

Table of Contents

A Framework for Evaluating the Technical Quality of PACE	3
Is the PACE Assessment System Valid?	5
Overview of the NH PACE System	6
Communication with USED.....	8
2016-2017 Waiver Approval Letter.....	8
December 2016 Update to USDOE on Criteria for Success and Milestones	10
2015-2016 Student Performance and Participation Results	15
NH PACE Theory of Action	18
Building Local Capacity.....	21
Three-Tiered System.....	21
Tier 1 Capacity Building.....	23
Summary	26
Comparability-Based Framework for Validating the System of Assessments	28
Overview of Validity Evidence for the NH PACE System	28
Within-District Comparability in Expectations for Student Performance	31
Cross-District Comparability In Expectations of Student Performance	51
Comparability of Annual Determinations across Assessment Systems	44
Summary	65
External Evaluation of System Success	66
HumRRO Executive Summary Report (2016-2017).....	66
Effects of PACE on 8 th Grade Student Achievement Outcomes (2014-2016).....	72
Appendix A: NH PACE Readiness Tool.....	73
Appendix B: NH PACE Task Development Template.....	74
Appendix C: Think Aloud Protocol	76
Appendix D: High Quality Assessment Review Tool	79
Appendix E: Summary of 2016 PACE Common Task Review	84
Appendix F: Scaffolding Brief.....	87
Appendix G: NH PACE 2016-2017 Data Collection Protocols.....	90
Appendix H: Grade 3 ELA ALDs, PACE to SBAC Map	99



A Framework for Evaluating the Technical Quality of PACE

This technical manual provides comprehensive and detailed evidence in support of the validity of the NH PACE Assessment and Accountability System. Validity refers to the accuracy and defensibility of the inferences drawn from the assessment scores about what students know and can do and the appropriateness of the assessment results for their intended uses. This manual focuses on validity related to annual determinations of student proficiency in English language arts and mathematics in grades 3-8 and high school when those determinations are not made using a standardized achievement test. The demonstration and evaluation of validity is an ongoing process; it is not a simply yes/no answer. The collection of validity evidence provided in this technical manual is from the first three years of the NH PACE pilot (2014-15, 2015-16, and 2016-17 school years).

Many different reports, briefs, documentation, and resources were gathered to create this technical manual. The text of those materials was left unchanged except to remove redundancies. The particular audience is noted if necessary for interpreting the tone used in the material.

The *Standards for Educational and Psychological Testing*¹, hereafter referred to as the *Standards*, was used as the foundation for developing the necessary validity evidence. The *Standards* is the authoritative document in educational measurement for evaluating the technical quality of tests and other measurement tools. Assuming the NH DOE applies to the U.S. Department of Education (USED) for an innovative assessment pilot as part of the Section 1204 Demonstration Authority under the Every Student Succeeds Act (ESSA), it will need to demonstrate how its system will meet assessment quality requirements. The innovative pilot eventually will be subject to a peer review process outlined in Section 1111 of ESSA². The assessment quality criteria outlined in the peer review guidance closely mirror the expectations of the *Standards*. Specific elements of technical quality that are included in the NH PACE system include the following:

- ✓ **Alignment** to the full breadth and depth of the state academic content standards.
- ✓ **Validity** or accuracy of the inferences drawn from the assessment scores about what students know and can do and the appropriateness of the assessment results for their intended uses.

¹ American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Tests*. Washington, DC: AERA.

² There are two standards for demonstrating technical quality of the assessment system in the statute for the innovative assessment pilot. Initially, states will have to demonstrate in their application how their system will meet the technical quality requirements outlined in Section 1111 for all state assessment. At the end of the Demonstration Authority, states will have to demonstrate how their implemented system meets all of the technical quality requirements outlined in Section 1111.



- ✓ **Reliability** or consistency of the scoring tools and the generalizability of the inferences about students’ knowledge and skills.
- ✓ **Comparability** of the assessment results for students within the pilot districts and, while the system is not yet statewide, across pilot and non-pilot districts.
- ✓ **Fairness** of the assessments with regard to accessibility for all students and minimizing bias.

In addition, characteristics of high-quality assessments and assessment systems were used in the design phase of the NH PACE system to support the efficacy of inferences made about student, teacher, school, and district performance. The NH PACE system is not simply a collection of assessment experiences for students, but instead a coherent system that has a planned flow for how information resulting from different assessments will work together to support the intended interpretations and uses. For example, the NH PACE assessment system is *comprehensive*, *coherent*, and *continuous*. These concepts of a high quality assessment system are not new, but are drawn from the National Research Council’s *Knowing What Students Know*³ and can be reviewed in greater detail from that resource or from a recent discussion of assessment system design⁴.

Comprehensive –The NH PACE system includes a range of measurement approaches “to provide a variety of evidence to support educational decision making”⁵. In this way, it is comprehensive because it allows students to demonstrate their competency in a variety of ways. This helps to ensure the validity and fairness of the inferences drawn from the assessments. Comprehensiveness also means that the assessment system, as a whole, reflects the breadth and depth of college and career ready standards and learning practices adopted by the state.

Coherence – This component of the NH PACE system is intricately linked with its theory of action. The NH PACE pilot is not simply a different form of assessment and accountability, but reflects a systemic educational approach to promote deeper and more meaningful learning for students. Thus coherence refers to assessments compatible with the methods of teaching and learning and to the underlying model of learning.

Continuity – Finally, the NH PACE system measures student learning over time. This element of an assessment system ensures that student progress can be monitored so that educators can make appropriate instructional decisions for students.

³ Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

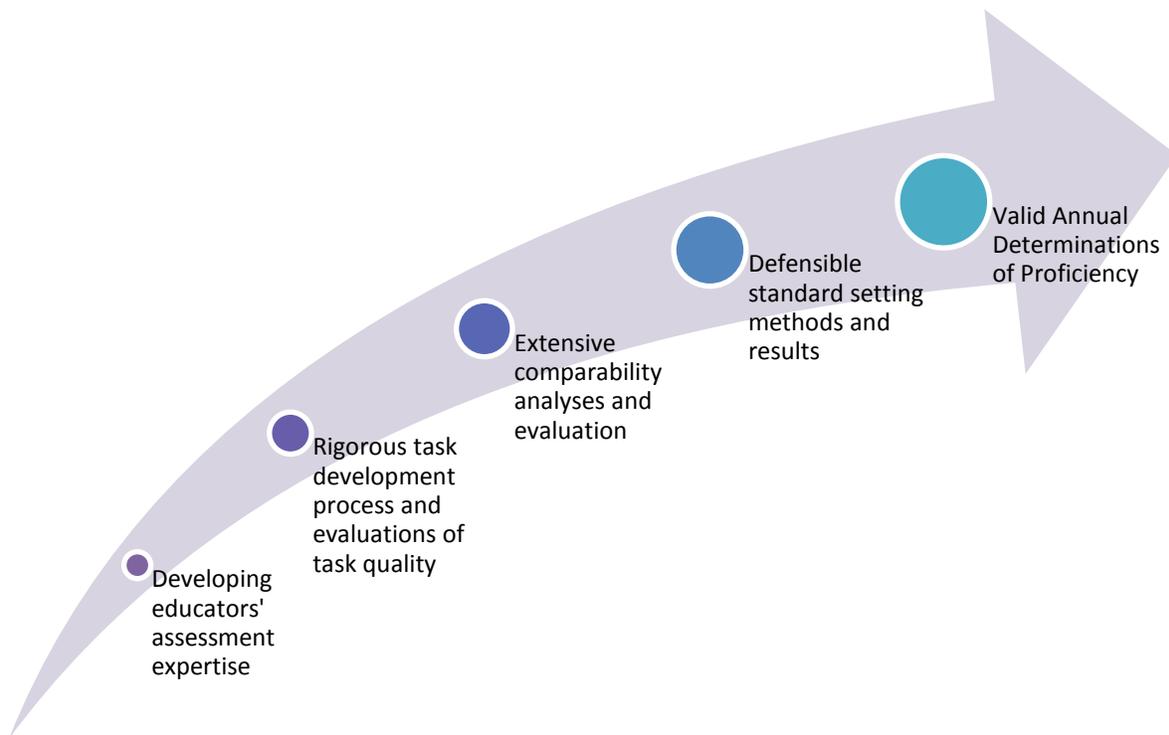
⁴ Chattergoon, R., & Marion, S. F. (2016). Not as easy as it sounds: Designing a balanced assessment system. *National Association of State Boards of Education*, 16(1), 6–9.

⁵ Pellegrino, et al., 2001, p. 253.



Is the PACE Assessment System Valid?

As explained above, validity is not a true/false question. Rather, validity involves marshalling evidence and logic to evaluate the extent to which the intended interpretations and uses of assessment scores are supported. In many ways, a validity evaluation is analogous to the way that an attorney builds a civil case to convince the jury that a preponderance of evidence supports the plaintiff's or defendant's claims. The extensive presentation of technical and logical evidence in this report builds a validity argument for PACE. Like a good argument, the evidence presentation follows a story from the theory of action or the logic model guiding PACE to ultimate claims supporting the defensibility of the PACE annual determinations of proficiency. The following graphic highlights the key aspects of the evidence presented later in this report that helps weave together the validity argument in support of the PACE assessment system. The documentation presented in this report is a "preponderance of evidence" supporting the validity of inferences from PACE assessments for the intended uses in the PACE system.



Overview of the NH PACE System

New Hampshire was awarded permission from the U.S. Department of Education in March 2015 to pilot an accountability system designed to support deeper learning for students and powerful organization change for schools and districts. The accountability pilot, referred to as Performance Assessment of Competency Education or PACE, is grounded in a competency-based educational approach designed to ensure that students have meaningful opportunities to achieve critical knowledge and skills. NH PACE began as a two-year pilot (2014-2016) and was granted an additional one-year waiver (2016-2017).

The core of the NH PACE assessment and accountability system is locally-developed, locally-administered performance assessments tied to grade and course competencies determined by local school districts. Additionally, in each grade and subject without a state assessment (a total of 17 subjects and grades), one, common complex performance task called the PACE Common Task is administered by all participating districts. The PACE Common Task is NOT a state test! Rather, it is developed collaboratively among the participating districts and is used to ensure that each teacher's evaluation of student performance is comparable to the evaluations made by other teachers. Finally, Smarter Balanced is administered in grade 3 (English language arts), 4 (math), and grade 8 for both ELA and math. The SAT is administered to all grade 11 students.

In a competency-based system, students' opportunities are judged by the outcomes they achieve and not by "inputs" such as seat time. Therefore, students must achieve identified learning targets before moving on to the next goals and/or graduating from high school. If they do not, school districts are expected to work with families to support additional learning opportunities to ensure that students have legitimate opportunities to master the necessary knowledge and skills.

High-quality performance assessments play a crucial role in the NH PACE system because of the need to measure the depth of student understanding of these complex learning targets. Performance assessments are used both to inform teachers and students of how the learning activities are working and what might need to be adjusted (formative) along with serving to help document what students have actually learned (summative).

Participating School Districts for 2016-2017 include: **Sanborn** Regional School District, **Rochester** School District, **Epping** School District, **Souhegan** School District, **Concord** School District, **Pittsfield** School District, **Seacoast Charter** School, **Monroe** School District, and the **White Mountains** School District. Instructional and assessment accommodations are available for students with disabilities as well as students for whom English is not their native language. A fundamental value of NH PACE is that the system should be designed to maximize the learning opportunities for each individual student.



Several other school districts are currently building their capacity to become fully participating NH PACE districts in subsequent years. Districts must demonstrate readiness to participate in the pilot and must make certain commitments to continue with the pilot. This process is described in more detail under the “Building Local Capacity” section.

One of the New Hampshire PACE team’s major areas of focus is developing a process and the capacity to scale such efforts to all those NH schools and districts that elect to participate in PACE. The current NH PACE accountability system is based on a voluntary proof of concept pilot. The NH DOE is committed to supporting the development of local leadership and capacity to help low performing schools implement the NH PACE system with fidelity.



Communication with USED

2016-2017 Waiver Approval Letter



UNITED STATES DEPARTMENT OF EDUCATION

OFFICE OF ELEMENTARY AND SECONDARY EDUCATION

OCT 06 2016

The Honorable Virginia M. Barry
Commissioner of Education
New Hampshire Department of Education
101 Pleasant Street
Concord, NH 03301

Dear Commissioner Barry:

I am writing in response to the New Hampshire Department of Education's (NHDOE) request submitted on August 5, 2016, along with the clarifying e-mail sent on August 19, 2016, to extend the waiver allowing implementation of the New Hampshire Performance Assessment of Competency Education (PACE) Pilot. On March 5, 2015, the U.S. Department of Education (Department) approved a waiver, subject to certain conditions and commitments, to permit NHDOE to pilot the PACE assessment system, a system of locally developed, competency-based assessments, in four local educational agencies (LEAs) in 2014-2015 and eight LEAs in 2015-2016, in place of the statewide assessments administered consistent with section 1111(b)(2) of the Elementary and Secondary Education Act (ESEA), as amended by the No Child Left Behind Act (NCLB) of 2001.

I am pleased to grant, pursuant to my authority under section 8401 of the Elementary and Secondary Education Act of 1965 (ESEA), as amended by the Every Student Succeeds Act (ESSA), a one-year extension of the waiver, for the 2016-2017 school year, for implementation of the PACE pilot and expansion to include nine LEAs and 36 schools in total within those LEAs, of the following statutory requirements under Title I, Part A of the ESEA and their associated regulatory provisions: sections 1111(b)(1)(B) and 1111(b)(3)(C)(i) of the ESEA, as amended by NCLB, which require a State educational agency to use the same academic achievement standards and assessments, respectively, for all public school children in the State. NHDOE requested this extension so that the students in the nine pilot LEAs may take PACE Pilot assessments in reading/language arts, mathematics, and science in lieu of the statewide assessment in certain grades in 2016-2017. The LEAs will administer the NHDOE State assessments in reading/language arts and mathematics once each in elementary, middle, and high school and will administer PACE Pilot assessments in every other grade in which assessments are required under Title I and NHDOE will ensure comparable achievement levels across all LEAs in the State. This waiver extension is granted because NHDOE sufficiently demonstrated that the innovative assessment system being developed will advance student academic achievement, continue to provide assistance to the populations participating in the PACE Pilot, and maintain or improve transparency in reporting to parents and the public on student achievement and school performance, as required in section 8401(b)(1) of the ESEA, as amended by the ESSA.

