Survey (TJS), which was also based on the Achievement Level Descriptors. This analysis is an additional source of validity evidence to support the PACE innovative assessment system.

For the Body of Work (BOW) analysis, student bodies of work were kept for only those portfolios that also had a within-district Teacher Judgment Survey (TJS) that could be matched based on student ID, district, subject, and grade. Also, student bodies of work were kept for only those portfolios that had at least 2 raters. Table 54 below shows the number of matched BOW and TJS samples by grade, subject, and district (N=354).

Grade	N	Subject	N	District	N
3	47	ELA	142	Bethlehem	15
4	32	Math	128	Concord	63
5	55	Science	84	Epping	34
6	41	Total	354	Lafayette Regional	7
7	49			Lisbon Regional	20
8	25			Monroe	27
9	40			Pittsfield	26
10	65			Profile	5
Total	354			Rochester	55
				Sanborn Regional	60
				Seacoast Charter	18
				School	
				Souhegan	24
				Cooperative	
				Total	354

Table 54. Number of Matched Student Bodies of Work and Teacher Judgment Surveys by Grade, Subject, and District

Bodies of work in the analysis were rated by 2-6 raters because of the crossed design and issues with rater attendance. This resulted in median values that were sometimes not whole numbers, which were then difficult to compare to the whole number TJS achievement levels 1-4. Median values were therefore rounded up and rounded down to the nearest whole number and results are reported below for both. Table 55 shows the number of median values prior to rounding.

BOW Ratings			TJS Ratings		
Median Achievement			Achievement Level		
Level	N	%	Achievement Level	N	%
1.0	57	16.10	1.0	29	8.2
1.5	10	2.80	2.0	85	24.0
2.0	145	41.00	3.0	147	41.5
2.5	7	2.00	4.0	93	26.3
3.0	108	30.50	Total	354	100.0
3.5	6	1.70			
4.0	21	5.90			
Total	354	100.00			

Table 55. Number of BOW Ratings and TJS Ratings by Achievement Level Prior to Rounding BOW Ratings

Figure 31 illustrates the cross tabulation of BOW portfolio ratings and TJS ratings by achievement level when the BOW median ratings are rounded down (e.g., 1.5 to 1.0) in the left-hand panel the cross tabulation when the BOW median ratings are rounded up (e.g., 1.5 to 2.0) in the right-hand panel. Table 56 further parses this data by subject area and low/high median BOW values and reports on the correlation between the two sets of scores and the percent exact and adjacent agreement.

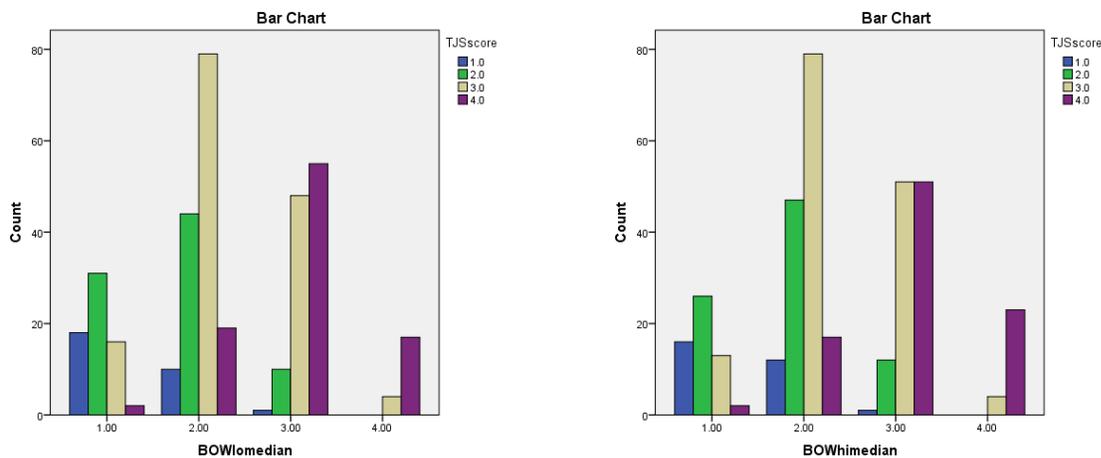


Figure 31. Distribution of BOW Low-Rounded Median Ratings (left-hand panel) or BOW High-Rounded Median Ratings (right-hand panel) by Teacher Judgment Survey Achievement Level

Subject	BOW Median (Low)				BOW Median (High)			
	Spearman Correl.	%Exact	%Adj	%Exact + %Adj	Spearman Correl.	%Exact	%Adj	%Exact + %Adj
ELA	.605***	35.9%	55.6%	91.5%	.609***	36.6%	56.3%	93.0%
Math	.656***	43.0%	49.2%	92.2%	.646***	46.9%	46.1%	93.0%
Science	.521***	25.0%	56.0%	81.0%	.518***	29.8%	53.6%	83.3%

Table 56. Spearman Correlation and % Agreement Rates by Subject and Low/High BOW Median Values

**Significant at the .001 level alpha level.

In general, there were greater degrees of agreement between the BOW ratings and Teacher Judgment Surveys than last year, which most likely reflects better quality BOW samples. The BOW method is well-known in the measurement literature to produce more rigorous cutscores than other standard setting methods and that was the case here as well. This means that upon evaluating the evidence of student work, summer teacher raters were more likely to give the student a rating that was lower than the within-district teacher given achievement level on the Teacher Judgment Survey. Many teachers reported that upon completion of this activity, they had a greater understanding of the purpose of collecting samples of student work throughout the year that are truly reflective of the students' achievement on the full range of competencies. Teachers found that the student work samples that had been selected to support this activity were of mixed quality, and therefore, made it difficult to find evidence to support a high achievement level. To support the collection of higher quality BOW samples that show evidence of the full range of student achievement relative to the competencies and Achievement Level Descriptors, the Center for Assessment will continue to provide additional training to educators on the purpose and nature of the bodies of evidence they should be collecting throughout the year. In addition, because the statewide assessments in elementary and middle school will change from Smarter Balanced (SBAC) to NH SAS starting in 2018, once the NH SAS Achievement Level Descriptors are available, the Center for Assessment will lead a review of the PACE Achievement Level Descriptors so that the two assessment systems are aligned.

Appendix G: Concurrent Analyses 2016 and 2017

Concurrent Analyses 2016

Figures 11 and 12 show 2016 performance on the two assessment systems (PACE and statewide) for the PACE districts as measured by percent proficient for ELA and math, respectively. The blue bars are PACE grades, the red bars are Smarter Balanced (SBAC) grades, and the green bars are the SAT grades. The figures reveal that the percentage of students deemed proficient across the assessment systems is remarkably consistent. But for the colors indicating the different assessments, student performance across the two systems would be indistinguishable.

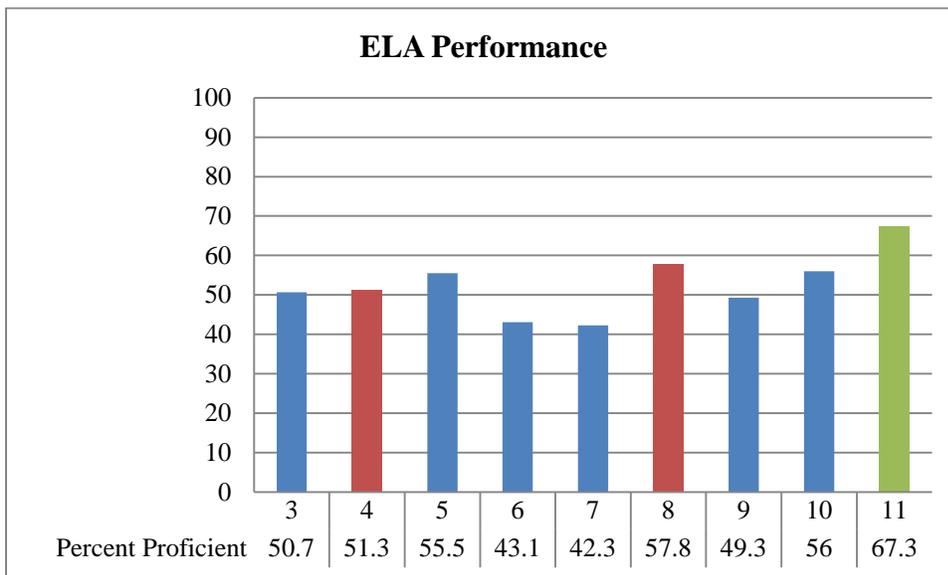


Figure 11. PACE District Performance in ELA across Assessment Systems

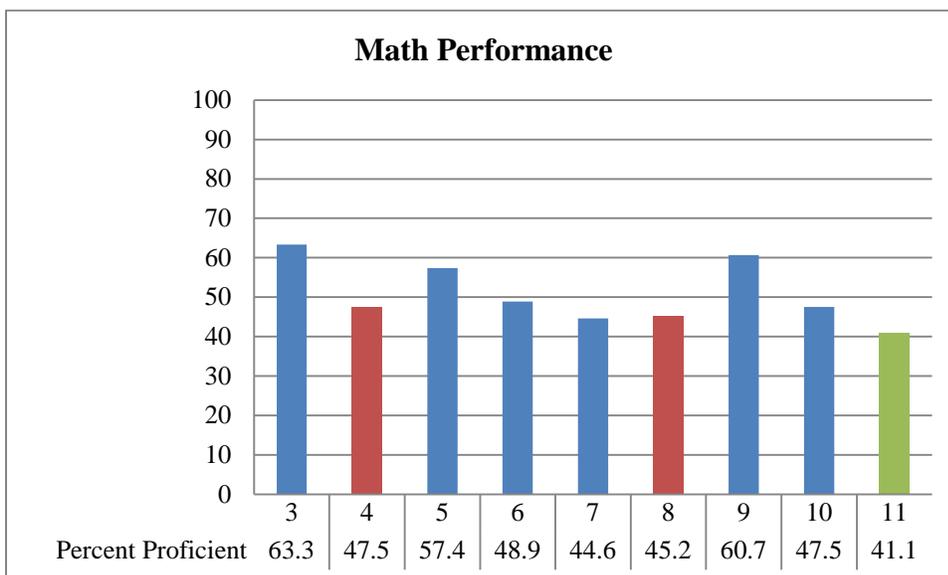


Figure 12. PACE District Performance in Math across Assessment Systems

Secondly, by calculating PACE annual determinations for the students taking SBAC this year, the state has both SBAC and PACE 2015-2016 annual determinations for students in grade 3 ELA, grade 4 math, grade 8 ELA and math, and grade 11 ELA and math. Though annual determinations were not reported for these subjects and grades for PACE and no common performance task was administered, the same procedure for producing annual determinations was used in these grade levels as for the PACE reported annual determinations. Figures 13-18 display the achievement level distributions for the two sets of annual determinations. The degree of similarity between the distributions provides further support the comparability of the interpretations of the reported achievement levels. Note: The, Figures 17 and 18 only include data from the students in Concord, Epping, Rochester, and Sanborn. The other districts either do not have grade 11 students or did not submit competency scores for grade 11.

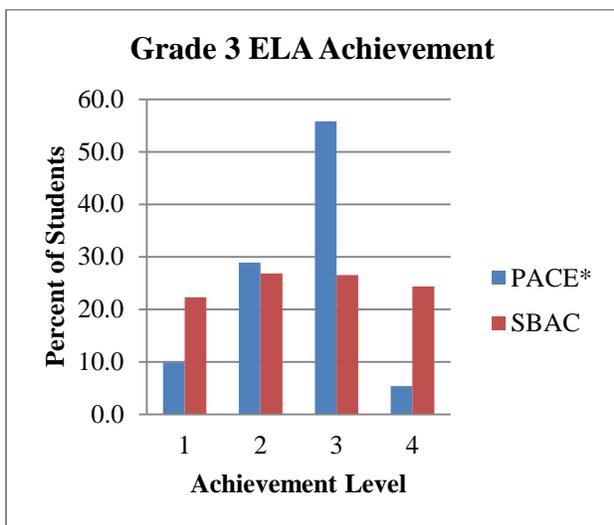


Figure 13. G3 ELA

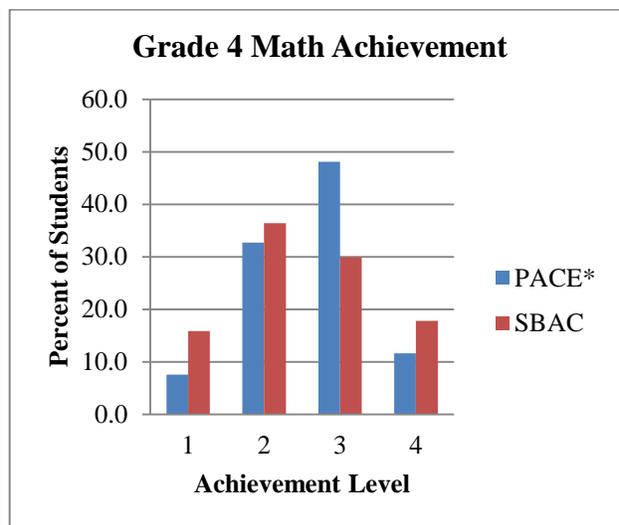


Figure 14. G4 Math

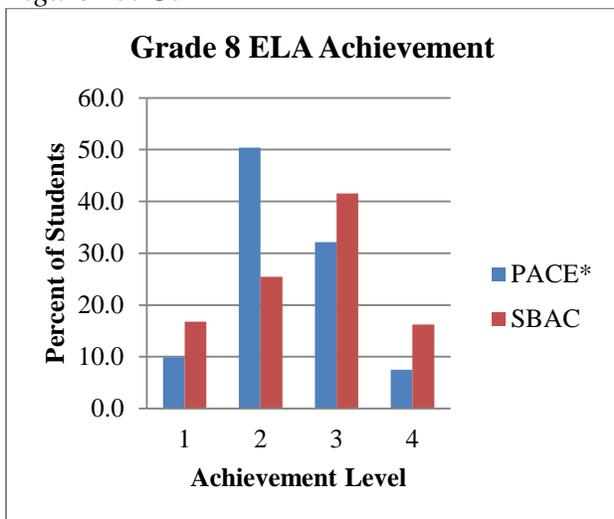


Figure 15. G8 ELA

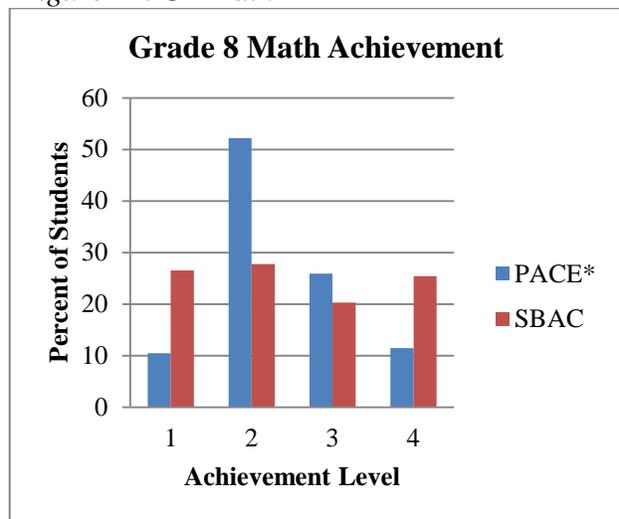


Figure 16. G8 Math

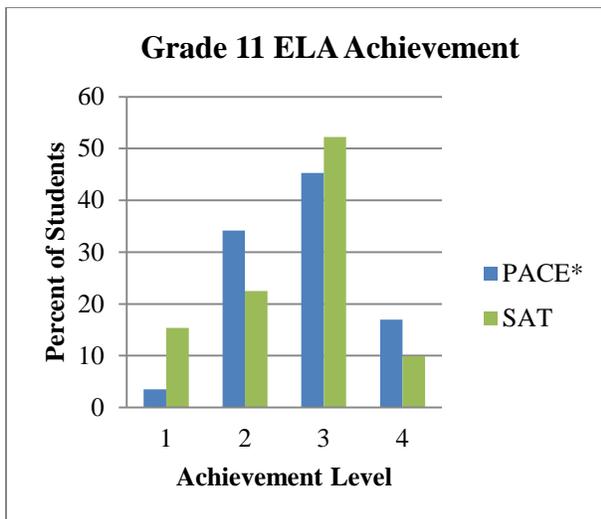


Figure 17. G11 ELA

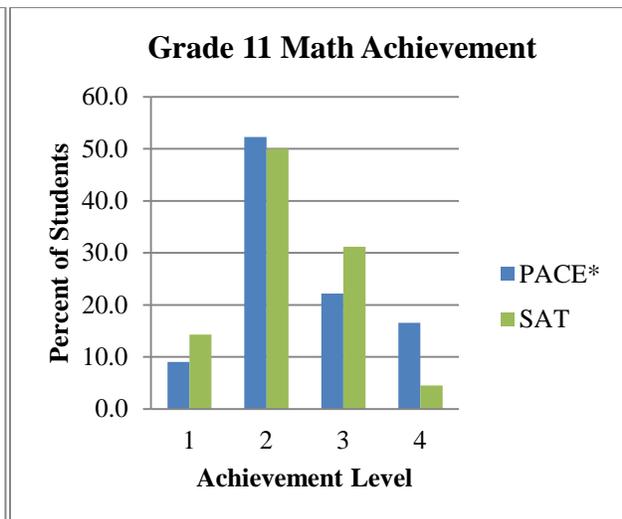


Figure 18. G11 Math

While the figures shown above are compelling, Tables 18-21 provide additional information regarding the classification accuracy by matching students across the assessment systems.

Table 18.

Classification Accuracy for SBAC ELA

		Proficient on SBAC	
		No	Yes
Proficient on PACE	No	34.6%	14.4%
	Yes	11.3%	39.7%

Table 19.

Classification Accuracy for SBAC Math

		Proficient on SBAC	
		No	Yes
Proficient on PACE	No	40.8%	10.4%
	Yes	12.6%	36.3%

Table 20.

Classification Accuracy for SAT ELA

		Proficient on SAT	
		No	Yes
Proficient on PACE	No	23.3%	14.4%
	Yes	14.5%	47.8%

Table 21.

Classification Accuracy for SAT Math

		Proficient on SAT	
		No	Yes
Proficient on PACE	No	48.2%	13.0%
	Yes	17.4%	21.4%

For all four comparisons presented in Tables 18-21, the classification accuracy is at least 70% agreement. While this agreement is high, there are a variety of reasons why there may be legitimate differences in the results produced by the different assessment systems. First, the degree of agreement is limited by the reliability of each assessment system. In other words, an assessment cannot correlate more with another assessment than it can with itself (i.e., reliability), so since both PACE and Smarter Balanced (or SAT) are not perfectly reliable, we are approaching the upper bound of the relationship between the two assessment systems. Additionally, New Hampshire's PACE assessment system is in place to measure the state-defined learning targets differently than they are measured in the statewide assessment system. The purpose is to measure the standards more deeply and authentically through performance-based assessments. Additionally, the PACE assessment system is intended to measure the set of standards more completely (e.g., including the listening and speaking standards). Therefore, perfect agreement between the two assessment systems would be an indication of failure on the part of the PACE assessment system. The demonstrated 70% agreement in proficiency classification across the two systems should be considered acceptable given the competing objectives of attaining comparability while designing and implementing an innovative assessment system that is intended to create meaningful changes to teaching and learning.

Table 22 shows the proficiency classification accuracies for the waiver-reported subgroups. The classification accuracies for the reported subgroups do not vary greatly from the overall classification accuracy of approximately 70%. Some variation around 70% is natural due to sampling error associated with the small sample sizes of many of the subgroups. The only subgroups with proficiency classification accuracies of less than 60% are African Americans and students who are two or more races (non-Hispanic). We will pay particular attention to those subgroups of students in next year's analyses to ensure this observation is not an indication of something systematic.

Table 22.

Concurrent PACE to SBAC Classification Accuracies for Subgroups

	SBAC ELA	SBAC Math	SAT ELA	SAT Math
American Indian or Alaskan Native	**	**	**	**
Asian	84.8%	78.8%	89.5%	**
Black or African American	73.3%	77.2%	52.6%	**
Hispanic or Latino	75.6%	83.3%	71.4%	**
Native Hawaiian or Pacific Islander	**	**	**	**
Two or more races (non-Hispanic)	64.3%	58.8%	**	**
White	73.9%	76.9%	70.9%	69.4%
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**
WaiverSubgroup - SWD and EconDis - Not EL	86.3%	90.2%	81.8%	**
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	81.9%	78.8%	69.4%	72.7%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	100.0%	**	**
WaiverSubgroup - EconDis and EL - Not SWD	89.3%	75.0%	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	68.4%	72.1%	66.0%	75.8%

**Sample size is <10

Concurrent Analyses 2017

Figures 10 and 11 show the percentage of students proficient or above on the two assessment systems (PACE and statewide) for the PACE districts for ELA and math, respectively. The blue bars represent PACE grades, the red bars represent Smarter Balanced (SBAC) grades, and the green bars represent the SAT grade. The figures reveal that the percentage of students deemed proficient or above across the assessment systems is remarkably consistent.