This waiver is granted to NHDOE subject to the following condition:

New Hampshire will submit a report on November 15, 2016, and October 1, 2017, on implementation in the previous school year, 2015-2016 and 2016-2017, respectively, that includes at least the following information:

- Overview of the PACE Pilot and results, including:



- o Achievement results and participation rates for each PACE LEA by subject and grade for the "all students" group and each subgroup;
- o Summary of implementation, including status on the State's criteria for success and identified milestone (as developed by the State under the initial waiver permitting NHDOE to pilot the PACE assessment system); and
- o Copies of key documents supporting PACE Pilot implementation (e.g., PACE guidebook, PACE manual for LEAs, task development template, and task review tool) and results from the State's independent evaluation of the PACE Pilot that is currently underway.
- Results of the analyses NHDOE identified in its extension request, including:
 - o Findings and analysis from the State's review of local assessment map;
 - o Summary of the State's review of the technical quality of the common performance tasks;
 - o Findings and analyses from the State's audit of local assessments, which include a review of both locally developed performance task and "unit summative assessments";
 - o Summary and results of the State's local task review, for which all PACE Pilot LEAs will submit their summative non-"common" locally developed performance tasks for review;
 - o Results of the State's planned comparability analyses, including comparisons of individual student-level results on the Statewide assessments taken the prior year and the PACE Pilot assessments taken the current year (e.g., mathematics grade 4 Smarter Balanced in 2014-2015 and grade 5 PACE Pilot in 2015-2016), and individual student-level results on PACE Pilot assessments taken the prior year and the Statewide assessments taken the following year (e.g., mathematics grade 3 PACE Pilot in 2014-2015 and grade 4 Smarter Balanced in 2015-2016). The State's analyses must also include evidence of the reliability of scoring for the PACE assessment; and
 - o Results of other implementation analyses of the PACE Pilot planned by NHDOE, such as the generalizability study.

I note that on July 11, 2016, the Department published a notice of proposed rulemaking (NPRM) regarding the innovative Assessment Demonstration Authority authorized in Title I, Part B of the ESEA, as amended by the ESSA. We are in the process of reviewing the comments submitted on the NPRM and revising the regulations, as necessary. These regulations will govern any future requests to implement innovative assessments in New Hampshire.

This Letter and the State's final renewal application will be posted on our website. Thank you for your continued commitment to enhancing education in New Hampshire, including your focus on developing new and innovative assessments to improve outcomes for New Hampshire's student. We look forward to continuing to work with you and learn from this pilot effort. If we can provide any further assistance as you implement the PACE Pilot under the waiver, please contact Collette Roney or Tawanda Avery of my staff at: OSS.NewHampshire@ed.gov.

Sincerely,



Ann Whalen

Senior Advisor to the Secretary

Delegated the Duties of Assistant Secretary for Elementary and Secondary Education

cc: Paul Leather, Deputy Commissioner



December 2016 Update to USED on Criteria for Success and Milestones

Commitment and Capacity

The PACE reciprocal accountability model is based upon a clear commitment on the part of the district leadership to continue the hard work of the PACE initiative. Further, the success of the PACE project depends on the capacity and expertise developed among district personnel. We argue that such capacity building is supported by high quality collaboration among participating districts. Because this information has been reported previously, we simply highlight any new information.

Criterion #1: Clear commitment from local educational leaders

Success Indicator: Clear written commitment from each district superintendent and/or charter school/management leader indicating the leader's willingness to commit the time and resources necessary to support successful implementation of PACE.

We have previously reported on the project's success meeting this criterion. Now that PACE has expanded to nine Tier 1 districts, we have continued to receive full commitment from local leaders. As important as the formal commitment, we continue to see this commitment manifest through active participation in leadership and task development meetings.

Criterion #2: Building of cross-district leadership and cross-district collaboration

Success Indicator: Documentation of meaningful participation by all PACE districts in the key activities of the PACE project. These activities include PACE Leads meetings, task development workshops in ELA, math, and science, achievement level descriptor writing, and cross-district calibration. Specifically, meeting this criterion means that all PACE districts are represented in essentially all key PACE activities.

One of the central tenets of the PACE theory of action is that collaborative work among the participating districts is a key mechanism for building capacity and expertise among local district personnel. We have documented the multiple levels of involvement previously, so we do not repeat that information here, except to note the new development of having teachers assume leadership roles in the task development work. There has been an impressive penetration in terms of the numbers of teachers who have been involved in at least some aspect of this kind of collaboration including task development workshops, cross-district comparability studies, and achievement-level descriptor writing. We report later in this document on the new leadership roles for teachers in terms of task development. We argue that this expertise building is a critical component for the sustainability of the project.

Development, Implementation and Scoring of Performance Assessments

NH DOE documented in several previous progress reports that we had successfully developed common PACE performance assessments to allow for the administration of at least one common task in all of the non-Smarter Balanced grades/subjects for both the 2014-2015 and 2015-2016 school years. The following criteria focus on the development, implementation, and scoring of performance assessments, which is really the heart of the pilot.



Criterion #3: Development of high-quality performance assessments

Success Indicator: All PACE common tasks meet the agreed upon criteria for high quality assessment as judged by the Center for Assessment’s performance assessment experts either initially or after revision. Further, the project will develop tools, procedures, and training to facilitate the development of high-quality local assessments.

As we reported in our previous reports, NH DOE’s goal was never to have just one performance assessment in all of the grades and subjects. Rather this one common task was just the beginning. The goal has always been to transform district assessment systems so that they were largely based on tasks that elicited deep thinking from students. The PACE districts had already been moving in this direction. We also reported on the significant number of tasks submitted to the performance task bank.

As discussed elsewhere in this report, PACE has adopted a multi-faceted approach to task development starting with the “over-development” of common performance tasks and carrying through with the submission of local tasks to national expert review from SCALE. The Center for Assessment continues to evaluate all of the common performance using a transparent review process based on the High Quality Assessment Review Tool. A summary of the Center’s reviews is presented later in this report. Briefly, the quality of the common performance tasks continues to improve, in fact, quite dramatically. We have also seen evidence that this improved quality is infusing the quality of locally-developed and administered tasks and assessments.

Criterion #4: Successful implementation of common performance assessments in Year 2

Success Indicator: All PACE Common tasks are implemented as described in the PACE Task Template including the use of appropriate accommodations for students with identified disabilities and English language learners.

As noted above and throughout this document, we have witnessed a notable improvement in the quality of the common performance tasks and we have also improved some administration shortcomings that had been discovered in Year 1. **Meeting this criterion would mean that all PACE tasks were administered as intended with relatively few adjustments required. We are pleased to report that this occurred in 2015-2016 at an even better rate than in 2014-2015.**

Criterion #5: Rates of participation in training and calibration

Success Indicator: All teachers involved in scoring the PACE Common tasks are trained on the general and task-specific training protocols either in person or via digital tools.

As documented in previous reports PACE district leaders used the directions contained in the PACE Calibration Protocol document to guide the training of their teachers and to ensure that all teachers scoring PACE common tasks have received appropriate training before scoring. PACE district leaders have documented that all teachers scoring PACE common tasks received the required training. Most districts did this through face-to-face training with their teachers, but at least one district (Rochester) supplemented the in-person training with web-based videos, created to ensure that a common message was available to all teachers. Based on this terrific idea from this one district, NH DOE and its consultants are planning to create training videos to ensure at



least a baseline of common scorer training across all districts for 2016-2017. This step will not only aid cross-district comparability, but also project sustainability. Finally, since scoring quality is enhanced tremendously with the use of agreed-upon anchor or benchmark papers, a key goal of this year's task development process is to develop a set of shared anchor papers for each task based on the small scale field test being conducted this spring. Having such anchor papers in advance will only enhance the already high-quality scoring.

Criterion #6: Inter-rater agreement within district

Success Indicator: Given the multi-dimensional nature of the PACE tasks, the target for rater consistency is 60 percent exact agreement rate for each dimension.

As documented in previous reports, we relied on the National Assessment of Educational Progress (NAEP) for guidance on appropriate interrater reliability targets. The National Center for Educational Statistics put forth the following for evaluating the rater consistency for constructed response items on NAEP.

Agreement percentages vary significantly across items. On a simple two-point mathematics item, agreement should approach 100 percent. On the other hand, when scoring a complex six-point writing constructed response item, an agreement of 60 percent would be considered an acceptable result⁶.

The PACE Common Performance Assessments are multi-dimensional compared to NAEP's unidimensional, constructed-response items, but PACE tasks are comprised of multiple, four-point rubrics compared with NAEP's single 6-point rubric (for the most complex items). Therefore, we thought that a 60 percent exact agreement rate would be an ambitious, but reasonable target for PACE.

Within district interrater consistency and across district consistency rates are reported extensively later in this document. We continue to be impressed with the level of scoring consistency exhibited both within and across districts.

Criterion #7: Cross-district calibration

Success Indicator: Fifty-four (54) percent exact agreement is the criterion for evaluating the cross-district comparability of rubric score dimensions or that the average rubric score does not deviate from the consensus score by more than 0.5 points on the rubric scale.

Again, using NAEP as the gold standard, NCES indicated that the cross-year rater consistency should be "within 8 percent of the prior year interrater agreement for two- and three-point items and within 10 percent of the prior year interrater agreement for four- to six-point items⁷." In the case of PACE, we are not evaluating (yet) rater agreement across years, but we are using the NAEP cross-year criterion as a proxy for cross-district comparability targets. Applying the guideline of a 10 percent reduction in cross-year compared to within-year rater agreement rates for 6-point rubrics, would result in a 54% agreement target. The evaluation of cross-district

⁶ See: https://nces.ed.gov/nationsreportcard/tdw/analysis/initial_itemscore.aspx.

⁷ https://nces.ed.gov/nationsreportcard/tdw/analysis/initial_itemscore.aspx.



scoring is more like an accuracy determination rather than a measure of consistency because the randomly assigned pairs of raters during the calibration workshops are intended to produce a “truth” score that is then compared to the score awarded by the teacher. This is an even more challenging comparison than the NAEP cross-year consistency comparison, but since these analyses compare average rubric scores for the entire task, we think it is fair to use NAEP’s criterion.

Having a “truth” score is an important first step in these analyses. As we documented in previous reports, the consensus scoring approach required randomly assigned pairs of raters from different districts to review samples of student work and then to assign a consensus (expert) score to that piece of work. We report on the cross-district calibration analyses and results later in this report.

Outcomes

The TAC emphasized that it is premature to use student outcomes as evaluation criteria for the project. Major educational reform efforts such as PACE require time to implement and take hold. Educational reform experts such as Fullan and Hargreaves note that it is not unreasonable to see little impact on student achievement in the first 3-5 years of a major reform. In fact, it is not uncommon to see some performance drops (what Fullan termed an “implementation dip”) early in a reform. That said, PACE district leaders know there is little patience for seeing declines in student performance relative to comparable districts and/or exaggerating equity gaps. Of course, the first step in evaluating student achievement is to ensure that such evaluations are based on comparable annual determinations.

Criterion #8: Produce “comparable” annual determinations

Success Indicators: This criterion relies on two indicators of success. The first, is that no district falls below the 54% exact agreement rate. The second is that the annual determinations across all PACE districts for grades and subjects without Smarter Balanced assessments are similar to the annual determinations produced by Smarter Balanced assessment results.

As noted in previous reports and in ED’s approval letter, being able to produce comparable annual determinations is critical to the project’s success. Also, as noted in our earlier progress reports, PACE annual determinations are based on a combination of local summative assessments tied to district-adopted competencies and the common PACE performance assessments using an examinee-centered judgmental standard-setting method called contrasting groups. Simply conducting a successful standard setting process does not necessarily lead to comparable annual determinations. For that, we need evidence that teachers from different districts evaluate student work similarly. **The cross-district calibration discussed above provides the evidence, at this point in the project, of comparability among districts in terms of evaluating the quality of student work.**

While having a shared understanding regarding the quality of an assessment product is a critical foundation for comparability, it is not enough to ensure annual determinations are comparable. The main outcome of the standard setting study was to produce these comparable outcomes such that if a student in District X is called “proficient” based on an examination of their work, then a student with similar levels of work produced in District Y will also be called proficient. **So**



rather than considering a single assessment, the standard setting methods rely on students' work generated throughout the year.

Methods for ensuring that the annual determinations provide for the same interpretations about what students know and can do across districts requires something in common. Obviously, we have the PACE common task in all grades and subjects and that is certainly one tool we can use. However, the reliability/generalizability shortcomings with any single task have been well-documented, so any judgments of comparability will be limited by the unreliability associated with a single task. However, the Smarter Balanced assessments are administered in multiple grades in all PACE districts. Therefore, we can compare the results from Smarter Balanced and PACE annual determinations to see how they “line up.” For example, we would be concerned if noticeably more or fewer students were classified in Level 3 and 4 on Smarter Balanced compared to PACE. Therefore, we suggest that evaluating the difference between average (across elementary and middle grades) Smarter Balanced and PACE results (students scoring at Levels 3 and 4). **The results presented later in this document indicate that the results comparing PACE and Smarter Balanced/SAT results are even stronger this year than last. It is clear that local districts are holding their students to even more rigorous standards than was the case with last year's impressive results.**

Criterion #9: “No harm” on Smarter Balanced

Success Indicator: Districts will not differ significantly from the “predicted” Smarter Balanced results.

As discussed with ED, we are evaluating the cross-year results as students move from Smarter Balanced and SAT annual determinations to annual determinations based on PACE. We discuss the results of these analyses later in this report.

Criterion #10: Ensuring Equitable Outcomes

Success Indicator: Achievement gaps between the major subgroups in PACE school districts (e.g., economically disadvantaged, students with disabilities) as measured by Smarter Balanced will not increase over time relative to each district's baseline (2015) and compared to statewide trends for the same groups as measured by Smarter Balanced.

As documented in last December's report, the differences in performance among major subgroups and the all students group were similar for both PACE and Smarter Balanced annual determinations. While that information is useful, it is more important to document what happens to these achievement gaps over time. Since the PACE results are not equated from year-to-year, we must rely on Smarter Balanced assessment results to be able to document any trends in achievement gaps. We will document the size of these gaps using an effect-size metric because of the well-known problems with trying to evaluate changes in achievement gaps using more simplistic approaches such as comparing the proportion of students scoring at or above some cutscore. The results of these analyses will be forthcoming in a subsequent report.



2015-2016 Student Performance and Participation Results

The following tables present the results for the PACE assessment system for the 2015-2016 school year. Achievement results are broken down by grade and performance level for ELA, math, and science in Tables 1-3, respectively. Participation by grade level is provided for ELA and math in Table 4. Table 5 contains the achievement results and participation rates for the student subgroups in the eight participating PACE districts.

Table 1.

English Language Arts: 2015-2016 PACE Results by Grade and Level

Grade	Percent at Level 1	Percent at Level 2	Percent at Level 3	Percent at Level 4	Percent at Level 3 & 4
4	8%	40%	40%	12%	52%
5	11%	33%	41%	14%	55%
6	10%	47%	30%	13%	43%
7	11%	46%	33%	9%	42%
9	12%	40%	34%	14%	48%
10	7%	35%	43%	16%	58%
All	14%	33%	38%	14%	53%

Table 2.

Mathematics: 2015-2016 PACE Results by Grade and Level

Grade	Percent at Level 1	Percent at Level 2	Percent at Level 3	Percent at Level 4	Percent at Level 3 & 4
3	9%	27%	57%	7%	64%
5	9%	35%	43%	14%	57%
6	15%	38%	29%	19%	48%
7	8%	47%	35%	9%	44%
9	7%	34%	37%	22%	59%
10	9%	37%	33%	21%	54%
All	15%	37%	35%	14%	48%



Table 3.
Science: 2015-2016 PACE Results by Grade and Level

Grade	Percent at Level 1	Percent at Level 2	Percent at Level 3	Percent at Level 4	Percent at Level 3 & 4
4	2%	47%	30%	21%	51%
8	9%	45%	40%	5%	45%
9	10%	33%	41%	17%	57%
10	10%	46%	32%	12%	44%
All	6%	46%	35%	13%	48%

Table 4.
2015-2016 PACE Participation

Grade	Math	ELA
3	99%	Smarter Balanced
4	Smarter Balanced	92%
5	99%	99%
6	99%	99%
7	97%	97%
All	96%	96%



Table 5.

2015-2016 PACE Results by Subgroup (students are only counted in one (1) category)

Race/Ethnicity	Percent Scoring at Levels 3 & 4			Participation	
	ELA	Math	Science	ELA	Math
Race - American Indian or Alaskan Native (Non-Hispanic)	30%	42%	**	97%	100%
Race - Asian (Non-Hispanic)	53%	54%	45%	96%	98%
Race - Black or African American (Non-Hispanic)	28%	18%	30%	92%	92%
Race - Hispanic	44%	38%	35%	98%	96%
Race - Native Hawaiian or Pacific Islander (Non-Hispanic)	**	**	**	**	**
Race - Two or more races	44%	36%	29%	93%	89%
Race - White (Non-Hispanic)	54%	50%	50%	97%	96%
WaiverSubgroup - EconDis and EL - Not SWD	23%	23%	28%	90%	93%
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	44%	41%	37%	96%	96%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	49%	48%	62%	98%	90%
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	20%	22%	29%	94%	94%
WaiverSubgroup - SWD and EconDis - Not EL	9%	11%	16%	93%	92%
WaiverSubgroup - SWD and EconDis and EL	5%	14%	**	100%	100%
WaiverSubgroup - SWD and EL - Not EconDis	**	**	**	**	**
All	53%	48%	48%	96%	96%



NH PACE Theory of Action

The NH PACE theory of action is grounded in the latest advances related to how students learn⁸, how to assess what students know⁹, and how to foster positive organizational learning and change¹⁰. Figure 1 illustrates a version of the PACE theory of action with system design features on the left to outcomes on the right. The purpose of this theory of action is to demonstrate broadly how implementation of the PACE system is intended to impact the instructional core, thereby advancing college and career readiness. This figure differs from other theories of action depicting the PACE system that explicate how the system is intended to function at a more granular, component-based level (e.g., the PACE theory of action developed by HUMRRO for use in the external, formative evaluation). In its most basic form, the theory of action postulates that system design features drive changes to the instructional core of classroom practices such that teachers will focus on the depth and breadth of key competencies (or content standards). These changes in instruction then lead to improved student achievement outcomes for all students; specifically, that students will be college or career ready.

There are four main system design features with embedded assumptions of how those design features will lead to changes in the instructional core of classroom practices. The first design feature is that local education leaders are explicitly involved in designing and implementing their own accountability system. This fosters positive organizational learning and change by supporting the internal motivation of educators. This is in contrast to all-too-common top-down accountability and extrinsic approaches where the goals and methods of the accountability system are defined at the state or federal levels and districts are simply expected to comply. The second design feature is that local education leaders are provided reciprocal support and capacity building to support their development of key capacities related to designing and implementing the system. This means the NH DOE and its technical partners provide high-quality professional development, training, and support to local districts in the technical, policy, and practical issues related to the system design and implementation. The third design feature is the use of competency-based approaches to learning, instruction, and assessment. These approaches structure learning opportunities for students to gain meaningful knowledge and skills at a depth of understanding that they can transfer to new real-world situations. These approaches also

⁸ Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school (Expanded Edition)*. Washington, DC: National Academy of Sciences.

Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29, 7, 4-14.

⁹ Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

¹⁰ Elmore, R. F. (2004). Moving forward: Refining accountability systems. In Fuhrman, S. H. & Elmore, R. F. *Redesigning accountability systems for education* (pp.276-296). New York, NY: Teachers College Press.

Fullan, M. (2001). *Leading in a culture of change*. San Francisco: Jossey Bass.

Pink, D. H. (2009). *Drive: The surprising truth about what motivates us*. New York, NY: Riverhead Books.



improve student motivation and engagement because they allow students more voice and choice in their own learning. The fourth design feature is the use of locally designed and curriculum-embedded performance assessments throughout the year. These high-quality assessments signal high learning expectations, monitor student learning, and provide specific feedback to teachers and students on their performance relative to the grade and subject competencies. Since these rich, cognitively demanding assessment experiences are curriculum-embedded, teachers can adjust their instruction in real-time to meet students where they are at and help them grow towards proficiency. The PACE Common Task serves as an exemplar for teachers of a high-quality performance assessment, rubric, and scoring protocols and procedures. As more PACE Common Tasks are designed, there will be a bank of high-quality performance tasks and rubrics with anchor papers at different levels of performance to help drive positive instructional changes. The ultimate goal of NH PACE, as seen in the theory of action below, is that student achievement outcomes will improve and that all students will be college or career ready upon graduation from high school.



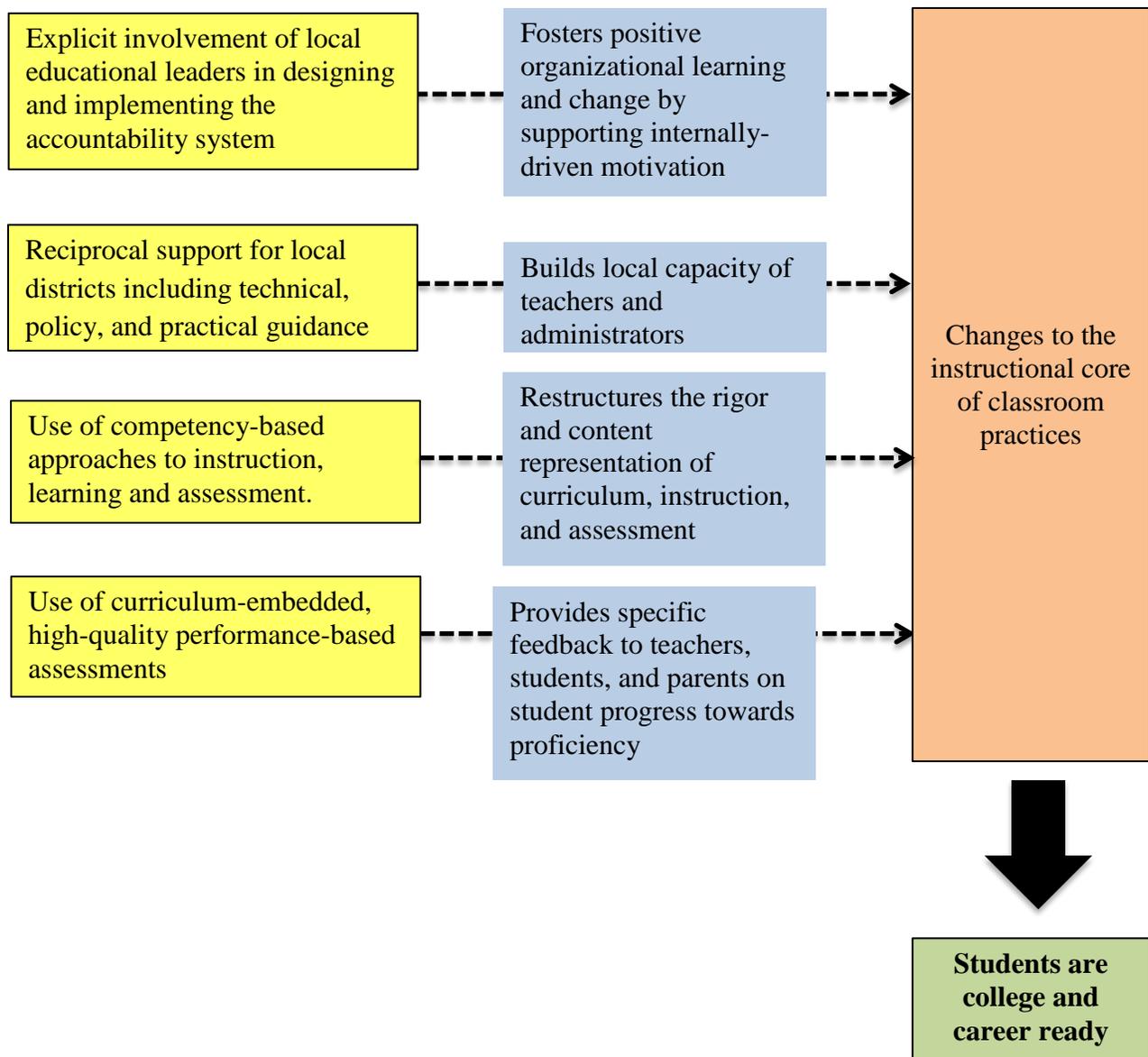


Figure 1. NH PACE Theory of Action



Building Local Capacity

A key premise of the NH PACE theory of action is that local education leaders are supported by NH DOE and each other in creating the expertise necessary to implement the system with fidelity. There are many ways in which the NH PACE pilot builds local capacity both prior to and while implementing the NH PACE system. The following section begins with a detailed description of the three-tiered system that prepares districts with the key capacities to implement the NH PACE system as intended. The section ends with an in-depth discussion about the differentiated Tier 1 professional development and support offered by the NH DOE and its technical partners to local education leaders.

Three-Tiered System

The process for school districts to be accepted for inclusion in the PACE pilot is based on three-tiers of cohorts. Districts are selected for participation in one of the cohorts based on their application to the NH DOE, which includes a readiness tool (Appendix A) related to competency-based education and performance-based assessment. This process allows districts to enter at their current level of preparation and also helps the NH DOE identify areas of professional development and support necessary for districts to become fully implementing PACE districts (Tier 1). This means districts do not have to enter at Tier 3; districts can skip Tiers 2 and 3 completely and just begin implementing as a Tier 1 district—it all depends on their level of readiness.

Table 6 provides specific definitions for each tier and an explanation of the targeted supports offered to districts by the NH DOE for each of the three tiers. Tier 1 districts are those districts that are implementing PACE. Tier 1 districts have reported implementing competency-based education in classrooms and have some experience and capacity with performance assessments of competencies. The state provides targeted assistance to districts to help them move toward Tier 1; however, districts ultimately decide when they are “ready” to move into Tier 1. Tier 2 includes districts that report at least course level or school-wide competencies in place, but do not have a lot of experience with performance assessments. Tier 3 districts are at the “less advanced” development stage in terms of competency-based education and performance assessment and need more targeted assistance and support. In general, Tier 3 includes districts that report limited competency-based learning environments, do not implement competencies at the classroom level, and have no background with performance assessments.



Table 6.

NH PACE Three-Tiered System and Targeted Support Provided by the NH DOE

	Definition	Targeted Support Provided by NHDOE to the District
Tier 1	Districts that are implementing the PACE pilot have reported implementation of local competencies in school-wide and classroom settings, and some experience with performance assessment in a competency-based learning environment. Evidenced a commitment to transitioning to implementing performance assessment of competencies for accountability purposes district-wide (K-12), and have articulated a beginning plan of how to best accomplish that transition in their community.	<p>The district Superintendent and PACE team leader will have the opportunity to meet monthly with PACE state-level leadership for policy and project management discussions.</p> <p>Access to workshop days throughout the year facilitated by experts, consultants, and coaches allowing cross-school learning of performance assessments within specific content areas and across grade-spans that support curriculum-embedded competency-based task design for formative and summative assessment purposes, scoring, and calibration.</p> <p>Coaching and guidance from experts in the development and implementation of common performance assessment tasks for accountability, based on readiness.</p>
Tier 2	Districts that have reported to have course level and school-wide competencies in place and have at least some implementation of competencies in classroom settings. Competency-based learning environments may be evidence in some places in the district. Experience with task-based performance assessment for competency attainment may be limited to extended learning opportunities or may not have been attempted in any systematic way.	<p>Access to intense Quality Performance Assessment (QPA) training.</p> <p>Access to professional development from state and national experts on performance assessment literacy, beginning levels of performance task development, depth of knowledge levels, how to analyze at student work, reliable scoring, and local structures such as professional learning communities. Districts are also introduced to the NH PACE implementation protocols.</p>
Tier 3	Districts that have reported no or few local active competency based learning environments, do not implement the competencies at the classroom level with students (though they may or may not have written competencies), and have no background experience with performance assessment of competencies.	<p>Access to school-level coaching from New Hampshire Learning Initiative (NHLI)-contracted expert consultants on the topics of developing and implementing competencies and working with the state model competencies.</p> <p>Planning activities with other Tier 3 districts to prepare for greater involvement in performance assessment district-wide.</p>



Tier 1 Capacity Building

There is differentiated training and support offered to Tier 1 districts depending upon when they started implementing the PACE system. This allows professional development to be targeted to the specific needs of incoming local education leaders and more advanced professional development to be offered to local education leaders that have been implementing for a few years. The training is described below in three categories: high-quality performance task development training; advanced teacher leader training; and Summer Institute training and professional activities.

High-Quality Performance Task Development Training

The development of high-quality PACE Common Tasks is grounded in the training of local educators. Teams of teacher leaders from all NH PACE districts who receive advanced assessment coaching are responsible for leading much of the task development work with their fellow teachers. Teachers from all NH PACE districts collaborate in grade and subject area teams and follow a disciplined process of task development. Figure 2 illustrates the PACE Common Task development and pilot-testing process.

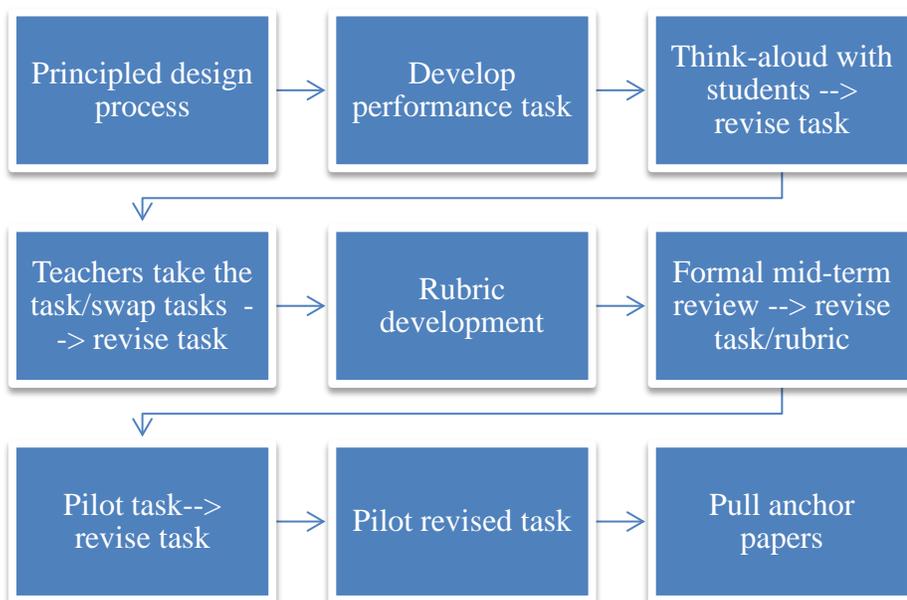


Figure 2. PACE Common Task Development and Pilot-Testing Process

The process begins with a **principled assessment design process**, which means the task is developed based on 1) what students should know and at what depth of knowledge, 2) what evidence is necessary to demonstrate that the student has the desired knowledge, and 3) what tasks will allow students to demonstrate and communicate the desired knowledge. A “backward design” model¹¹ template is used to provide guidance on the characteristics of a high-quality task

¹¹ Wiggins, G., & McTighe, J. (2005). *Understanding by design: Expanded 2nd edition*. New York, NY: Pearson.



and PACE expectations (see Appendix B). This template is used by educators to initially **develop multiple performance tasks** for each grade and subject area, which are designed to provide data on how students are progressing toward the NH competencies for English language arts, math, and science. In addition to the performance templates, there are a number of supports available to teachers regarding high quality task development, including a scaffolding brief that outlines appropriate levels of scaffolding within tasks to ensure the performance assessments are true measures of what students know and are able to do *independently* (Appendix F). Eventually one PACE Common Task is chosen to implement in each grade and subject area the following school year.