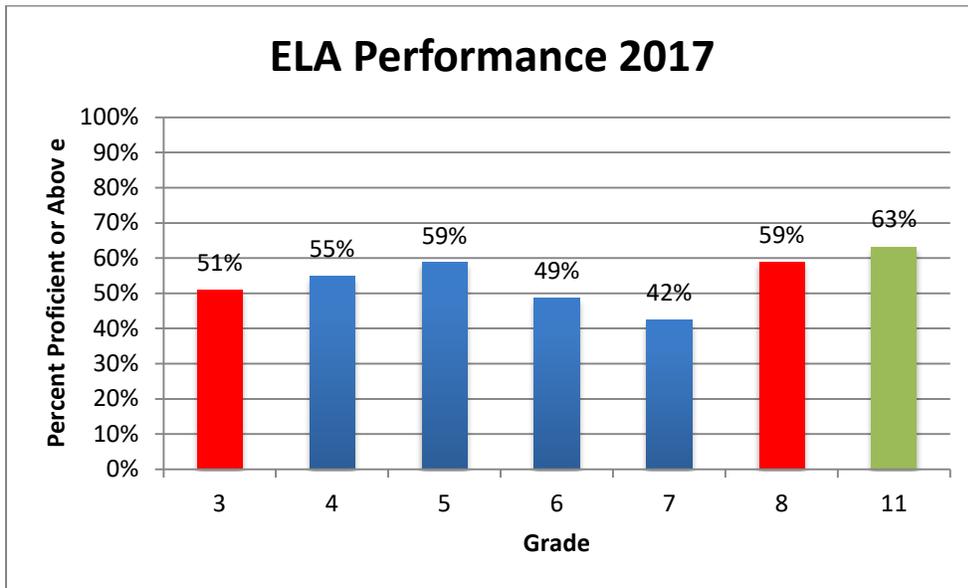


Figure 10. PACE District Performance in ELA across Assessment Systems

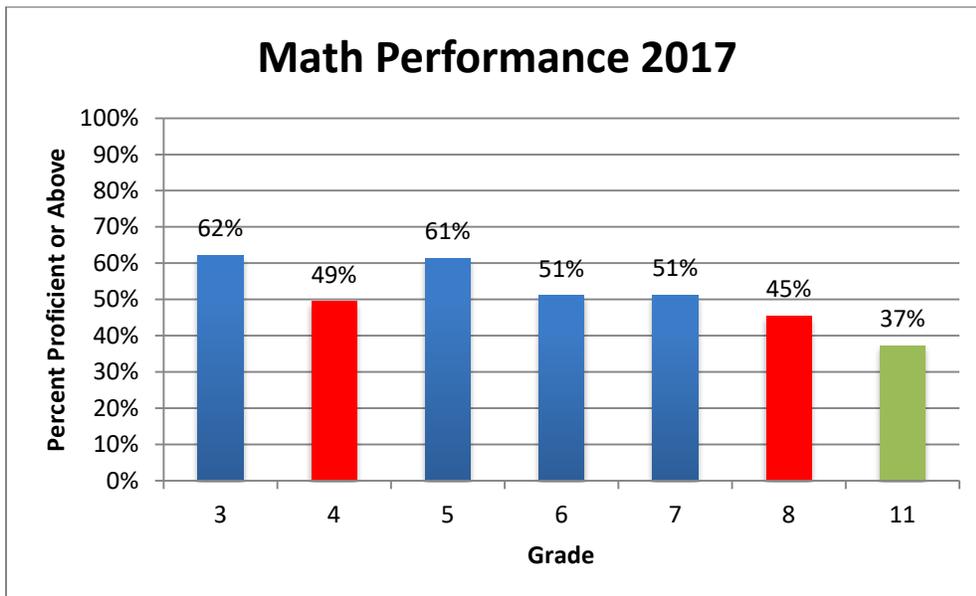


Figure 11. PACE District Performance in Math across Assessment Systems

Secondly, PACE annual determinations were calculated for the students taking SBAC this year. This means the state has SBAC and PACE 2016-17 annual determinations for students in grade 3

ELA, grade 4 math, grade 8 ELA and math, and for the SAT in grade 11 ELA and math. Though annual determinations were not reported for these subjects and grades using the PACE results and no common performance task was administered, the same procedure for producing PACE annual determinations (i.e., contrasting groups survey and competency scores) was used in these grade levels as for the PACE reported annual determinations. Table 18 shows the number of matched students by subject, grade, and district included in the analyses below (N=4,339).

Subject	Grade	District	N
ELA	3	Concord	313
		Landaff	2
		Monroe	6
		Pittsfield	21
		Rochester	260
		Sanborn	97
		Seacoast	19
	8	Concord	329
		Epping	73
		Monroe	6
		Profile	34
		Rochester	273
		Sanborn	130
		Seacoast	22
	11	Concord	247
		Pittsfield	23
		Profile	38
		Sanborn	138
		Souhegan	224
Math	4	Bethlehem	13
		Concord	316
		Epping	79
		Monroe	5
		Pittsfield	33
		Rochester	285
		Sanborn	101
	Seacoast	35	
	8	Concord	330
		Monroe	6
		Pittsfield	15
		Profile	32
		Rochester	266
		Sanborn	129
		Seacoast	22
	11	Concord	113
		Pittsfield	14
		Profile	37
		Sanborn	129
Souhegan		124	

Table 18. Number of matched students by subject, grade, and district (N=4,339)

Figures 12-13 display the overall percent of students scoring proficient or above in ELA and math across the two assessment systems. As before, the blue bars represent PACE, red bars

represent SBAC, and green bars represent SAT. The degree of similarity in the percentage of students deemed proficient or above across the two assessment systems further supports the comparability of proficiency designations between assessment systems.

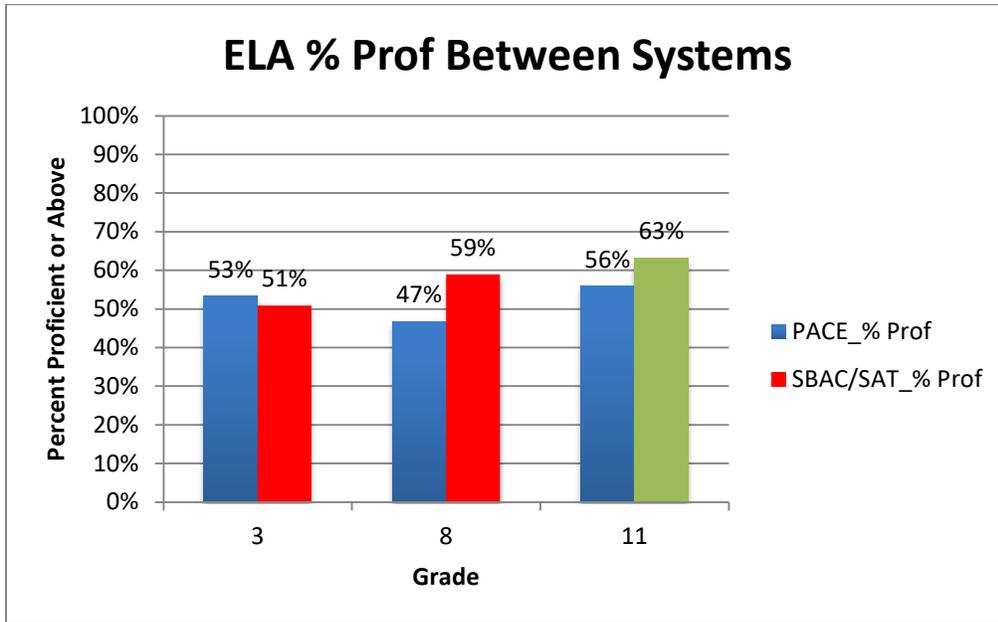


Figure 12. Percentage of students proficient or above in ELA between the PACE and SBAC/SAT assessment systems by grade level

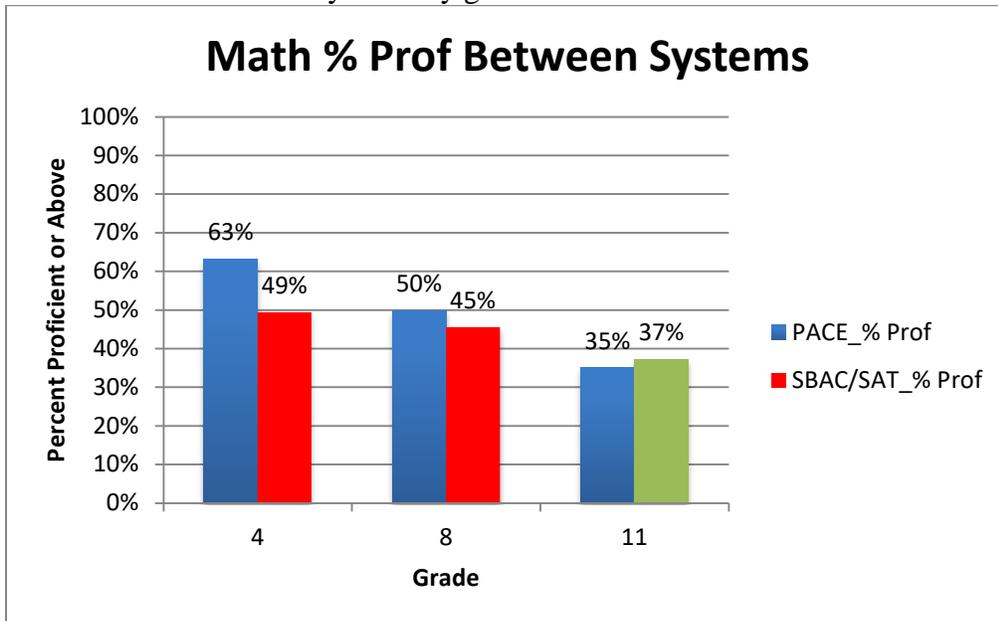


Figure 13. Percentage of students proficient or above in Math between the PACE and SBAC/SAT assessment systems by grade level

Figures 14-19 display the achievement level distributions for the two sets of annual determinations followed by tables 19-24 that provide the number of students included in the figures. The degree of similarity between the distributions provides further support regarding the high degree of comparability of the students scoring at the reported achievement levels.