Once the performance tasks are initially developed, **cognitive laboratories** (also known as think aloud protocols) **are used with students** to collect evidence about task quality and the thinking processes that students employ when interacting with the task (see Appendix C). Tasks are then revised based upon student feedback. **Teachers then take the performance task themselves and swap performance tasks** in order to examine task quality and gather suggestions for revision. **Task specific, multi-dimensional rubrics are developed** to describe student performance on key competencies. The Center for Assessment then conducts a **mid-term review of the tasks and rubrics** using the High-Quality Assessment Review Tool (see Appendix D). This tool was developed using the criteria for high-quality assessments from the *Standards for Educational Psychological Testing*. This tool identifies areas of strength and provides recommendations for revisions. This feedback is provided to the educators who created the tasks and they are revised as necessary prior to pilot testing (see Summary of 2016 Common Task Review in Appendix E).

Teachers conduct small-scale pilots to evaluate and refine task quality. Teams of teachers from all PACE districts then convene to discuss task and rubric quality and understandability. Revisions are made to the tasks or rubrics as necessary. **The revised tasks are then re-piloted** in some classrooms and **identify anchor papers to support reliable scoring**.

At the end of the task development process, one PACE Common Task and student anchor papers per grade and subject area is chosen for operational use and to aid scoring during the next school year. This cycle repeats each year and builds a bank of prior PACE Common Tasks for teachers within PACE districts to use to support their local assessment purposes throughout the year.

There are three main purposes for the common tasks across districts: 1) to help measure the degree of cross-district comparability of scoring, 2) to serve as models of high quality tasks to support local task development, and 3) to contribute to the long-term goal of building a large task bank from which districts can draw for local assessment purposes. The first purpose is discussed



in greater detail in the comparability sections of this manual. The second purpose centers on the ideas that the common tasks are designed to provide districts with examples of high-quality performance tasks. As described above, the common tasks are run through an extensive development and review process before being approved by the NHDOE for operational use. The result is the set of operational tasks that provide models for designing rich, authentic assessment experiences that measure deep learning. The tasks are designed and reviewed specifically to allow for independent student inquiry, multi-step problem solving and argument building, and typically allow for multiple possible solutions. Part of the theory of action of PACE is that by requiring complex thinking on assessments, educators will need to prepare students to think deeply in order to perform well. The common tasks are one mechanism the help realize the PACE goals. Additionally, one of the goals of the PACE pilot is that by providing these models for high-quality performance assessments, local assessment capacity will increase. Local capacity is not only increased by preparing for and administering the common tasks, but by acutely engaging teachers in the common task development process. Cross-district teams of teachers come together for multiple, multi-day intensive sessions throughout the academic year and summer months to develop and refine the common tasks. The teachers who participate in this process are receiving hands-on professional development about best practices in assessment design to bring back to their respective districts.

The third purpose of the common tasks is to support one of the long-term goals of the PACE project which is to maintain a task bank of performance assessments. By rotating the competencies that are assessed by the common tasks each year, former common tasks can continue to be used as local tasks. Previously operational tasks will have the additional benefit of coming with annotated samples of student work to serve as anchor papers to calibrate scoring. This task bank can then be used by local educators to support their classroom assessment needs. As the number of PACE districts grows, the capacity of the cross-district teams of teachers to develop multiple assessments per year becomes more realistic.

Advanced Teacher Leader Training

The PACE pilot provides additional support to teacher by offering two kinds of leadership opportunities. First, the PACE pilot has received a three-year grant from the National Educator Association in New Hampshire to fund a cohort of teacher leaders. These teacher leaders carry out key communication and implementation functions associated with the PACE project. The roles for these teacher leaders are jointly state and locally defined. The primary intention of creating teacher leaders within the PACE districts is to build local capacity for integrating the PACE theory of action into the PACE districts and communicating about it to the public.

In addition to teacher leaders, the PACE pilot has invested in teachers to become content leads. Content leaders are responsible for the following duties:

- Support their colleagues in the development of the pilot and operational tasks.



- ☑ Facilitate the task development process; organize materials and send them to be posted on the LibGuide (an online intranet for PACE teachers).
- ☑ Review the LibGuide to make sure the most up to date materials are posted.
- ☑ Act as a liaison to the assessment experts to help resolve questions regarding assessment quality.
- ☑ Plan the task design process to meet deadlines.
- ☑ Communicate and Share the feedback to teachers from task review.
- ☑ Encourage positive, collaborative behavior amongst the teachers in the team.
- ☑ Communicate the goals of the next meeting and the tasks each teacher representative needs to complete.
- ☑ Lead the review of student work from the pilot to improve the task.
- ☑ Protects the project materials by not sharing passwords to guides with anyone outside of the project.

There are at least three areas of training used to deepen the expertise of content leads. First, content leads receive *advanced performance assessment training*, including discussions of validity theory and principled assessment design. Secondly, content leads receive additional support regarding *depth of knowledge* so that they can understand how to increase cognitive complexity—a critical factor in increasing the rigor of instructional and assessment practices. Lastly, teacher leaders receive training on the *facilitation* of adult learner to help them work with their colleagues to support the development of high-quality common performance tasks.

Summer Institute Training and Professional Activities

Teachers from Tier 1 districts gather each summer to review and score student work from other districts. These cross-district scoring opportunities provide a rich professional development opportunity for teachers as they discuss student work with colleagues from other districts and align their understanding of student performance using evidence from student work samples. Many teachers comment each year on evaluations of the Summer Institute that it is the best professional development they have ever received.

There is also new teacher and leadership training that takes place at the Summer Institute. Districts that will be implementing PACE as a Tier 1 district in the following school year send teams of teachers and administrators. Teachers from these districts mock score the PACE Common Tasks and also receive training in the design and implementation of high-quality performance tasks. District leaders receive training in how to support their teachers and schools through the process of implementing a new assessment and accountability system.

Summary

This section detailed the three-tiered system of supports from the NH DOE and its technical partners that prepares districts to implement the PACE system with fidelity. Specific and detailed information about the differentiated capacity building training and support offered to



implementing (Tier 1) districts was also provided. In the next section, a comparability-based framework for validating the NH PACE system of assessments is explained and the collected validity evidence is methodically detailed.



Comparability-Based Framework for Validating the System of Assessments

Overview of Validity Evidence for the NH PACE System

The NH DOE has developed a comprehensive plan for collecting and synthesizing validity evidence to support the uses of the NH PACE system results. This section situates the collected validity evidence within a comparability-based framework. The NH DOE has designed a system that ensures annual determinations of student proficiency are comparable within pilot districts, among pilot districts, and across pilot and non-pilot districts. The NH DOE engages in *comparability by design* to promote and evaluate the intended claims.

The validity of the NH PACE assessment and accountability system primarily rests on both **internal** comparability—i.e., the degree to which the assessment scores for a given grade and subject area within districts are comparable, as well as the local assessments among the PACE districts provide for comparable inferences regarding what students know and can do—and **external** comparability of PACE results to the other assessment systems used in the state for school accountability.

Defining Comparability

Comparability is a judgment based on an accumulation of evidence to support claims about the meaning of test scores and whether scores from two or more tests or assessment conditions can be used to support the same interpretations and uses. In this way, assessments are not dichotomously determined to be comparable or not, but like validity, comparability is a judgment about the strength of the theory and evidence to support the comparability of score interpretations for a given time and use. This means that evidence used to support claims of comparability will differ depending on the nature (or grain-size) of the reported scores. For example, supporting claims of raw score interchangeability—the strongest form of comparability—would likely require the administration of a single assessment form with measurement properties that are the same across all respondents (i.e., measurement invariance). Most state assessment systems with multiple assessment forms fail to meet this level of score interchangeability. Instead, the design of most state assessment systems aims to be “comparable enough” to support scale score interchangeability. This level of comparability typically requires that the multiple tests forms are designed to the same blueprint, administered under almost identical conditions, and scored using the same rules and procedures. Still, many states continue to struggle to meet this level of comparability due to challenges with multiple modes of administration—paper, computer, and devices (see DePascale, Dadey & Lyons, 2016). In this way, comparability is an evidence-based argument, and the strength of evidence needed will necessarily depend on the type of score being supported. As shown in Figure 3, comparability lies on a continuum and rests on two major critical dimensions: the comparability of content and the comparability of scores, and that each of these may exist at different degrees of granularity.



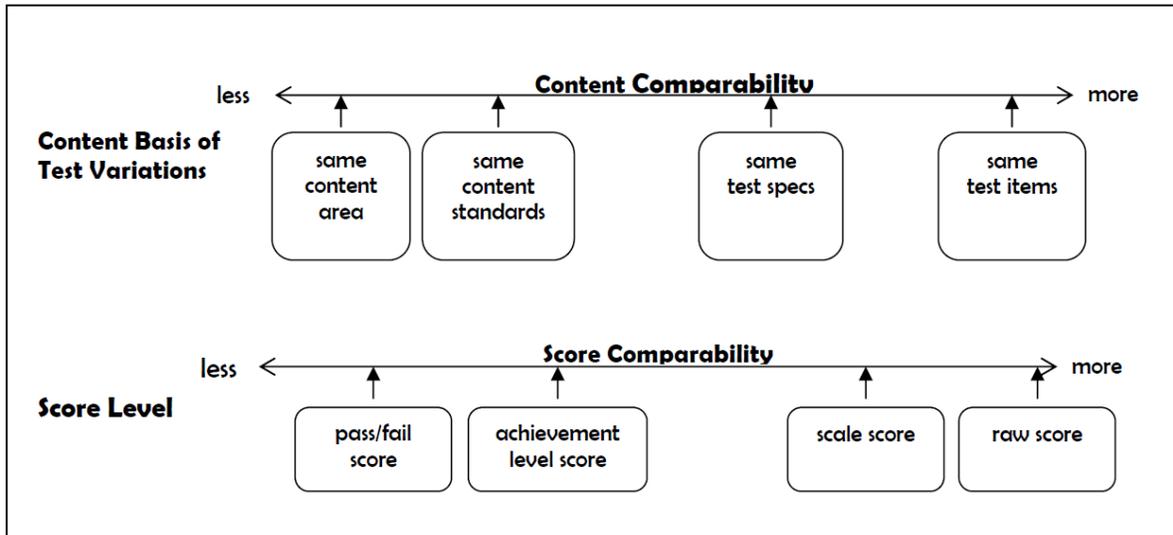


Figure 3. Comparability Continuum (Winter, 2010, p. 5)

Reiterating our earlier recommendation, comparability must be required at the level of the annual determinations. This means that evidence is provided to support the notion that if a student is determined to be “proficient” in one district, had that student been assigned to another district’s assessment system (either pilot or non-pilot) he or she could expect to also be deemed proficient.

Supporting Claims of Comparability across Assessment Systems

The issue of comparability across the two state assessment systems is of primary concern for two reasons. First, because NH must incorporate assessment results from the pilot districts into the state accountability system alongside the results generated from the non-pilot districts, the assessment systems must produce results that are comparable enough to support their simultaneous use in the single statewide accountability system. Secondly, requiring that the assessment systems produce comparable results ensures that the state and district will not view the innovative assessment and accountability demonstration authority as a way to relax the rigorous expectations established under the current state assessment systems. The innovative assessment system must be aligned to the intended content standards and produce annual summative determinations that are consistent across the two assessment programs. This does not require scale score comparability, but does require the ability to meaningfully compare the achievement level classifications for use in the accountability system.

To address these two major concerns, NH generates and evaluates ample evidence of comparability of assessment results. Evidence of comparability supports the notion that in general, schools that are participating in the innovative assessment system could be expected to have similar distributions of students into performance classifications had the school instead participated in the statewide standardized assessment system. This is not to say that we would expect all districts that participate in the innovative pilot to exhibit similar levels of achievement



as the non-pilot districts—because pilot districts will be most certainly a non-random sample, or, the innovative learning model associated with the assessment system should influence achievement—but that the performance standards support the same interpretations relative to the level of achievement of the learning targets.

Overview of Comparability Methods

As mentioned previously, there are three main levels of comparability used to validate the NH PACE system of assessments: within-district comparability, cross-district comparability, and comparability across state assessment systems. Examples of the activities and audits that occur at the three levels are summarized in Figure 4 and described in detail below going from the lowest level to the highest level. Gathering evidence at each of these levels is essential for supporting the claims of comparability, and ultimately supporting the validity of the system as a whole.

The 2016-2017 Data Collection Protocols alongside the additional resources listed into those Protocols (see Appendix G) explain the types of data districts submit in order to examine comparability within and across districts. The state provides the data necessary to examine comparability across the two state assessment systems.

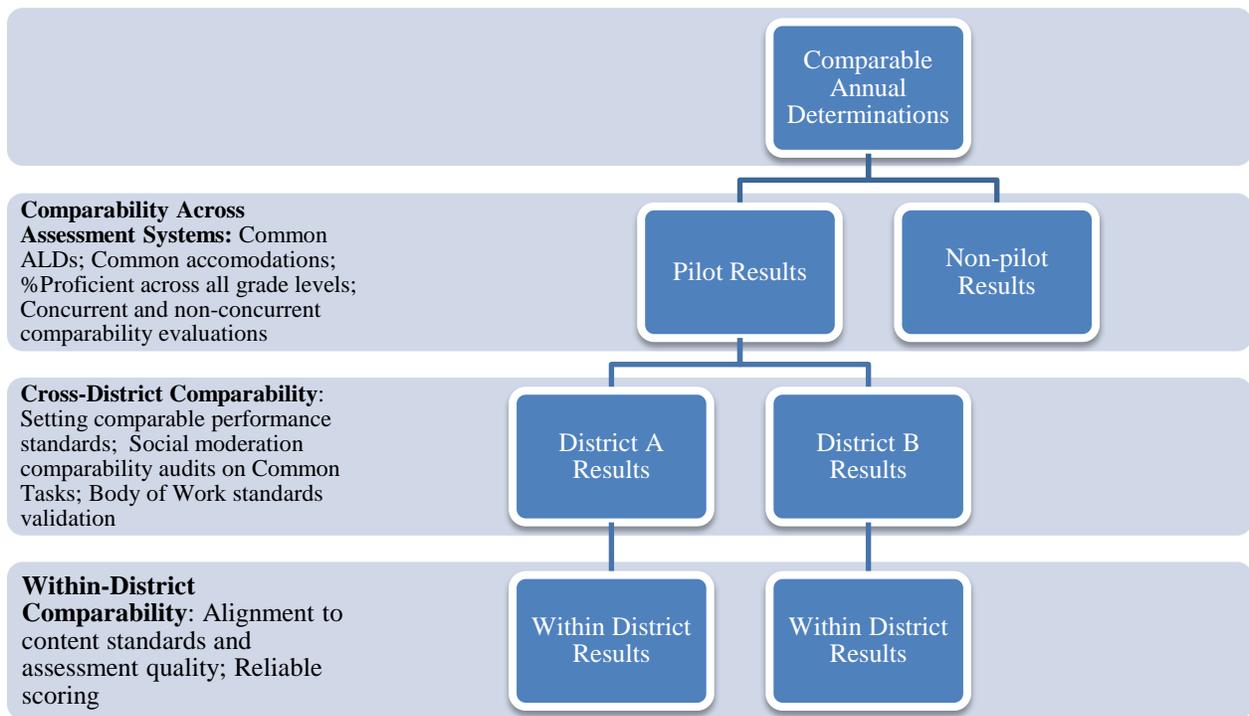


Figure 4. Establishing an Evidence-Base for Comparable Annual Determinations



Within-District Comparability in Expectations for Student Performance

There are two main sources of within-district comparability evidence: A) alignment and assessment quality and B) reliable scoring. Evidence regarding alignment and assessment quality comes from 1) reviews of local assessment maps and 2) two-part reviews of local task quality. Evidence regarding reliable scoring comes from process-based evidence (e.g., principles of scoring student work, calibration and anchor paper protocols for the PACE Common Task and local tasks, double scoring protocols), as well as audits on inter-rater reliability and the generalizability of local assessment scores. Each of these is discussed in detail below.