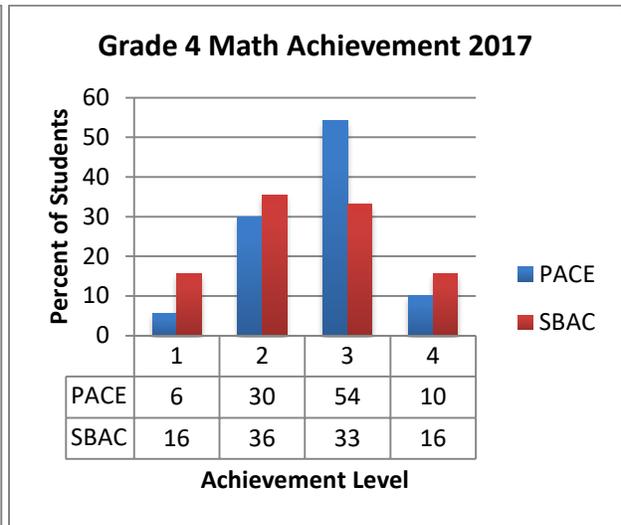
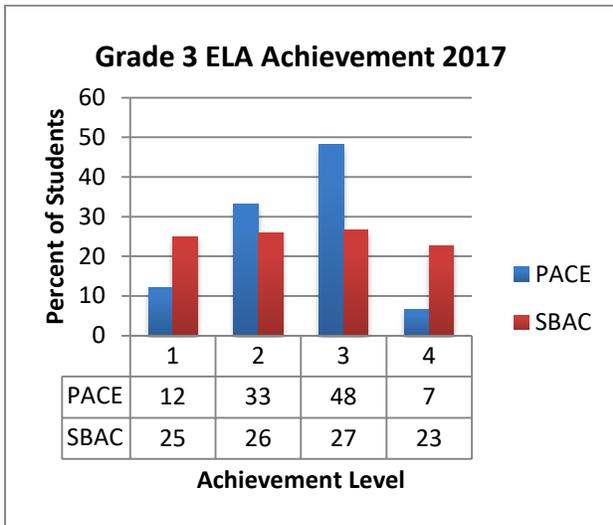


Figure 14. G3 ELA Figure 15. G4 Math

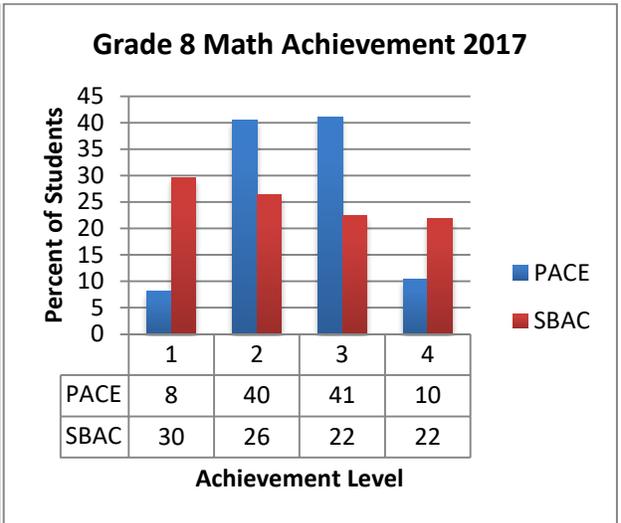
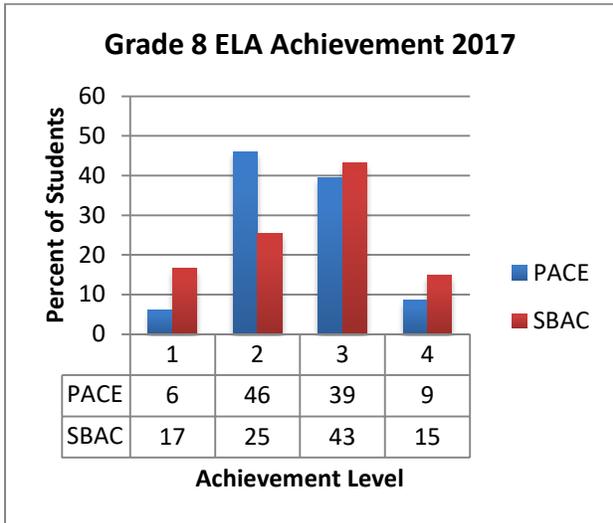


Figure 16. G8 ELA Figure 17. G8 Math

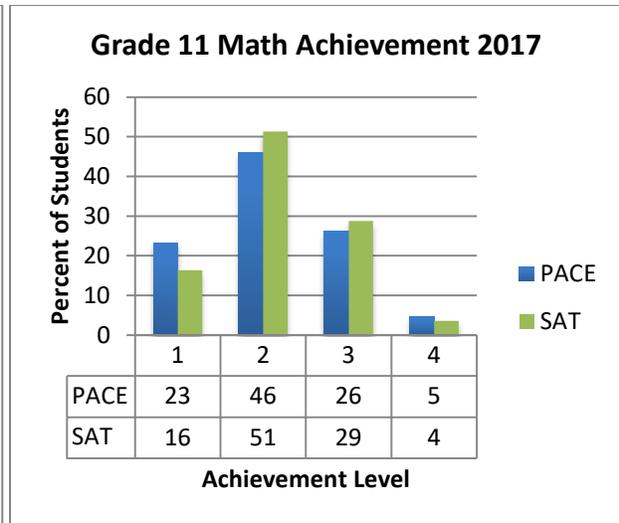
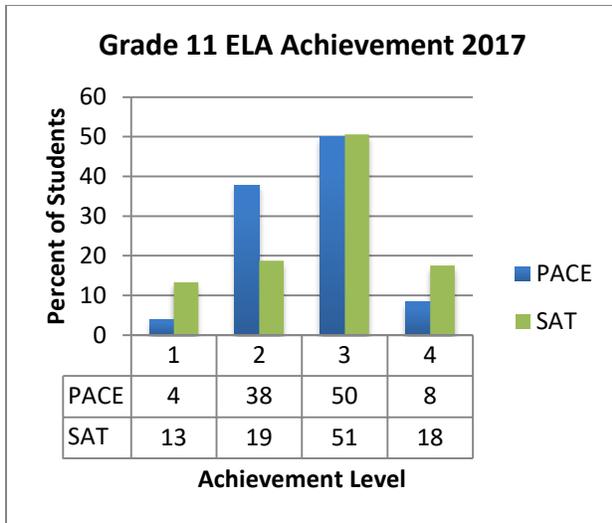


Figure 18. G11 ELA Figure 19. G11 Math

		2017 SBAC			
		1	2	3	4
2017 PACE	1	60	14	9	3
	2	96	90	43	9
	3	22	78	127	120
	4	0	4	13	30

Table 19. Grade 3 ELA Crosstabs (n counts) for 2017 PACE and 2017 SBAC

		2017 SBAC			
		1	2	3	4
2017 PACE	1	35	11	2	0
	2	80	140	37	3
	3	21	150	218	82
	4	0	7	31	50

Table 20. Grade 4 Math Crosstabs (n counts) for 2017 PACE and 2017 SBAC

		2017 SBAC			
		1	2	3	4
2017 PACE	1	23	13	17	1
	2	98	159	126	15
	3	22	48	199	72
	4	0	0	33	41

Table 21. *Grade 8 ELA Crosstabs (n counts) for 2017 PACE and 2017 SBAC*

		2017 SBAC			
		1	2	3	4
2017 PACE	1	47	13	4	1
	2	147	104	52	20
	3	40	86	108	95
	4	3	7	15	58

Table 22. *Grade 8 Math Crosstabs (n counts) for 2017 PACE and 2017 SBAC*

		2017 SAT			
		1	2	3	4
2017 PACE	1	8	7	11	1
	2	62	68	113	10
	3	19	49	195	71
	4	0	1	20	35

Table 23. *Grade 11 ELA Crosstabs (n counts) for 2017 PACE and 2017 SAT*

		2017 SAT			
		1	2	3	4
2017 PACE	1	27	58	12	0
	2	36	111	44	1
	3	5	44	51	9
	4	0	1	13	5

Table 24. *Grade 11 Math Crosstabs (n counts) for 2017 PACE and 2017 SAT*

Table 25 aggregates the crosstabs above showing the percentage of exact agreement and percentage of exact or adjacent agreement by grade and subject area. Importantly, there is over 90% exact or adjacent agreement on achievement levels between the two assessment systems.

	%Exact Agreement	%Exact or Adjacent Agreement
Grade 3 ELA	42.8%	93.5%
Grade 4 Math	51.1%	96.2%
Grade 8 ELA	48.7%	93.7%
Grade 8 Math	39.6%	90.6%
Grade 11 ELA	45.7%	93.7%
Grade 11 Math	46.5%	95.4%

Table 25. Percent Agreement Across 2017 PACE and 2017 SBAC/SAT

Tables 26-31 provide additional information regarding the classification accuracy across the assessment systems. “Classification accuracy” refers to the percentage of students who received the same proficiency classification (i.e., ‘proficient’=Yes or ‘not proficient’=No) across the two years. Note: these analyses assume no student growth across years.

		2017 SBAC (% Proficient or above)	
		No	Yes
2017 PACE (% Proficient or above)	No	36.2%	8.9%
	Yes	14.5%	40.4%

Table 26. *Grade 3 ELA classification accuracy for 2017 PACE and 2017 SBAC*

		2017 SBAC (% Proficient or above)	
		No	Yes
2017 PACE (% Proficient or above)	No	33.8%	18.3%
	Yes	8.1%	39.8%

Table 27. *Grade 8 ELA classification accuracy for 2017 PACE and 2017 SBAC*

		2017 SAT (% Proficient or above)	
		No	Yes
2017 PACE (% Proficient or above)	No	21.6%	20.1%
	Yes	10.3%	47.9%

Table 28. *Grade 11 ELA classification accuracy for 2017 PACE and 2017 SBAC*

		2017 SBAC (% Proficient or above)	
		No	Yes
2017 PACE (% Proficient or above)	No	30.7%	4.8%
	Yes	20.5%	43.9%

Table 29. *Grade 4 Math classification accuracy for 2017 PACE and 2017 SBAC*

		2017 SBAC (% Proficient or above)	
		No	Yes
2017 PACE (% Proficient or above)	No	38.9%	9.6%
	Yes	17.0%	34.5%

Table 30. *Grade 8 Math classification accuracy for 2017 PACE and 2017 SBAC*

		2017 SAT (% Proficient or above)	
		No	Yes
2017 PACE (% Proficient or above)	No	55.6%	13.7%
	Yes	12.0%	18.7%

Table 31. Grade 11 Math classification accuracy for 2017 PACE and 2017 SAT

For all six comparisons presented in Tables 26-31, the classification accuracy is at least 70% agreement. While this agreement is high, there are a variety of reasons why there may be legitimate differences in the results produced by the different assessment systems. First, the degree of agreement is limited by the reliability of each assessment system. In other words, an assessment cannot correlate more with another assessment than it can with itself (i.e., reliability). Therefore, because both PACE and Smarter Balanced (or SAT) are not perfectly reliable, we may be approaching the upper bound of the relationship between the two assessment systems. Additionally, New Hampshire’s PACE assessment system is in place to measure the state-defined learning targets differently than they are measured in the statewide assessment system. The purpose is to measure the standards more deeply and authentically through performance-based assessments. Additionally, the PACE assessment system is intended to measure the set of standards more completely (e.g., including the listening and speaking standards). Therefore, perfect agreement between the two assessment systems would be an indication of failure on the part of the PACE assessment system. The demonstrated 70% agreement in proficiency classification across the two systems should be considered acceptable given the competing objectives of attaining comparability while designing and implementing an innovative assessment system that is intended to create meaningful changes to teaching and learning.