A. Evidence of Alignment and Assessment Quality

1. Reviews of local assessment maps. For the 2016-2017 academic year, the NH DOE is collecting and reviewing assessment maps from all PACE districts for all grades and subjects covered under the PACE pilot as a way to document that all content standards are addressed in the assessment system. **The purpose of reviewing the assessment maps is to ensure all students are provided with a meaningful opportunity to learn the required grade level content standards. Every district submitted their assessment maps for state review. The assessment maps went under review in June and July 2017 to ensure that the entire breadth and depth of the state standards are being assessed to inform the competency determinations throughout the year.** The information provided in the assessment maps includes:

- The competencies assessed in each course
- The alignment of the state standards to the competencies
- The number, type, and timing of the summative assessments administered for each competency.

Figure 5 contains an example of the assessment map submitted on November 1, 2016 by the Monroe school district for fourth grade ELA.



Competency	Standards	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
Foundational Reading Skills	RF.4.3 RF.4.4 L.4.3 L.4.4 L.4.5 L.4.6		PBA - Characters Unit 2		PBA - Voices to Remember		Bi-Weekly Story Tests			PBA - Grand Canyon	
Reading Literature	RL.4.1-7 RL.4.9-10										
Reading Informational Texts	RI.4.1-7 RI.4.9-10										
Narrative Writing	W.4.3 W.4.4 W.4.5								Small Moment in their Life - On Demand Prompt		
Informational Writing	W.4.2 W.4.4 W.4.5 W.4.6 W.4.8 W.4.9						Informational Topic of their Choice	PACE Common Task		PBA - Grand Canyon	
Opinion & Argument Writing	W.4.1 W.4.4 W.4.5 W.4.8 W.4.9		PBA - Characters Unit 2	PBA - Persuasive Letter				PACE Common Task			PBA - Opinion School Lunch
Speaking, Listening, & Language	LS.4.1 LS.4.2 LS.4.3 LS.4.6 L.4.1 L.4.3-6				PBA - Voices to Remember			PACE Common Task		PBA - Grand Canyon	
Inquiry, Investigation, & Research	W.4.7 W.4.8 W.4.9 W.4.10 SL.4.1-6										

Figure 5. Monroe G4 ELA Assessment Map

The bullets below provide illustrative examples of the type of **feedback** we provided upon review of this map:

- ✓ G4 ELA standard RI.4.8 appears to be missing from this assessment map. This standard requires the integration of knowledge and ideas for reading informational texts. Specifically, students should be able to explain how an author uses reasons and evidence to support particular points in a text. Please provide the state with an example of when students are asked to demonstrate this skill on a summative assessment and update the assessment map to reflect that addition.
- ✓ G4 ELA standard L.4.2 appears to be missing from this assessment map. This standard covers the conventions of Standard English. Specifically, students should be able to demonstrate command of the conventions of Standard English capitalization, punctuation, and spelling when writing. Please provide the state with an example of when students are asked to demonstrate this skill on a summative assessment and update the assessment map to reflect that addition.
- ✓ All competencies are assessed multiple times with the exception of Reading Informational Texts. You may consider adding another assessment opportunity for your students in this competency.
- ✓ The total number of assessments that will factor into the end-of-year annual determination is approximately 16. Based on our preliminary generalizability analyses, this number should be sufficient for generating reliable estimates of student achievement.



- ✓ All of your competencies are assessed by at least one performance-based assessment (PBA). This shows dedication and fidelity to the vision for the PACE assessment system.

Assessment Map Review Protocol

<input type="checkbox"/> Formatting is clear and follows the data collection protocols
<input type="checkbox"/> All assessments are summative in nature
<input type="checkbox"/> All standards are addressed
<input type="checkbox"/> Each competency is measured with multiple assessments
<input type="checkbox"/> There are greater than 15 summative assessments for the full competencies
<input type="checkbox"/> All competencies are assessed by at least one performance assessment that measures deeper levels of understanding
<input type="checkbox"/> The types of assessments match the skills/standards being assessed
<input type="checkbox"/> Additional feedback for the district

Summary of Results of 2016-17 Assessment Map Review

The comprehensive assessment map review included a careful evaluation of each district’s assessment map using specific review criteria intended to provide formative feedback to districts for each grade level or course. The review criteria include: clarity, alignment between state standards and standards assessed, coherence between the PACE theory-of-action and the types of assessments listed, and generalizability based on the number of assessments.

Assessment maps for all courses that receive PACE annual determinations were reviewed for each Tier 1 district/school in June and July 2017 by the Center for Assessment. SAU 35 is a unit comprised of separate districts, so maps for each of the five SAU 35 schools were reviewed separately: Bethlehem, Lafayette, Landaff, Lisbon, and Profile. A total of 140 assessment maps were reviewed. Though maps varied in quality and completeness, all but four maps were missing in Science, and one map missing in each of math and ELA. The tables below display the number and type of assessment maps reviewed for each Tier 1 district/school.



Concord				Epping				Monroe			
Subject	Grade	Map Submitted	Reviewed	Subject	Grade	Map Submitted	Reviewed	Subject	Grade	Map Submitted	Reviewed
ELA	4	yes	JT	ELA	4	yes	JT	ELA	4	yes	SL/JT
	5	yes	JT		5	yes	JT		5	yes	JT
	6	yes	JT		6	yes	JT		6	yes	JT
	7	yes	JT		7	yes	JT		7	yes	JT
	9	yes	JT		9	yes	JT		9		
	10	yes	JT		10	yes	JT		10		
Math	3	yes	SL	Math	3	yes	SL	Math	3	yes	SL
	5	yes	SL		5	yes	SL		5	yes	SL
	6	yes	SL		6	yes	SL		6	yes	SL
	7	yes	SL		7	yes	SL		7	yes	SL
	9	yes	SL		9	yes	SL		9		
	10	yes	SL		10	yes	SL		10		
Sci	4	yes	CE	Sci	4	yes	CE	Sci	4	yes	CE
	8	yes	CE		8	yes	CE		8	yes	CE
	9	Yes	CE		9	yes	CE		9		
	10	yes	CE		10	yes	CE		10		

Pittsfield				Rochester				Sanborn			
Subject	Grade	Map Submitted	Reviewed	Subject	Grade	Map Submitted	Reviewed	Subject	Grade	Map Submitted	Reviewed
ELA	4	yes	JT	ELA	4	yes	JT	ELA	4	yes	JT
	5	blank	JT		5	yes	JT		5	yes	JT
	6	yes	JT		6	yes	JT		6	yes	JT
	7	yes	JT		7	yes	JT		7	yes	JT
	9	yes	JT		9	yes	JT		9	yes	JT
	10	incomplete	JT		10	yes	JT		10	yes	JT
Math	3	NO	NA	Math	3	yes	SL	Math	3	yes	SL
	5	Yes	SL		5	yes	SL		5	yes	SL
	6	Yes	SL		6	yes	SL		6	yes	SL
	7	yes	SL		7	yes	SL		7	yes	SL
	9	yes	SL		9	yes	SL		9	Yes	SL
	10	yes	SL		10	yes	SL		10	Yes	SL
Sci	4	NO	NA	Sci	4	yes	CE	Sci	4	yes	CE
	8	yes	CE		8	yes	CE		8	yes	CE
	9	yes	CE		9	yes	CE		9	yes	CE
	10	yes	CE/CE		10	yes	CE		10	yes	CE



Seacoast				Souhegan				Bethlehem			
Subject	Grade	Map Submitted	Reviewed	Subject	Grade	Map Submitted	Reviewed	Subject	Grade	Map Submitted	Reviewed
ELA	4	yes	JT	ELA	4			ELA	4	yes	JT
	5	yes	JT		5				5	yes	JT
	6	yes	JT		6				6	yes	JT
	7	yes	JT		7				7		
	9				9	yes	JT		9		
	10				10	yes	JT		10		
Math	3	yes	SL	Math	3			Math	3	yes	SL
	5	yes	SL		5				5	yes	SL
	6	yes	SL		6				6	yes	SL
	7	yes	SL		7				7		
	9				9	yes	SL		9		
	10				10	yes	SL		10		
Sci	4	NO	NA	Sci	4			Sci	4	yes	CE
	8	NO	NA		8				8		
	9				9	yes	CE		9		
	10				10	yes	CE		10		

Lafayette				Landaff				Lisbon			
Subject	Grade	Map Submitted	Reviewed	Subject	Grade	Map Submitted	Reviewed	Subject	Grade	Map Submitted	Reviewed
ELA	4	yes	JT	ELA	4			ELA	4	yes	JT
	5	yes	JT		5				5	yes	JT
	6	yes	JT		6				6	yes	JT
	7				7				7	yes	JT
	9				9				9	yes	JT
	10				10				10	yes	JT
Math	3	yes	SL	Math	3	yes	SL	Math	3	yes	SL
	5	yes	SL		5				5	yes	SL
	6	yes	SL		6				6	yes	SL
	7				7				7	yes	SL
	9				9				9	yes	SL
	10				10				10	yes	SL
Sci	4	NO	NA	Sci	4			Sci	4	yes	CE
	8				8				8	yes	CE
	9				9				9	yes	CE
	10				10				10	yes	CE



Profile			
Subject	Grade	Map Submitted	Reviewed
ELA	4		
	5		
	6		
	7	yes	JT
	9	yes	JT
	10	yes	JT
Math	3		
	5		
	6		
	7	yes	SL
	9	yes	SL
	10	yes	SL
Sci	4		
	8	yes	CE
	9	yes	CE
	10	yes	CE

A range of minor to major revisions were needed on the assessment maps in order to ensure appropriate documentation of the assessment of all content standards. The most common suggested revisions include the following: 1) identifying the standards associated with the competency or in some cases filling missing gaps in the completeness of the standards assessed, 2) ensuring multiple opportunities for students to demonstrate proficiency toward each competency, 3) ensuring that there are at least 15 summative assessments for the full set of competencies in order to generate reliable estimates of student achievement, and 4) ensuring that all or most of the competencies are assessed by at least one performance-based assessment.

In order to continue to build the strength of the local assessment systems, a workshop has been included in the Leadership Strand of the NH PACE Summer Institute on Local Assessments and Assessment Maps. This workshop will provide an opportunity for school/district administrators to bring and share their assessment maps and their local performance assessments, and receive custom feedback on their maps that resulted from the review. This workshop will be facilitated by Center for Assessment staff so that key quality control issues can be discussed.

Given that the districts can greatly benefit from learning from their peers, the 2017-2018 data collection protocols have been updated to include a district peer review of the local assessment maps and aligned assessments. The rationale behind this change is to generate rich conversations about the structure of assessment systems across districts. The hope is that this district interaction will lead to increased sharing and leveraging of best practices to tackle common challenges. As the pilot scales, it is not efficient or feasible for the monitoring of the quality of the local assessment maps and aligned assessments to fall centrally on the shoulders of the state. The buddy district peer review system creates a system of internal accountability, with audits and checks by the state, all while continuing to build local capacity and agency. During the first year of district peer review, we will ask districts to share and review their maps and aligned assessments from a sample of the PACE courses. Given the high degree of similarity in the maps across grade levels and courses within districts, this sample should reveal any systematic issues



in the quality of local assessment systems while also reducing the burden on the districts. The district peer review protocols are included in the next section.

District Peer Review Protocols for 2017-2018

The District Peer Review represents an opportunity for your district to receive feedback from a peer district regarding a sample of your local course assessment maps and assessments. Each year the pairs of buddy districts will rotate along with the sample of grades and content areas sampled for review. Providing clear documentation of your assessment maps and samples of assessments to your buddy district will help you receive better and more helpful feedback. The district peer review system is designed to monitor the quality and alignment of the local assessments and provide formative feedback to districts on the state of their assessments and assessment systems.

Buddy Districts

Epping	Souhegan HS
Rochester	Concord
Sanborn	SAU 35
Seacoast Charter	Monroe
Pittsfield	Amherst Middle School
Laconia Elementary Schools	SAU 23 Elementary Schools
Plymouth Elementary School	Richards Elementary School

Submission Process

- Work with your buddy district to organize the transfer of materials for peer review by January 15, 2018. The materials should include one assessment map and **three (3)** aligned summative assessments for each of the following courses:

Grade	Subject Area
3	Math
4	Science
5	ELA
6	Math
7	ELA
8	Science
HS	Algebra
HS	Grade 10 ELA
HS	Life Science

- All of the state standards should be mapped to at least one competency. The summative assessments for each competency should be labeled by type and mapped by time of administration. Anything included in the assessment map may be subject to a state audit to ensure assessments are aligned to intended standards and are high quality.



- For each course, three summative assessments should be submitted along with any scoring guides/rubrics and any other information teachers might need to help evaluate the quality of the assessment (e.g., samples of student work).

Example Grade 3 Assessment Map

Competency	Standards	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
1. Algebraic Thinking	CC.3.OA.1	Short Summative		PBA	Unit Test						
	CC.3.OA.2										
	CC.3.OA.3										
	CC.3.OA.4										
	CC.3.OA.5										
	CC.3.OA.6										
	CC.3.OA.7										
	CC.3.OA.8										
	CC.3.OA.9										
2. Number Operations	CC.3.NBT.1		Short Summative					PBA	Short Summative	PBA	
	CC.3.NBT.2										
	CC.3.NBT.3										
3. Fractions and Proportional Reasoning	CC.3.NF.1			Short Summative		Short Summative	Unit Test	PBA			
	CC.3.NF.2										
	CC.3.NF.2a										
	CC.3.NF.2b										
	CC.3.NF.3										
	CC.3.NF.3a										
	CC.3.NF.3b										
	CC.3.NF.3c										
	CC.3.NF.3d										
4. Data	CC.3.MD.3									Short Summative	Unit Test
	CC.3.MD.4										
5. Geometry and Measurement	CC.3.MD.1				Short Summative			Short Summative	PACE Common Task		
	CC.3.MD.2										
	CC.3.MD.5										
	CC.3.MD.6										
	CC.3.MD.7										
	CC.3.MD.7a										
	CC.3.MD.7b										
	CC.3.MD.7c										
	CC.3.MD.7d										
	CC.3.MD.8										
	CC.3.G.1										
	CC.3.G.2										

Review Process

Note: Trainings for district and school personnel who will be reviewing assessment maps and local assessments will be available in February and March of 2018.

- Arrange for a team of teachers, curriculum coordinators, and other instructional support staff to jointly review the submitted course packets (assessment map and aligned assessments) for the following courses:

Grade	Subject Area
3	Math
4	Science



5	ELA
6	Math
7	ELA
8	Science
HS	Algebra
HS	Grade 10 ELA
HS	Life Science

- Your district need only review those courses that you offer within your district. For example, if your PAC E Tier 1 schools are only elementary, you need only review the elementary courses submitted to you.
- For each course packet that your district reviews, complete the Assessment Map and Assessment Peer Review Feedback Sheet.

Assessment Peer Review Feedback Sheet

District Being Reviewed: _____

Reviewing District: _____

Grade Level and Content Area: _____

Names of Reviewers: _____

Assessment Map Review Checklist:

- Formatting is clear and follows the data collection protocols.
- All assessments included in map are summative in nature—i.e., can be used to support competency determinations for students.
- All grade-level standards are assessed within at least one competency. If not, list the standards that are missing below:

- There are multiple assessment opportunities are available for every competency.
- There are at least 15 summative assessments for the full set of competencies—i.e., the set of assessments is likely to yield generalizable inferences about what students know and can do.
- All or most of the competencies are assessed by at least one performance assessment that measures deeper levels of understanding.

Assessment Quality Review Feedback:

This is an opportunity for the team of peer review teachers to comment on the quality of the assessment and suggest opportunities for improvement. Key factors of quality that should be reviewed include alignment to grade-level content expectations, depth of knowledge, alignment to principles of universal design for learning, and quality of scoring guides/rubrics.



<u>Assessment #1</u>	<u>Assessment #2</u>	<u>Assessment #3</u>
(Note the size of these spaces have been reduced for the purposes of the technical manual).		

Feedback Submission to Buddy District and State

- Once your course reviews are complete, send a copy of the completed Assessment Map and Assessment Peer Review Feedback Sheet for each course to your buddy district and to Mariane Gfroerer at Mariane.Gfroerer@doe.nh.gov by **April 13, 2018**.

2. Two-Part Review Protocol for Local Assessments. In addition to the review of the assessment maps, the NH DOE will also be contracting with the Center for Assessment to review the quality of one major assessment per competency for each course in every district. The results of this assessment audit will be available for the Fall 2018 progress update to USED. These assessments will be reviewed for technical quality with formative feedback provided to the districts. Alignment, with a focus on the depth at which the learning is measured, is the most important review criterion. The NH PACE theory of action postulates that having students engage in rich, cognitively-demanding assessment experiences, instruction and student achievement will improve. State audits will help ensure that students have an opportunity to learn the content standards and they are being assessed at a high depth of knowledge. If the audit reveals any systematic problems in the local assessment quality, state leaders will support those districts with professional capacity building opportunities.

✓ Part 1: NHDOE Assessment Review

The PACE Director at the NH DOE will review all of the submitted assessments and document the grade level, source, alignment to standards and assessment map, and the highest depth of knowledge reached by the assessment. This review will be used as a preliminary audit of identifying assessments that are of low quality and should undergo an in-depth review for technical quality by the Center for Assessment.

✓ Part 2: In-depth Center for Assessment Review

The Center for Assessment will provide a more rigorous technical review of the assessments flagged as potentially low quality by the PACE Director. The purpose of the in-depth review will be to provide constructive feedback to districts regarding best practice for assessment design in the PACE system. The assessment review will provide specific feedback on the individual assessments. If applicable, the Center will provide narrative formative feedback to the districts on how they can improve their assessment practices in general. The review will focus on elements of quality such as:



- Alignment to the ***content*** and ***rigor*** of the assessed standards
- Assessment design and item writing best practice (clarity, cognitive load, format, length)
- Bias/sensitivity

B. Evidence of Reliable Scoring

1. Principles of Scoring Student Work. All PACE districts hold grade-level calibration sessions for the scoring of the PACE Common Task. Teachers bring samples of their student work from the PACE Common 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. The 2016-2017 Data Collection Protocols document (see Appendix G) contains detailed instructions about calibration and anchor paper protocols for PACE Common Tasks and double scoring protocols for samples collection from PACE Common Tasks.

2. Inter-Rater Reliability Estimates. We externally audit the consistency in scoring by asking each district to submit a sample of papers from each PACE Common 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. For example, in 2015-2016, all participating PACE districts were asked to have 18 student work samples for each PACE Common Task scored by two teachers independently, thereby producing within-district 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 7.

Table 7.
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 8). 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 8.

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

Task	N	Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5	
		%Exact	%Adj								
<i>ELA</i>											
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		
<i>Math</i>											
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				
<i>Science</i>											
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 9). 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 9.
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:



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

where Pr(a) is observed agreement and Pr(e) is the probability of chance agreement. Table 10 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 10.
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 11 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 11.

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

3. Generalizability Analysis. In the NH PACE 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.

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 12 for the number of students and assessment tasks.

Table 12.

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 13 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 x t interaction; (2) sources of error variability in the competency score measurement that the one-facet p x t design has not captured, or (3) both. A large variance component for the p x 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 13.

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 14 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 14.

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. Figures 6 and 7 show sample plots showing estimated relative and absolute error generalizability coefficients as a function of the number of assessments by grade and subject.

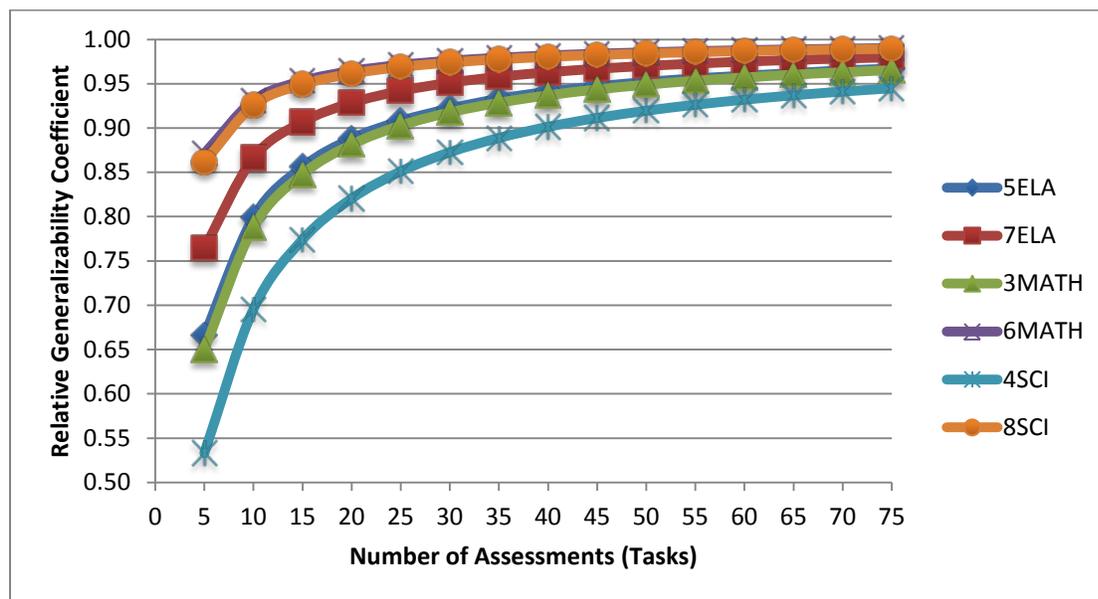


Figure 6. Sample plots showing estimated $E\rho^2$ scores as a function of the number of assessments by grade and subject



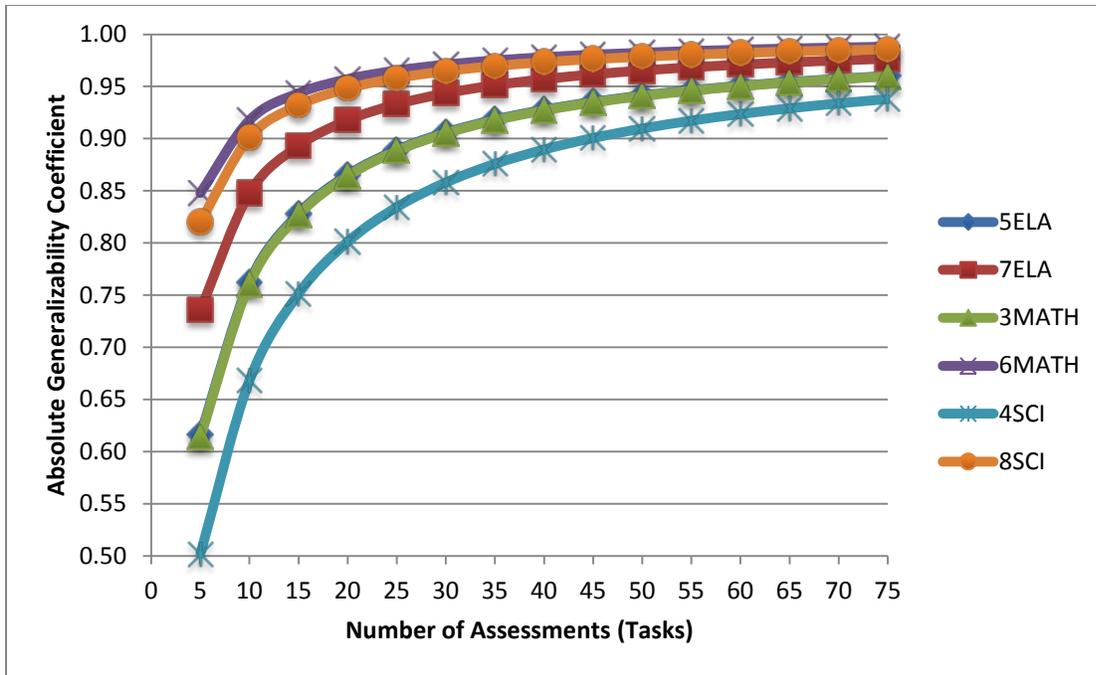


Figure 7. Sample plots showing estimated ϕ coefficient as a function of the number of assessments by grade and subject

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 15.



Table 15.

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.

Cross-District Comparability In Expectations of Student Performance

There are three main sources of cross-district comparability evidence: A) setting comparable performance standards, B) social moderation comparability audits using the PACE Common Tasks, and C) Body of Work standard validation. Each will be discussed in turn.

A. Setting Comparable Performance Standards

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.¹² 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. The only exception to this was in Grade 8 Science in Epping where scores above the 100-point scale were maintained since the one teacher in Grade 8 Science in Epping used extra credit opportunities to further differentiate among students at the high end of the competency scale.

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).

¹² We recommend that once the assessment maps are submitted for AY 2016-2017, districts work with the data team to establish a weighting scheme for the competencies that is defensible.



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 lower 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 2016, there was one case where there were multiple sources of evidence indicating an issue of incomparability in local scoring. See the following section entitled, "Social Moderation Comparability Audits on PACE Common Tasks" for more information.

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 Leve 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.

B. Social Moderation Comparability Audits on PACE Common Tasks (and adjustments to performance standards).

The PACE innovative assessment system uses PACE Common Tasks across districts to evaluate the degree of comparability in local scoring. These analyses rest on the assumption that patterns in scoring for the PACE Common Task is representative of district relative stringency and leniency in scoring of the local performance tasks and assessments. This assumption has been supported by evidence of generalizability (see Generalizability analyses above). 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 Common 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 calibration 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 occurred 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 Task from 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 16.

Table 16.

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 17 shows the average observed discrepancy by district.

Table 17.

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 PACE 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 (Table 18) 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 18.

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 8, 9, and 10 below.