Table 32 shows the proficiency classification accuracies for the waiver-reported subgroups. The classification accuracies for the reported subgroups do not vary greatly from the overall classification accuracy of approximately 70%. Some variation around 70% is natural due to sampling error associated with the small sample sizes of many of the subgroups. The only subgroups with proficiency classification accuracies of less than 60% are 1) African Americans (G8 Math), 2) WaiverSubgroup—Engl Learner (EL) only—Not EconDis, Not SWD (Gr 8 ELA and Math), and 3) WaiverSubgroup—EconDis and EL—Not SWD (G8 ELA and Math). We will pay particular attention to those subgroups of students in next year’s analyses to ensure this observation is not an indication of something systematic. A comparison with last year’s concurrent classification accuracies by subgroup does not reveal any systematic patterns.

Table 32. Concurrent 2017 PACE to 2017 SBAC or SAT Classification Accuracies for Subgroups by Grade and Subject Area						
	G3 ELA	G8 ELA	G11 ELA	G4 Math	G8 Math	G11 Math
American Indian or Alaskan Native	**	**	**	**	**	**
Asian	87.0%	62.5%	81.3%	96.4%	77.5%	**
Black or African American	75.0%	66.7%	63.2%	81.0%	57.9%	80.0%
Hispanic or Latino	75.9%	65.2%	70.0%	79.3%	90.9%	**
Native Hawaiian or Pacific Islander	**	**	**	**	**	**
Two or more races (non-Hispanic)	64.3%	**	**	84.6%	**	**
White	76.8%	74.6%	69.5%	73.4%	73.3%	74.3%
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**	**
WaiverSubgroup - SWD and EconDis - Not EL	83.9%	87.3%	87.5%	84.5%	82.6%	90.0%
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	78.4%	81.1%	72.2%	80.0%	79.7%	82.7%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	90.0%	40.9%	60.0%	81.3%	50.0%	70.0%
WaiverSubgroup - EconDis and EL - Not SWD	88.0%	33.3%	**	80.0%	40.0%	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	72.7%	73.7%	68.9%	75.1%	75.8%	82.8%

**Sample size is <10.

Appendix H: Non-Concurrent Analyses 2016 and 2017

Non-Concurrent Analyses 2016

Comparisons between 2015 PACE and 2016 SBAC

Since students participate in SBAC once per grade span, we have compared last years' performance on PACE with this years' performance on SBAC for students in grade 8 in ELA, and in grades 4 and 8 in Math. Figure 1 shows the percent proficient for the matched cohort of students across years. The blue bars represent math achievement while the red bars indicate ELA. The math achievement is more stable across years, while the percent proficient in ELA rose from PACE in 2015 to SBAC in 2016.

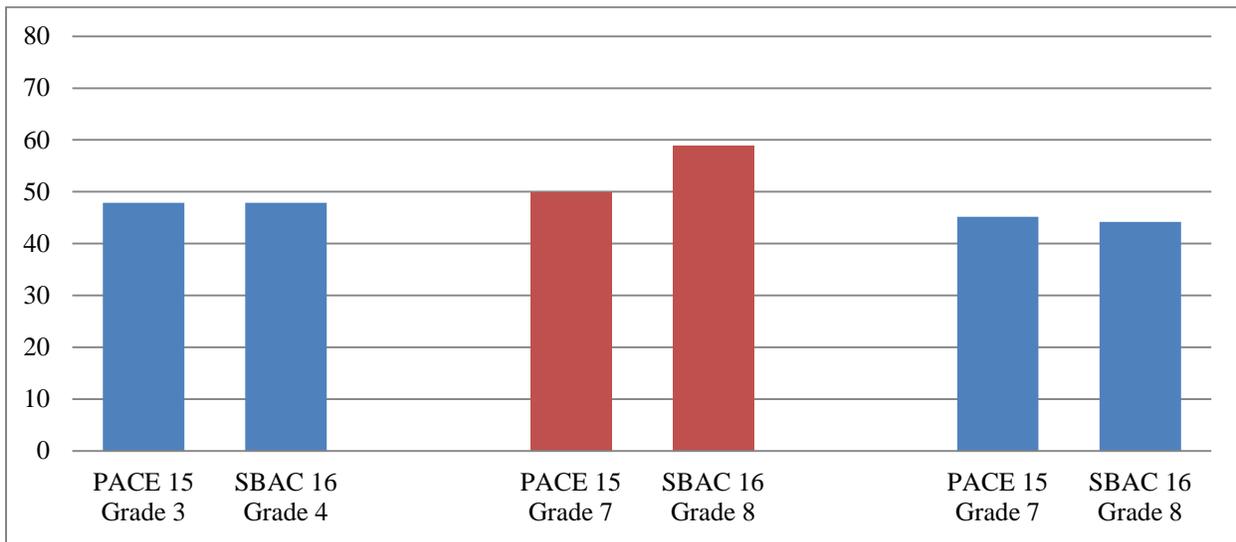


Figure 1. Cohort % Proficient across years and assessment systems

Classification tables are provided in Tables 1-3. “Classification accuracy” refers to the percentage of students who received the same proficiency classification (i.e., ‘proficient’ or ‘not proficient’) across the two years. In this case, classification accuracy may be a misnomer since students can and do legitimately change in their classifications across years

Table 1.
Classification Accuracy for G4 Math

		Proficient on SBAC	
		No	Yes
Proficient on PACE	No	38.2%	13.8%
	Yes	13.8%	34.1%

Table 2.
Classification Accuracy for G8 ELA

		Proficient on SBAC	
		No	Yes

Proficient on PACE	No	31.8%	18.3%
	Yes	9.0%	40.9%

Table 3.
Classification Accuracy for Grade 8 Math

		Proficient on SBAC	
		No	Yes
Proficient on PACE	No	43.9%	10.4%
	Yes	11.9%	33.7%

The classification accuracies across the three comparisons are all above 70%. Additionally, the observed differences in proficiency classifications for are fairly evenly distributed between students moving from proficient to non-proficient and students moving from non-proficient to proficient.

Table 4 shows the proficiency classification accuracies for the waiver-reported subgroups. The classification accuracies for the reported subgroups do not vary greatly from the overall classification accuracy of approximately 70%. Some variation around 70% is natural due to sampling error associated with the small sample sizes of many of the subgroups. The only subgroup with a potentially problematic proficiency classification accuracy is African Americans students. This pattern was also observed in the non-concurrent analyses comparing 2015 SBAC scores with 2016 PACE scores. However, there is no evidence to suggest that one assessment system is systematically rating African American students lower or higher than the other system, instead, the variations in proficiency classification are evenly spread across students moving from non-proficient to proficient and proficient to non-proficient across the two analyses. We will pay particular attention to this subgroup of students in next year’s analyses to ensure this observation is not an indication of something systematic.

Table 4.
2015 PACE to 2016 SBAC Classification Accuracies for Subgroups

	SBAC ELA	SBAC Math
American Indian or Alaskan Native	**	**
Asian	**	84.6%
Black or African American	**	60.0%
Hispanic or Latino	75.0%	69.2%
Native Hawaiian or Pacific Islander	**	**
Two or more races (non-Hispanic)	**	72.7%
White	72.2%	75.7%
WaiverSubgroup - SWD and EL - Not EconDis	**	**
WaiverSubgroup - SWD and EconDis - Not EL	90.3%	91.5%
WaiverSubgroup - SWD and EconDis and EL	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	76.9%	76.4%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	**
WaiverSubgroup - EconDis and EL - Not SWD	**	**
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	69.1%	71.7%

**Sample size is <10, note since 2015 PACE data is only for 4 districts, the n counts are smaller than for the non-concurrent analysis using 2015 SBAC and 2016 PACE.

Tables 5-7 show the results of comparing the 2015 SBAC annual determinations to the 2016 PACE annual determinations across the four achievement levels. Because the 2015 PACE data is only available for 4 districts, the n counts are smaller than for the non-concurrent analysis using 2015 SBAC and 2016 PACE. This information is also provided graphically after the tables.

Table 5.

Crosstabs (n counts) for 2015 PACE and 2016 SBAC ELA

		2016 SBAC ELA			
		1	2	3	4
2015 PACE ELA	1	13	17	6	1
	2	57	86	82	8
	3	4	40	109	38
	4	1	2	22	32

Table 6.

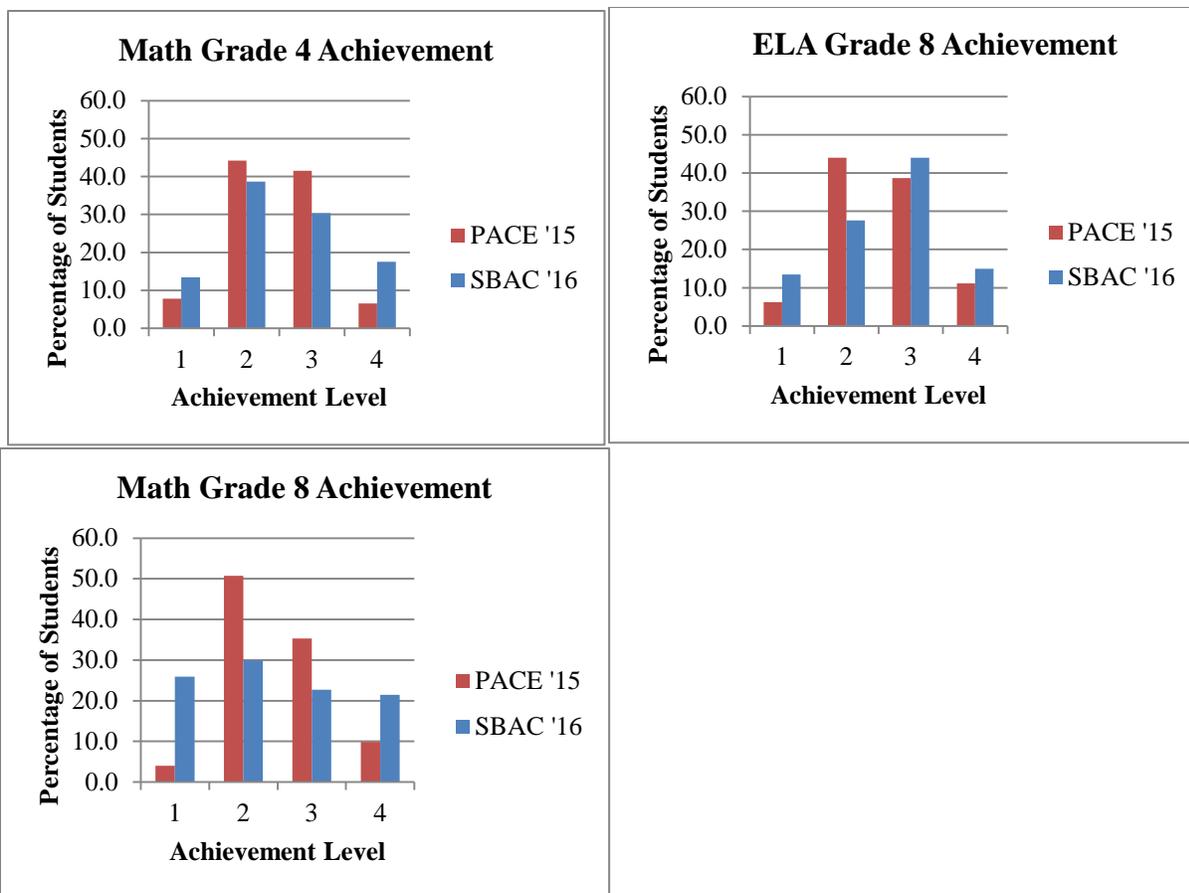
Crosstabs (n counts) for 2015 PACE and 2016 SBAC Math

		2016 SBAC Math			
		1	2	3	4
2015 PACE Math	1	40	15	4	0
	2	136	192	83	26
	3	14	105	135	102
	4	0	5	22	52

Table 7.

Percentage Non-Concurrent Agreement Across PACE and SBAC

	%Exact Agreement	%Exact or Adjacent Agreement
ELA	46.3	95.8
Math	45.0	94.7



As shown in the results above, while there is variation across the two assessment programs, the degree of agreement is high, with above 90% exact or adjacent agreement. The correlations between the two assessment programs across years are $r = 0.538$ for ELA and $r = 0.585$ for math (both statistically significant at the $\alpha=.01$ level). These correlations are remarkably high given that the HUMRRO evaluation report recently reported cross-year reliabilities for the 2015 and 2016 PACE scores ranging from $r = 0.483$ to $r = 0.630$.¹³ Because no assessment is likely to correlate more highly with a different assessment than with itself, the strength of the correlations between 2015 PACE and 2016 SBAC are remarkably high.

Non-concurrent analyses could not be conducted for the SAT given that PACE did not report annual determinations for high school students in 2015.

¹³ See the forthcoming HUMRRO evaluation report for these analyses.

Non-Concurrent Analyses 2017

We conducted two non-concurrent comparability evaluations because students participate in SBAC once per grade span: SBAC 2016 to PACE 2017 and PACE 2016 to SBAC 2017. Each analysis is discussed in a separate section below. Non-concurrent analyses could not be conducted for the SAT given that PACE did not report annual determinations for high school students in 2016 or 2017 as per the requirements of the original waiver.

SBAC 2016 to PACE 2017

The first analysis compares last years' performance on SBAC in grade 3 ELA, grade 4 math, and grade 8 ELA and math with this years' performance on PACE for students in grade 4 ELA, grade 5 math, and grade 9 math and ELA. Only students with an SBAC achievement level in 2016 and a PACE achievement level in 2017 are used for these analyses (N=2,859). Figure 20 shows the percent proficient or above for the matched cohort of students across years. The red bars indicate SBAC and the blue bars represent PACE. In two out of the four grades and subject areas, the percent proficient rose from SBAC 2016 to PACE 2017 and in two grades and subject areas the percent proficient either went down or stayed about the same. In other words, the results demonstrate remarkable consistency of expectations for the same students.

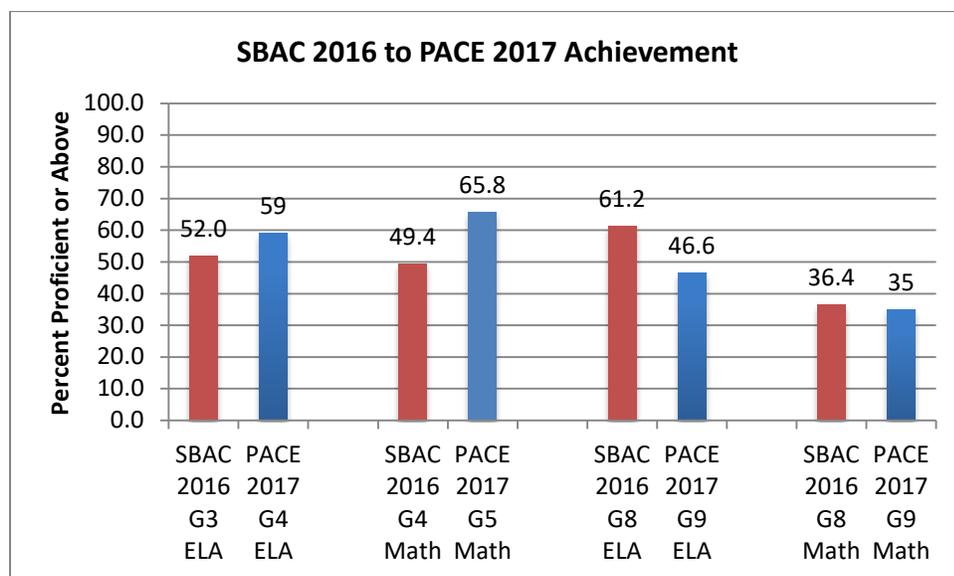


Figure 20. Cohort Percent Proficient or Above across SBAC 2016 to PACE 2017

Figures 21-24 display the achievement level distributions for SBAC 2016 and PACE 2017 by grade level and subject area, followed by Tables 33-36 that provide the number of students included in the figures. There does not appear to be any common pattern with regards to changes in achievement levels across years by subject area or grade level except to note that PACE generally has fewer students scoring at Level 4 than is the case for the state assessment system.

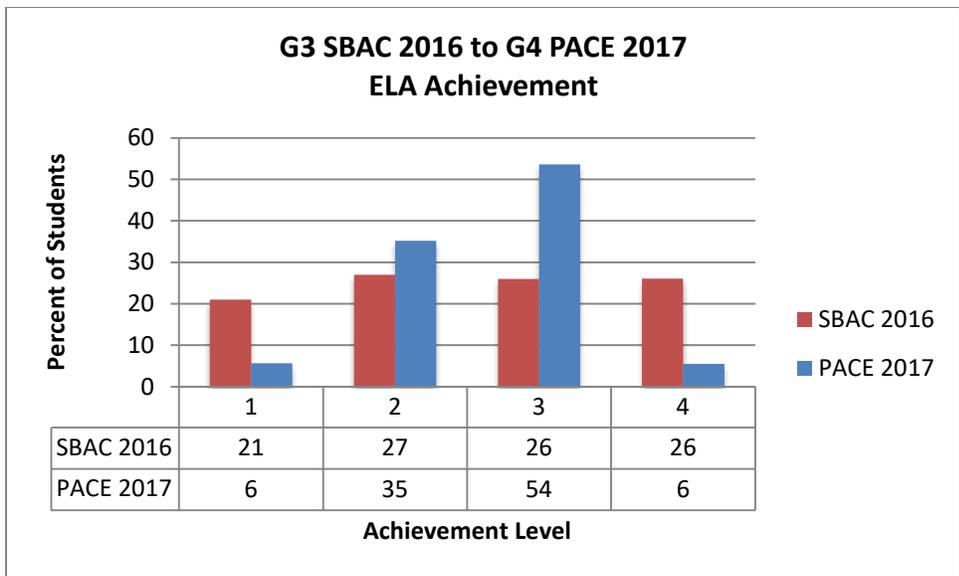


Figure 21. Percent of Students by Achievement Level for Grade 3 ELA SBAC 2016 to Grade 4 ELA PACE 2017

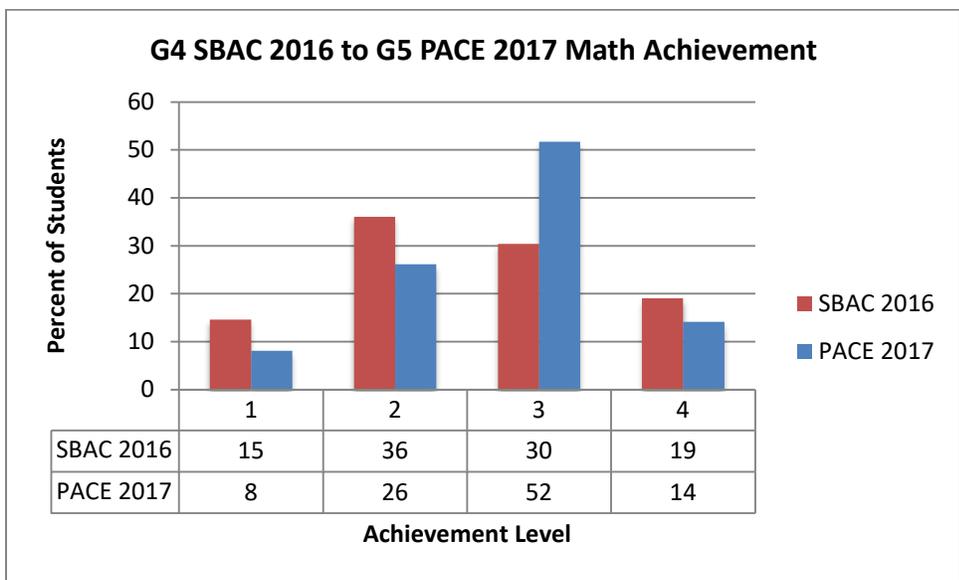


Figure 22. Percent of Students by Achievement Level for Grade 4 Math SBAC 2016 to Grade 5 Math PACE 2017