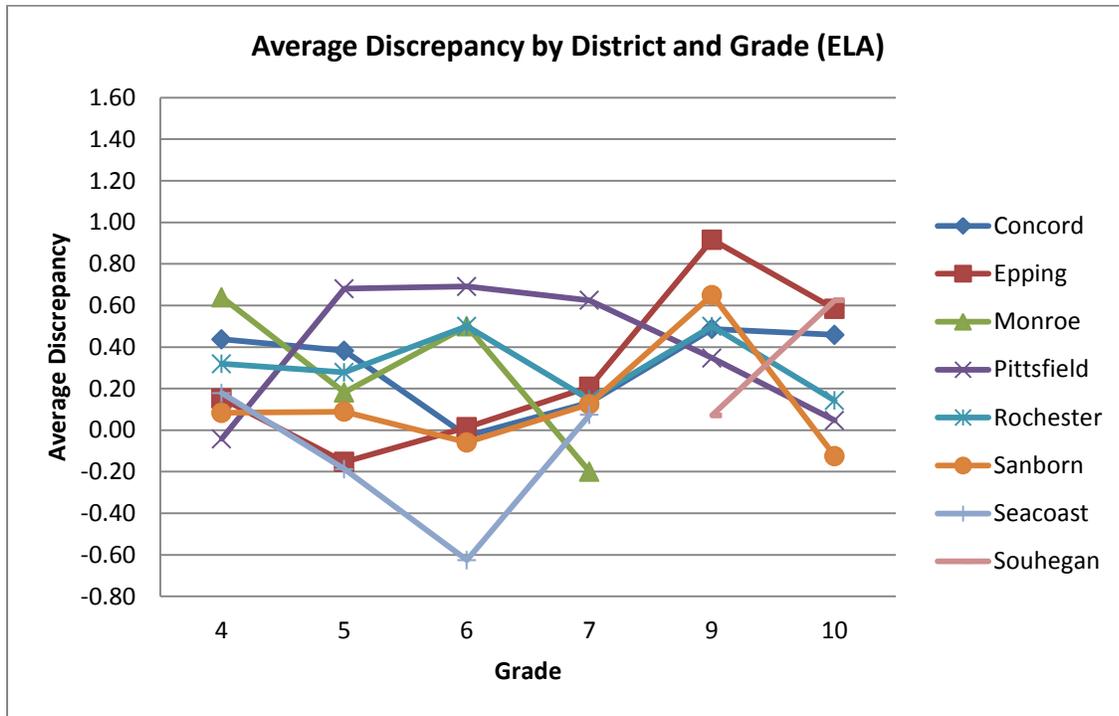


Figure 8. Marginal Means for ELA

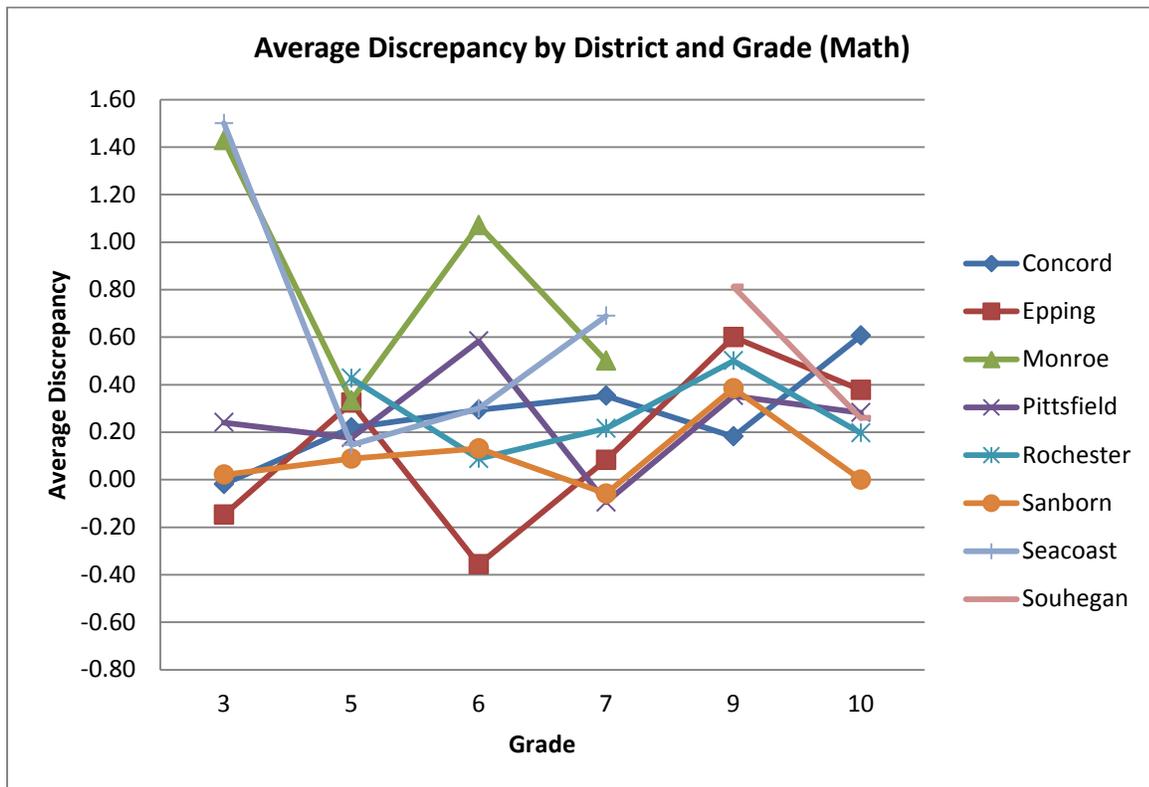


Figure 9. Marginal Means for Math

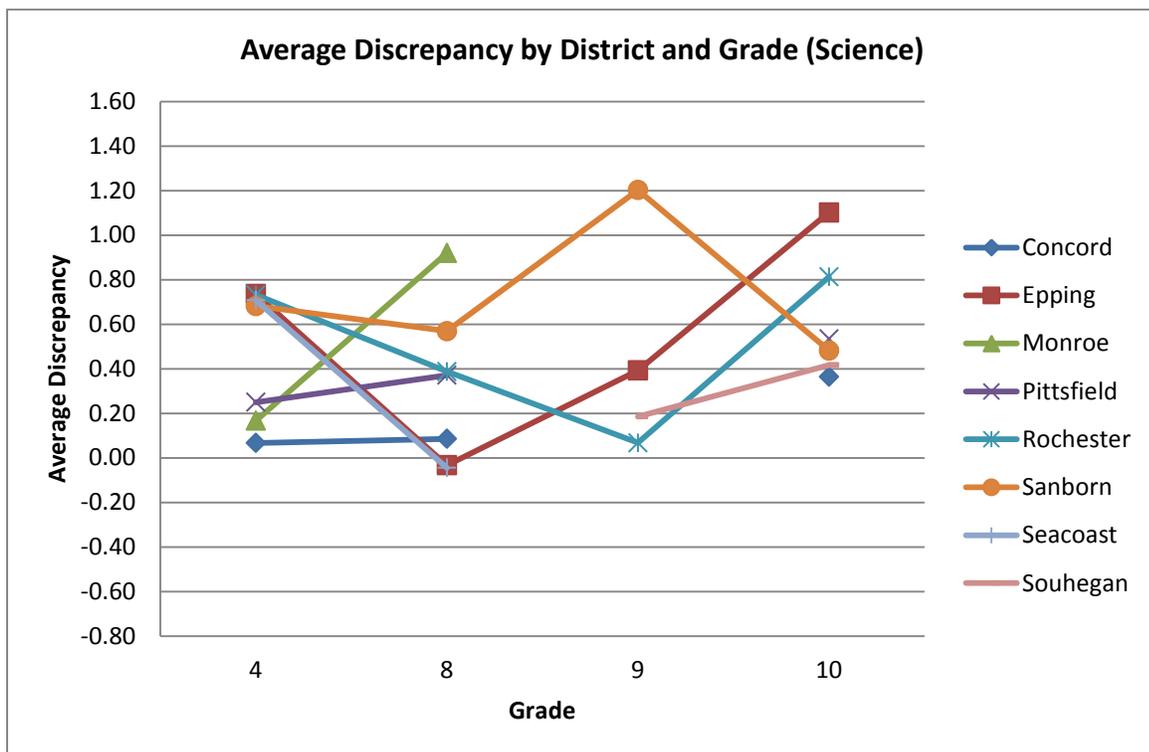


Figure 10. Marginal Means for Science

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 (Table 19). The equality of variance assumption was met for all combinations except fifth grade math for which the appropriate *t* value correction was made.

Table 19.
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 20).

Table 20.

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 11-15.

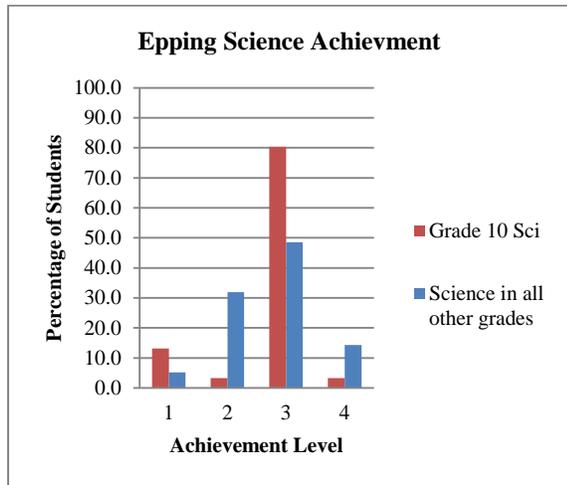


Figure 11. *Epping G10 Science Comparison*

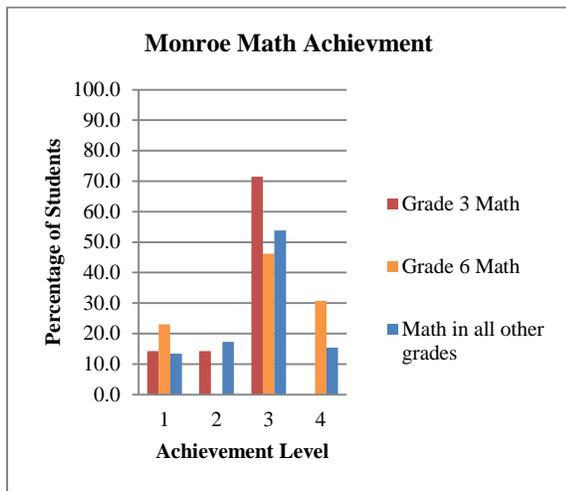


Figure 12. Monroe G3 & G6 Math Comparisons

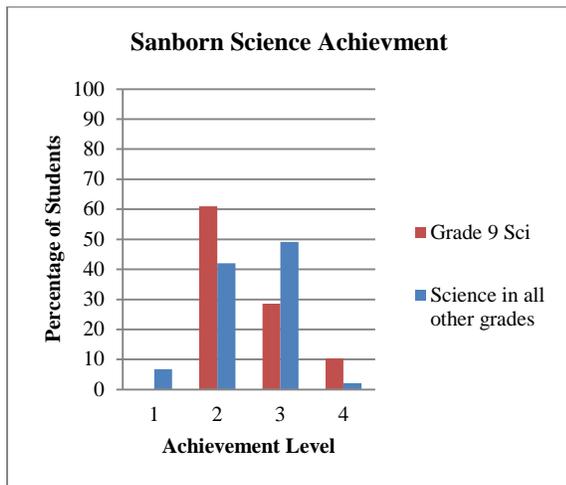


Figure 13. Sanborn G9 Science Comparison

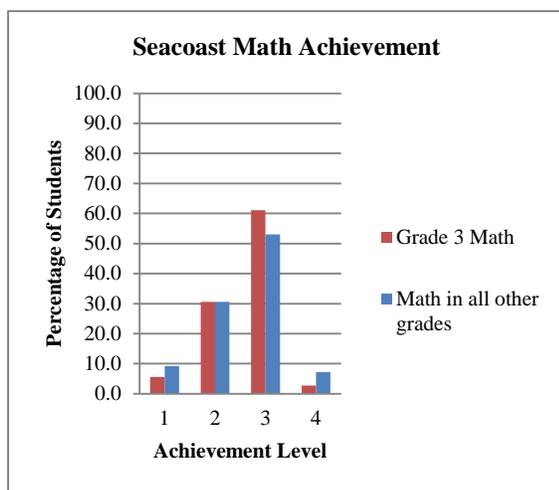


Figure 14. Seacoast G3 Math Comparison

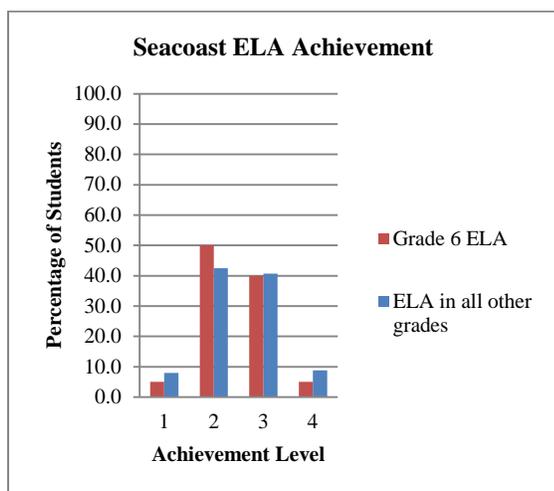


Figure 15. Seacoast G6 ELA Achievement

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 (see Table 21).

Table 21.
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 equipercentile standard setting technique using Grade 9 science achievement at the reference distribution. The Table 22 and Figure 16 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 22.

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

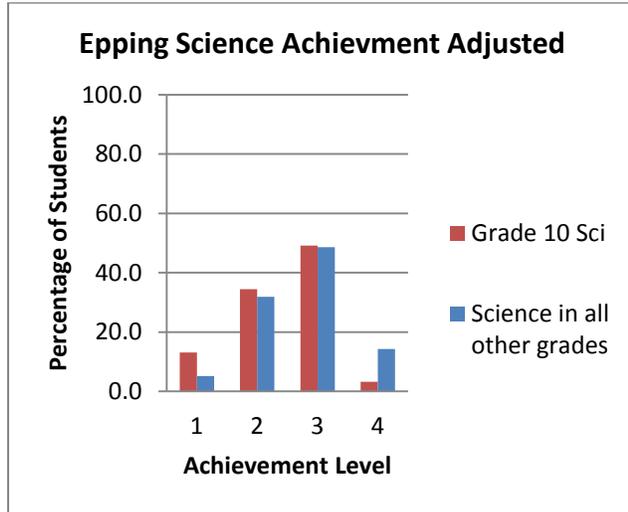


Figure 16. Resulting G10 Science Distribution Comparison

As we have noted in this technical manual, 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.

C. Body of Work (BOW) Standards Validation

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 (see Appendix G for more information). 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 (BOW) 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 scorable (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 17 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 23 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.

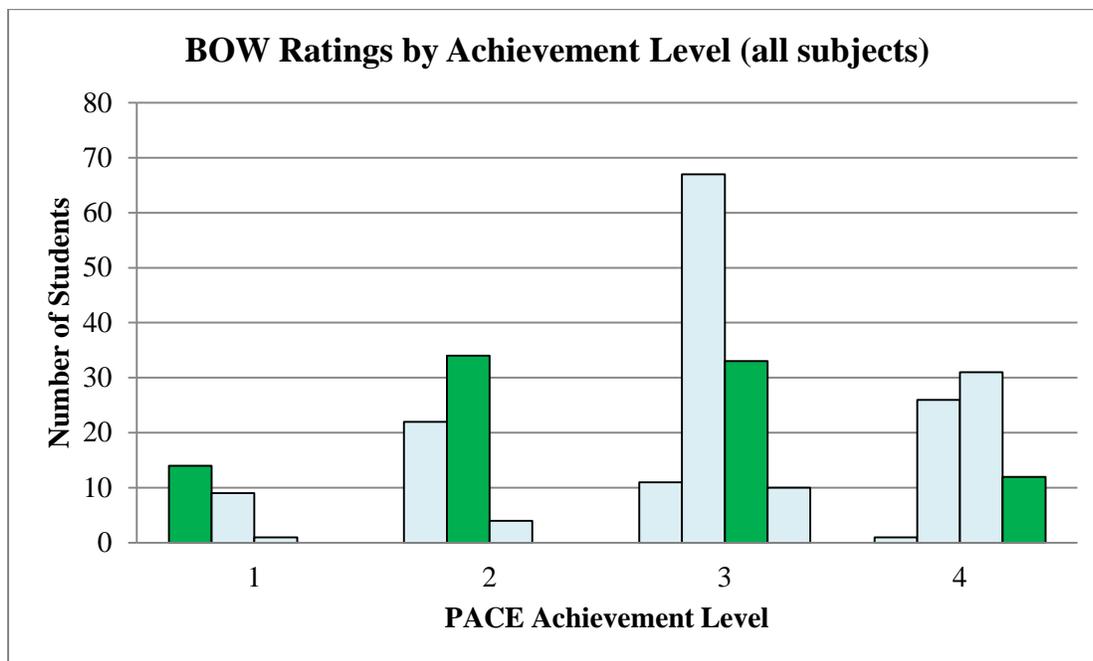


Figure 17. Distribution of BOW ratings by PACE Achievement Level

Table 23.

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 .01 level alpha level.

In general, the agreement between the BOW ratings and the PACE annual determinations is not as strong as expected. Figure 17 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 are 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.

Comparability of Annual Determinations across Assessment Systems

The accountability uses for the assessment system results rests on the comparability of annual determinations. Therefore, the comparability claims for the innovative pilot will apply to the reported performance levels (as opposed to scale scores for more traditional assessment models). The comparability processes and audits that occur at both the local, within-district level and the cross-district level are all in an effort to support the claim of comparability in the annual determinations. However, if the pilot is not statewide, a major ESSA comparability requirement is that the pilot system results are comparable with the non-pilot district results. The following procedures are used to formally promote and evaluate the comparability of the annual determinations across both pilot and non-pilot districts: common Achievement Level Descriptors (ALDs) and ALD development process; percent proficient across all grade levels; concurrent comparability evaluations; and non-concurrent comparability evaluations. Before detailing these sources of evidence for the PACE system, we discuss reasonable expectations for comparability across the two state assessment systems.