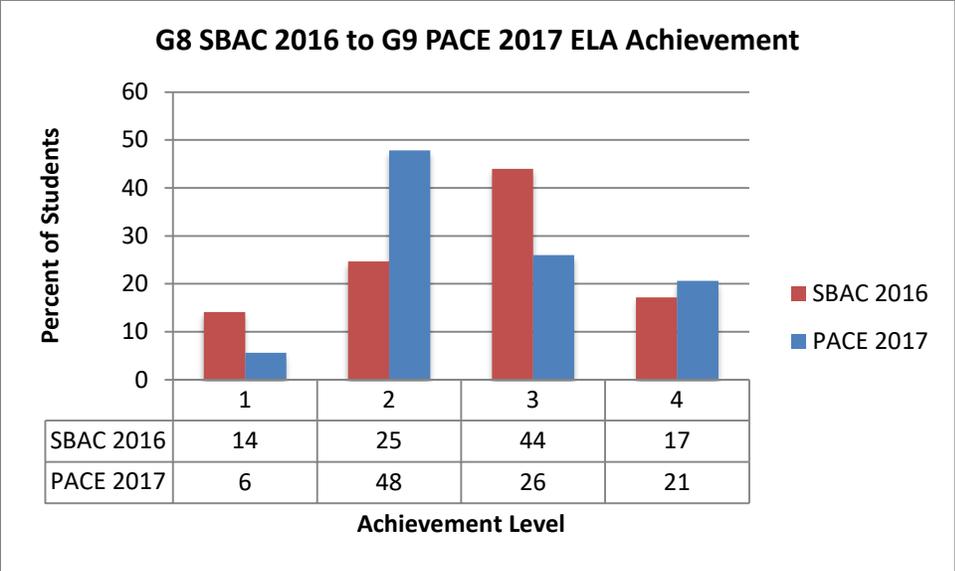


Figure 23. Percent of Students by Achievement Level for Grade 8 ELA SBAC 2016 to Grade 9 ELA PACE 2017

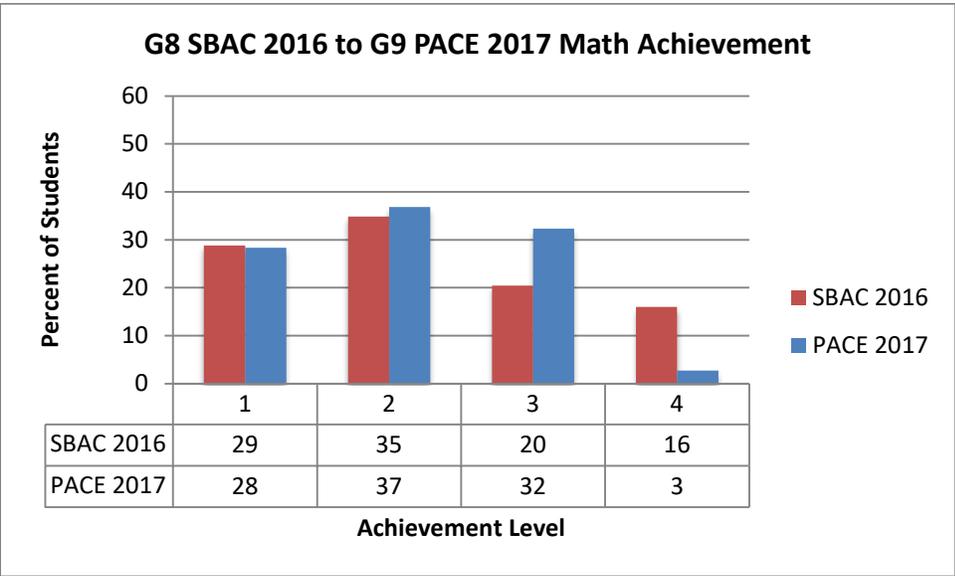


Figure 24. Percent of Students by Achievement Level for Grade 8 Math SBAC 2016 to Grade 9 Math PACE 2017

		PACE 2017			
		1	2	3	4
SBAC 2016	1	31	104	30	0
	2	13	103	95	1
	3	0	52	143	9
	4	1	18	153	33

Table 33. *Grade 3 ELA SBAC 2016 to Grade 4 ELA PACE 2017 Crosstabs (n counts) by Achievement Level (N=786)*

		PACE 2017			
		1	2	3	4
SBAC 2016	1	39	57	19	0
	2	22	110	147	4
	3	2	31	164	42
	4	1	7	76	65

Table 34. *Grade 4 Math SBAC 2016 to Grade 5 Math PACE 2017 Crosstabs (n counts) by Achievement Level (N=786)*

		PACE 2017			
		1	2	3	4
SBAC 2016	1	19	69	5	10
	2	16	123	22	20
	3	6	139	118	59
	4	0	19	45	62

Table 35. *Grade 8 ELA SBAC 2016 to Grade 9 ELA PACE 2017 Crosstabs (n counts) by Achievement Level (N=732)*

		PACE 2017			
		1	2	3	4
SBAC 2016	1	78	55	27	0
	2	56	79	55	3
	3	19	44	44	6
	4	4	26	53	6

Table 36. *Grade 8 Math SBAC 2016 to Grade 9 Math PACE 2017 Crosstabs (n counts) by Achievement Level (N=555)*

Table 37 aggregates the crosstabs above showing the percentage of exact agreement and percentage of exact or adjacent agreement by grade and subject area across the assessment systems from SBAC 2016 to PACE 2017. Importantly, while there is variation across the two assessment programs over two years, the degree of agreement is high across years ranging from 86-96% exact or adjacent agreement. The correlations between the two assessment programs across years are $r=0.505$ ($p<.001$) for ELA and $r=0.537$ for math ($p<.001$). The strength of the correlations between SBAC 2016 and PACE 2017 are quite high given the intentional differences in design and purpose. Also, these analyses assume that students did not change their performance levels across years when, in fact, we know that not to be true

	%Exact Agreement	%Exact or Adjacent Agreement
G3/G4 ELA	39.4%	93.6%
G4/G5 Math	48.1%	95.8%
G8/G9 ELA	44.0%	91.8%
G8/G9 Math	37.3%	85.8%

Table 37. Percent Agreement Across SBAC 2016 to PACE 2017

As was done with the concurrent comparability analyses, the 2x2 classification tables are provided in Tables 38-41. “Classification accuracy” refers to the percentage of students who received the same proficiency classification (i.e., ‘proficient’ or ‘not proficient’) across the two years. In this case, classification accuracy may be a misnomer since students can and do legitimately change in their classifications across years. In fact, schools are purposefully trying to improve the performance of students across years.

		G4 ELA PACE 2017 (%Prof+)	
		No	Yes
G3 ELA SBAC 2016 (%Prof+)	No	31.9%	16.0%
	Yes	9.0%	43.0%

Table 38. Grade 3 ELA SBAC 2016 to Grade 4 ELA PACE 2017 Classification Accuracy (N=786)

		G5 Math PACE 2017 (%Prof+)	
		No	Yes
G4 Math SBAC 2016 (%Prof+)	No	29.0%	21.6%
	Yes	5.2%	44.1%

Table 39. Grade 4 Math SBAC 2016 to Grade 5 Math PACE 2017 Classification Accuracy (N=786)

		G9 ELA PACE 2017 (%Prof+)	
		No	Yes
G8 ELA SBAC 2016 (%Prof+)	No	31.0%	7.8%
	Yes	22.4%	38.8%

Table 40. *Grade 8 ELA SBAC 2016 to Grade 9 ELA PACE 2017 Classification Accuracy (N=732)*

		G9 Math PACE 2017 (%Prof+)	
		No	Yes
G8 Math SBAC 2016 (%Prof+)	No	48.3%	15.3%
	Yes	16.8%	19.6%

Table 41. *Grade 8 Math SBAC 2016 to Grade 9 Math PACE 2017 Classification Accuracy (N=555)*

As would be expected, the classification accuracies across years are slightly lower than the classification accuracies observed for the concurrent year comparisons, ranging from 67.9% to 74.9%. The two elementary grades and subject areas (grade 3 ELA and grade 4 math) have the lowest percentage of students out of the four cells that move from proficient (“Yes”) in 2016 to not proficient (“No”) in 2017. This fits with what we would expect to see for a cohort across years—students either staying within the same cell or moving from non-proficient to proficient. It is unclear why the grade 8 to grade 9 classification accuracies has a higher percentage of students moving from proficient in one year to not proficient in the next year. There does not appear to be a consistent pattern in achievement changes across subject areas. Since this pattern was not observed in the concurrent analyses and it is something we will continue to closely monitor in the coming years.

PACE 2016 to SBAC 2017

The second analysis compares last years’ performance on PACE in grade 3 math and grade 7 ELA and math with this years’ performance on SBAC for students in grade 4 math and grade 8 ELA and math. Only students with a PACE achievement level in 2016 and an SBAC achievement level in 2017 are used for these analyses (N=2,344). Figure 25 shows the percent proficient or above for the matched cohort of students across years. The red bars indicate SBAC and the blue bars represent PACE. In one out of the three grades and subject areas, the percent proficient rose from PACE 2016 to SBAC 2017 and in the other two grades and subject areas the percent proficient either went down or stayed about the same, once again indicating that PACE is at least as rigorous if not more so compared to Smarter Balanced.

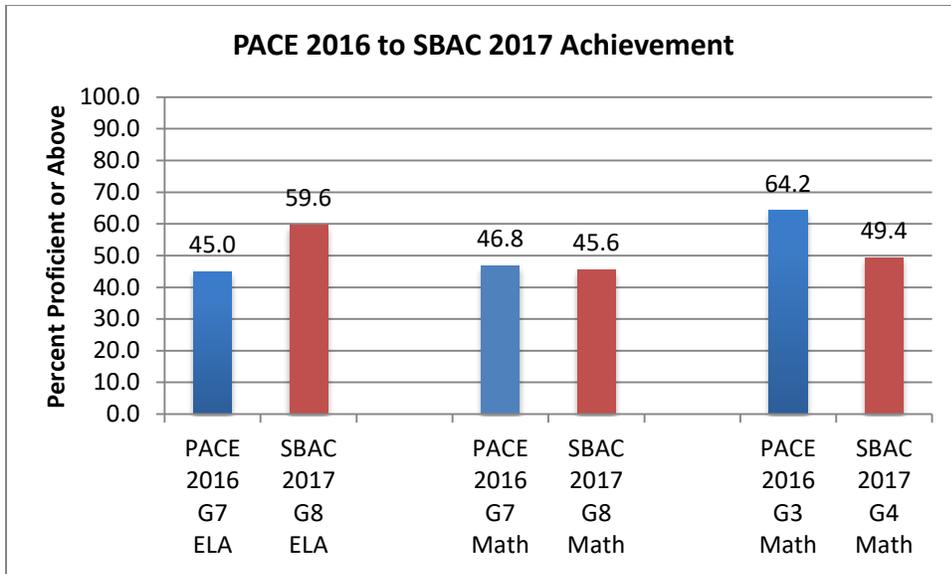


Figure 25. Cohort Percent Proficient or Above across PACE 2016 to SBAC 2017

Figures 26-28 display the achievement level distributions for PACE 2016 and SBAC 2017 by grade level and subject area followed by Tables 42-44 that provide the number of students included in the figures. There does not appear to be any common pattern with regards to changes in achievement levels across years by subject area or grade level, again indicating similar levels of expectations.

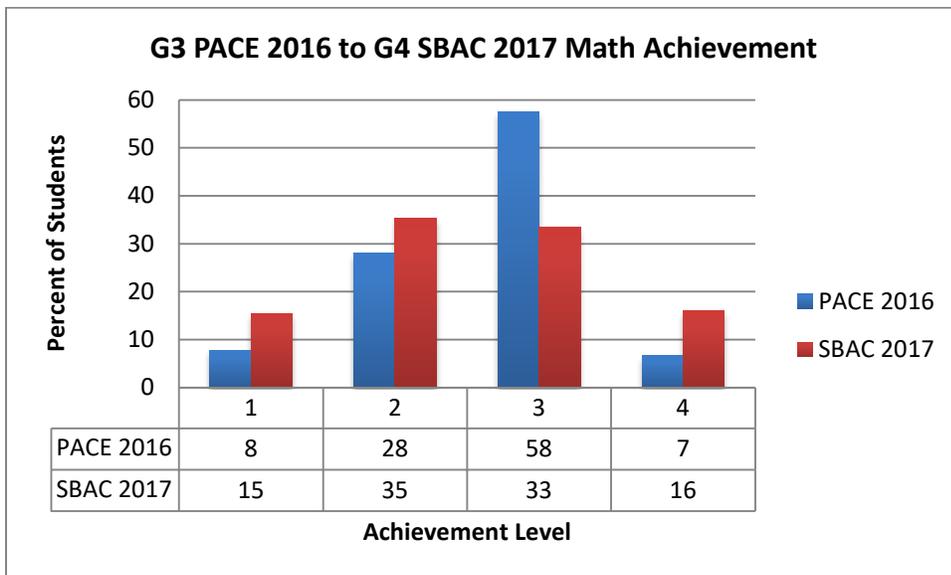


Figure 26. Percent of Students by Achievement Level for Grade 3 Math PACE 2016 to Grade 4 Math SBAC 2017

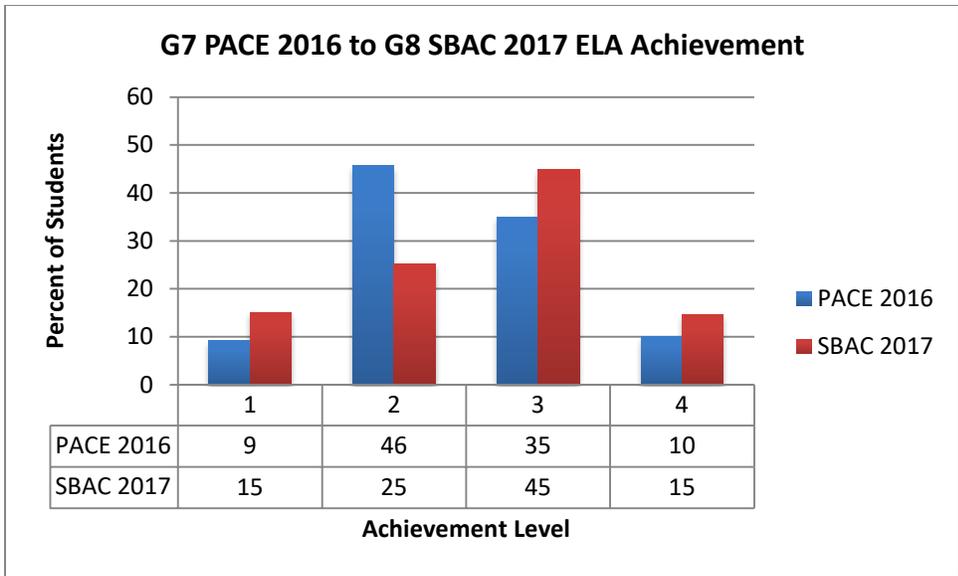


Figure 27. Percent of Students by Achievement Level for Grade 7 ELA PACE 2016 to Grade 8 ELA SBAC 2017

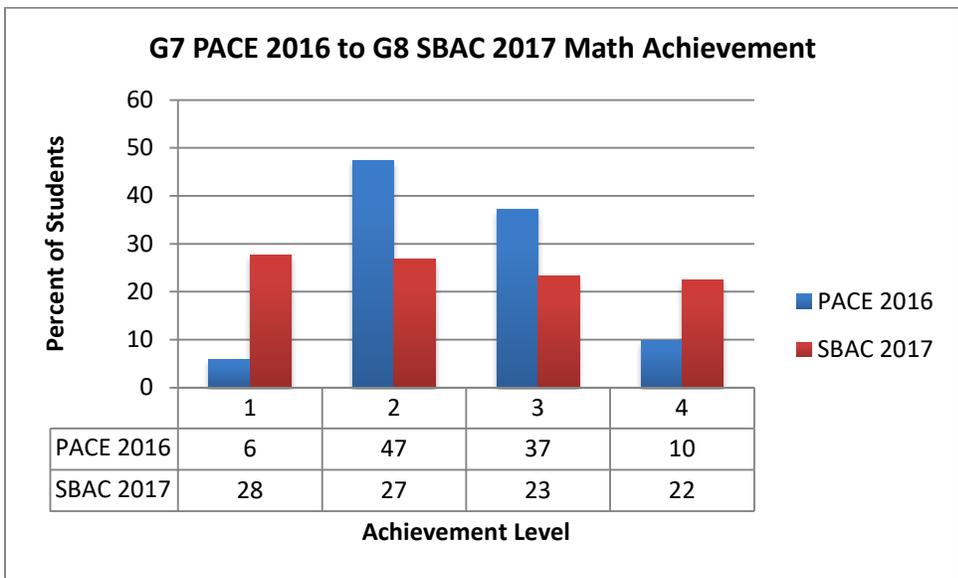


Figure 28. Percent of Students by Achievement Level for Grade 7 Math PACE 2016 to Grade 8 Math SBAC 2017

		SBAC 2017			
		1	2	3	4
PACE 2016	1	40	18	3	0
	2	59	112	45	2
	3	21	141	189	97
	4	0	3	23	26

Table 42. *Grade 3 Math PACE 2016 to Grade 4 Math SBAC 2017 Crosstabs (n counts) by Achievement Level (N=779)*

		SBAC 2017			
		1	2	3	4
PACE 2016	1	38	21	12	2
	2	74	125	147	15
	3	6	49	156	65
	4	1	5	39	34

Table 43. *Grade 7 ELA PACE 2016 to Grade 8 ELA SBAC 2017 Crosstabs (n counts) by Achievement Level (N=789)*

		SBAC 2017			
		1	2	3	4
PACE 2016	1	31	12	1	1
	2	165	123	62	18
	3	18	68	99	103
	4	1	4	18	52

Table 44. *Grade 7 Math PACE 2016 to Grade 8 Math SBAC 2017 Crosstabs (n counts) by Achievement Level (N=776)*

Table 45 aggregates the crosstabs above showing the percentage of exact agreement and percentage of exact or adjacent agreement by grade and subject area across the assessment systems from PACE 2016 to SBAC 2017. The degree of agreement is high across years ranging from 94.5% to 96.3% exact or adjacent agreement. The correlations between the two assessment programs across years are $r=0.523$ ($p<.001$) for ELA and $r=0.597$ for math ($p<.001$). As mentioned previously, given the fact that no assessment is likely to correlate more highly with a different assessment than with itself, the strength of the correlations between PACE 2016 and SBAC 2017 are remarkably high.

	%Exact Agreement	%Exact or Adjacent Agreement
G3/G4 Math	47.1%	96.3%
G7/G8 ELA	44.7%	94.8%
G7/G8 Math	39.3%	94.5%

Table 45. *Percent Agreement Across PACE 2016 to SBAC 2017*

The 2x2 classification tables for PACE 2016 to SBAC 2017 are provided in Tables 46-48.

Again, classification accuracy may be a misnomer since students can and do legitimately change their performance levels across years.

		G4 Math SBAC 2017 (%Prof+)	
		No	Yes
G3 Math PACE 2016 (%Prof+)	No	29.4%	6.4%
	Yes	21.2%	43.0%

Table 46. *Grade 3 Math PACE 2016 to Grade 4 Math SBAC 2017 Classification Accuracy (N=779)*

		G8 ELA SBAC 2017 (%Prof+)	
		No	Yes
G7 ELA PACE 2016 (%Prof+)	No	32.7%	22.3%
	Yes	7.7%	37.3%

Table 47. *Grade 7 ELA PACE 2016 to Grade 8 ELA SBAC 2017 Classification Accuracy (N=789)*

		G8 Math SBAC 2017 (%Prof+)	
		No	Yes
G7 Math PACE 2016 (%Prof+)	No	42.7%	10.6%
	Yes	11.7%	35.1%

Table 48. *Grade 7 Math PACE 2016 to Grade 8 Math SBAC 2017 Classification Accuracy (N=776)*

The classification accuracies across years are about the same as the classification accuracies observed for the concurrent year comparisons, ranging from 70.0% to 77.7%. Only grade 3 math PACE 2016 to grade 4 math SBAC 2017 had a relatively higher percentage of students that moved from proficient to not proficient across years. There does not appear to be a consistent pattern in achievement changes across subject areas. Again, since this pattern was not observed in the concurrent analyses and it is certainly something we will continue to closely monitor in the coming years.

Table 49 shows the proficiency classification accuracies for the waiver-reported subgroups for both cross-year analyses: PACE 2016 to SBAC 2017 and SBAC 2016 to PACE 2017. These statistics are disaggregated by subject but not by grade level in order to increase the likelihood of having cell sizes large enough to report. As with the concurrent analyses, the classification accuracies of the subgroups do not seem to vary greatly from the overall observed classification accuracies. The only subgroup with a proficiency classification accuracy of less than 60% is students who are classified as two or more races (non-Hispanic) in SBAC 2016 to PACE 2017 math. We will pay particular attention to this subgroup in next year’s analyses to ensure this is not indicative of something systemic.

	PACE 2016 to SBAC 2017		SBAC 2016 to PACE 2017	
	ELA	Math	ELA	Math
American Indian or Alaskan Native	**	**	**	**
Asian	75.0%	81.0%	82.5%	70.6%
Black or African American	75.9%	77.6%	76.3%	63.6%
Hispanic or Latino	68.4%	64.9%	73.8%	78.8%
Native Hawaiian or Pacific Islander	**	**	**	**
Two or more races (non-Hispanic)	**	75.0%	66.7%	50.0%
White	69.4%	75.1%	72.0%	71.2%
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**
WaiverSubgroup - SWD and EconDis - Not EL	87.3%	84.6%	79.6%	74.0%
WaiverSubgroup - SWD and EconDis and EL	**	**	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	80.0%	79.9%	79.9%	71.2%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	**	84.6%	81.8%	68.9%
WaiverSubgroup - EconDis and EL - Not SWD	**	80.0%	84.0%	61.8%
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	76.3%	73.7%	73.0%	69.8%

Table 49. Proficiency Classification Accuracies for Subgroups by Non-Concurrent Validity Analysis