There are a variety of reasons why there may be legitimate differences in the results produced by the two or more assessment systems. New Hampshire is taking advantage of the ESEA waiver for three reasons, 1) to measure the state-defined learning targets more flexibly (e.g., when students are ready to demonstrate “mastery”), 2) to measure the learning targets more completely and/or deeply, and 3) to measure targets from the standards that are not measured in the general statewide assessment (e.g., listening, speaking, extended research, scientific investigations). Therefore,

requiring the results produced across the old and new systems to tell the same story about student achievement has the very real potential to prevent meaningful innovation. To quote one of the leading experts on score comparability, Dr. Robert Brennan, when asked about comparability between the innovative and standardized assessment systems, “perfect agreement would be an indication of failure.”

Given this, *how comparable is comparable enough?* For example, if approximately 55% of the students were scoring in Levels 3 and 4 on the state standardized assessment, that does not mean we should expect exactly 55% of the students to be classified in Levels 3 and 4 in the PACE system. There could be very good reasons why the results would differ in either direction. For example, the PACE pilot system of assessments may be capturing additional information relative to real-world application and knowledge transfer that provides for more valid representations of the construct than possible with traditional standardized assessments. For this reason, we do not set a standard criterion, or comparability “bar”, because the intended uses and contextual factors surrounding the evaluation of comparability are critical.

However, it is worthwhile to consider what might be reasonable to expect for the amount of variability in proficiency classifications across the two assessment programs. We argue that a reasonable upper bound for comparability across pilot and non-pilot systems is the degree to which comparability is achieved across forms, modes, and years of administration for the statewide, standardized assessment system. This is akin to the axiom that a test cannot correlate any more with another test than it does with itself (i.e., its reliability). The literature is clear that there are significant effects associated with mode of administration (including paper/computer and across devices), accommodations, and forms across years.¹³ Due to the precedence for this type of variation within our current assessment systems, it may be reasonable to expect that the variability across the pilot and non-pilot would be at least as large as levels we see with current state testing programs. Again, when we refer to variability across assessment programs, we are not expecting that pilot and non-pilot districts exhibit the same levels of achievement—because districts are not randomly assigned to the pilot, the systems have potentially different emphases in measuring learning targets, and we hope that the innovation itself will improve achievement—but that the systematic effects of the assessment system on the achievement estimates likely will be larger than the effects of form, mode, device, and year that we see in our current assessment systems.

The unit of analysis for evaluating comparability must be at the school and subgroup levels, given the school accountability purposes of the assessment results. However, because the subgroups may involve small sample sizes, the tolerance for comparability needs to be greater for the subgroup analyses compared to the school level analyses. If school or subgroup differences across systems

¹³ e.g., DePascale, C., Dadey, N., & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices*. Retrieved from:

http://www.nciea.org/publication_PDFs/CCSSO%20TILSA%20Score%20Comparability%20Across%20Devices.pdf

are detected, the state should evaluate the practical implications of those differences for decision making within the accountability system. Figure 18 presents a series of questions that could determine whether or not the levels of comparability seen are appropriate for the intended purposes:

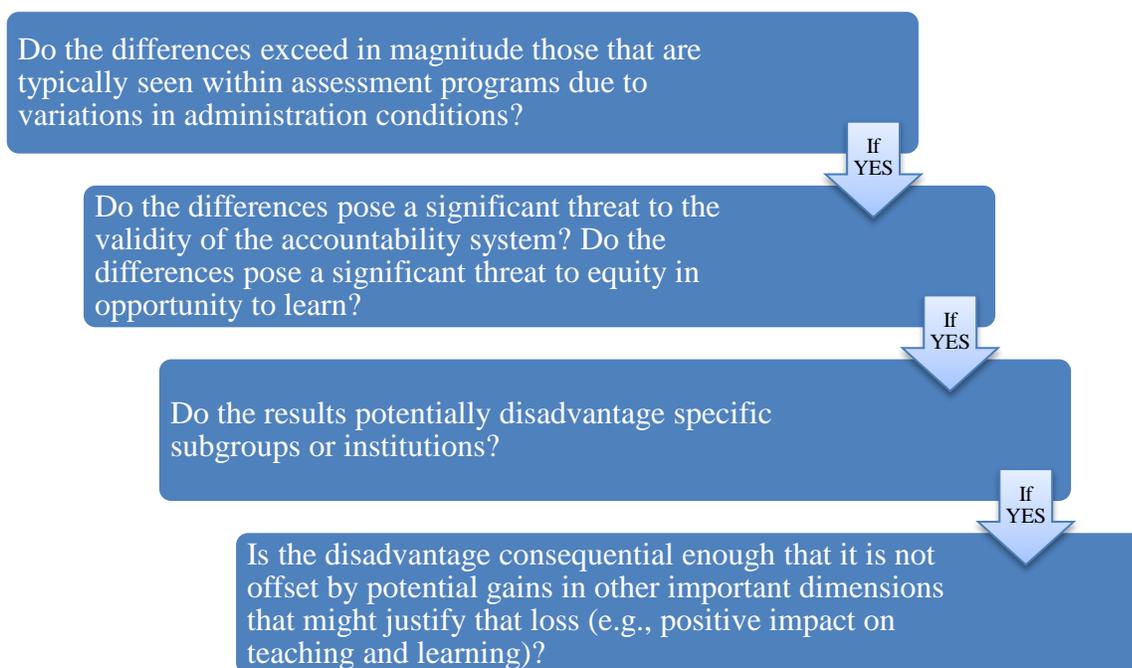


Figure 18. Decision Tree for Determining Degree of Comparability Achieved

If the answer to any of these questions is “no”, the assessment systems can be considered comparable enough to support their intended uses for the duration of the pilot. However, in the case where all of the answers above are “yes,” additional steps will need to be taken to improve the comparability of the achievement classifications to support their use in the statewide accountability system. To do so, the performance standards on either one of the assessment systems can be shifted or adjusted (such as equipercentile linking) to produce useable results for the duration of the demonstration authority, after which, standards can be re-set.

The first few years of the pilot are arguably the most important for demonstrating that results across pilot and non-pilot districts are comparable enough. As the innovation reaches critical mass and spreads across the state, comparability across the two assessment systems becomes less important than the comparability of results among districts within the innovative system of assessments.

The following evidence is present to support the comparability of the PACE pilot to the statewide assessment: A) the use of the common ALDs, B) common accommodations guidelines, C) consistency in percent proficient across assessment systems, D) concurrent comparability evaluations, and E) non-concurrent comparability evaluations.

A. Common ALDs and ALD Development Process

Achievement level descriptors (ALDs) are exhaustive, content-based descriptions that illustrate and define student achievement at each of the reported performance levels. ALDs are used to set criterion-referenced performance standards (i.e., cutscores) for an assessment program. One of the goals of the PACE project is to provide annual determinations that can be comparable across districts and between PACE and non-PACE districts. One of the ways to help instantiate this goal was to use the Smarter Balanced ALDs as the basis for the NH PACE ALDs. Therefore, this section describes the close relationship between the PACE and Smarter Balanced ALDs resulting from the PACE ALD development process (Figure 19).

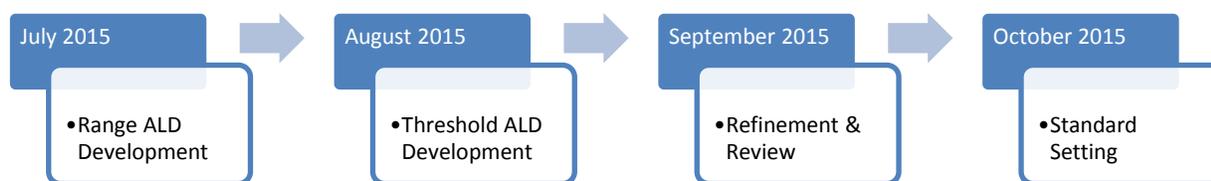


Figure 19. ALD Development Timeline

July 2015

On July 7, 2015 approximately 40 subject matter experts (SMEs) gathered to begin the Achievement Level Descriptor (ALD) development process. The SMEs were mainly comprised of teachers and administrators from five of the participating PACE Districts: Concord, Epping, Rochester, Sanborn, Souhegan. The ALD development process began with groups of SMEs working together to draft subject-specific range ALDs for each grade span. Range ALDs are typically developed at the beginning of a testing program to aid in item writing; in this case, the purpose of the range ALDs was to assist in the creation of the final threshold ALDs that would ultimately be used for standard setting. Range ALDs are designed to describe the knowledge, skills, and processes that students are expected to have by the end of the grade span at each of the four achievement levels. Before splitting into work groups, Drs. Christina Schneider, Scott Marion, and Jeri Thompson delivered training on ALDs that addressed their purposes and the desired outcome.

August 2015

On August 11, 2015 approximately 90 SMEs gathered in at Sanborn High School to draft the threshold ALDs to be used for standard setting. After training that covered the details of the standard setting method and the purpose and function of threshold ALDs within that process, teachers divided into grade-level teams by content area. The teachers used the PACE range ALDs developed on July 7th along with the Common Core State Standards (CCSS) and the Smarter Balanced threshold ALDs as resources to guide their work.

September 2015

After the two ALD development days in July and August, Center for Assessment staff reviewed the submitted work and refined the ALDs so that the format, language, and content were consistent and coherent within and across grade levels. Drs. Scott Marion, Jeri Thompson, and Susan Lyons took the lead on the Science, ELA, and Math content areas, respectively. Once the Center for Assessment content leads were satisfied with the threshold ALDs, a series of internal and external reviews were completed. As part of the external review, teachers and administrators who had participated on August 11th were given an opportunity to comment on the finished ALDs.

Because the development process, as described in this document, involved a high degree of local SME judgment and collaboration, the resulting ALDs are distinct from the Smarter Balanced (SBAC) threshold ALDs. Most notably, the PACE ALDs are not divided by groups of targets, but rather are written to describe student achievement in a more holistic manner. That being said, because the PACE and SBAC ALDs are both explicitly rooted in the CCSS, the similarity between the two sets of ALDs is clear. Appendix H provides snapshots of the ALDs for Grade 3 ELA. The content that is similar or identical across the two ALDs is connected with blue arrows.

October 2015

The final threshold ALDs were used in the standard setting as a critical step in defining the competency score ranges for the 2014-2015 PACE Annual Determinations. Center for Assessment associates regard the use of the threshold ALDs in the standard setting a success because teachers' placement of students into achievement levels were in high agreement with the student competency scores. See Table 24 below for correlations between achievement level placement based on ALD descriptions and average competency scores.

Table 24.

Correlations between ALD Placement & Competency Scores (2014-2015)

	ELA	Math	Science
Epping	0.777	0.641	0.734
Rochester	0.667	0.697	0.639
Sanborn	0.667	0.706	0.773

B. Common Accommodations

The PACE districts have established the following standards for administering accommodations on their local and common assessments. These standards are consistent with approved accommodations for other state-level assessments, including Smarter Balanced and NECAP. This coherence increases the comparability of results across assessment systems for students with disabilities and English learners.

Accommodation Standards for Common Summative Assessments	
Content Area	Approved Accommodation
Reading/ English Language Arts	No portion of the reading summative may be read (unless the summative requires a section to be read to ALL students being assessed). Written responses are allowed to be scribed* if in a student’s IEP/504 and/or ELL Plan AND if doing so does not impact the results of what is being assessed. ALL students can utilize word processing for written responses. ELL students may use a bilingual dictionary. Colored overlays, filters, or changes to lighting may be used. Students may use a ruler or writing utensil to track the text.
Mathematics	Text can be read, but symbols and numbers are not allowed to be read. Written responses are allowed to be scribed* if in a student’s IEP/504 and/or ELL Plan. ALL students can utilize word processing for written responses. Bilingual dictionaries may be used. Use of tools (calculators, number charts etc.) are only allowed if the summative assessment permits the use for ALL students.
Writing	Text can be read and graphic organizers provided, if in a student’s IEP and/or ELL Plan. Written responses may be scribed* if necessary. ALL students can utilize word processing for written responses. Students may have access to a dictionary, including a bilingual dictionary for ELL students, unless the assessment specifies otherwise.
Other Content Areas	Text can be read and written response scribed*, if in a student’s IEP/504 and/or ELL Plan. ALL students can utilize word processing for written responses.
Location	Any student can be assessed in an alternate location. ELL students may benefit from a location where they may read the assessment material out loud to themselves.
Time	Any student can have extended time, except in cases where reading fluency is being assessed. ALL students may take breaks when appropriate.
Number of Questions	Reducing the number of questions being assessed is not allowed. If this is required, it is considered to be a modification of the assessment, which means the student’s IEP reflects that his/her progress is reported through an off grade-level report card.
Changes to Font Size/Color	Allowed in all content areas for all students.

Reorganization of Questions	Any student can have the questions reorganized. For example, you may want to chunk all questions associated with one competency. You may choose to give all these questions at one time and then, the other questions at a different time. The key is that all parts of the assessment are administered.
------------------------------------	--

**Refer to the Scribing Standards document. These protocols must be applied when scribing.*

In addition to the table above, it is important to keep in mind your district’s definition of the terms grade-level and off grade-level. A student’s progress is measured to grade-level competencies unless the student has in his/her IEP the modification that he/she is working towards off grade-level competencies. In addition, one needs to distinguish the difference between instruction and assessment administration. As a teacher plans for and delivers grade-level content he/she uses differentiated instructional methods, but has the same learning target in mind for all grade-level students. The teacher scaffolds the learning for these students, which in some cases may require teaching off grade-level material in order to fill in gaps in the student’s learning, however, the goal and assessment for this student is still the grade-level material.

All students benefit from the use of highly effective instructional strategies as well as being taught how to use tools for their learning. Some examples include using graphic organizers to write, learning how to identify key words/phrases and then, highlighting/underlining them. These are good strategies and ones that we hope are in regular use throughout each classroom.

PACE Scribing Standards

Guide for the Scribe:

- Scribing is an accommodation that allows a student to access the general assessment and does not in any way alter the assessment expectation or production.
- The role of the scribe is to write exactly what a student dictates.
- Scribes may not question or correct student answers.

Scribing Procedures:

For All Content Areas

- For multiple-choice questions, the student may use his/her preferred mode of communication to indicate the correct answer choice, including, for ELL students, the student’s 1st language; a bilingual dictionary may be used; the scribe will then select the corresponding answer. For ELL students, a bilingual scribe should be used to the extent possible.
- For constructed response questions the scribe may handwrite, type, or use a computer.
- A scribe may draw a graph, diagram, or picture for the student as described by the student. *Note: in the case of an ELL student whose learning plan indicates scribing, the description may be in the student’s first language. A bilingual dictionary may be used.* The scribe will ask the student to edit the drawing. The scribe will ask the student to indicate if there are any changes they would like made.

- Students may proofread written answers and decide to edit punctuation or make changes to capitalization or spelling. The scribe will make all requested edits, even if incorrect.
- The student may dictate more than one sentence at a time and add punctuation after the fact when given the scribed sentences to proofread.
- After the scribe records the student's answer, the scribe shows the student the written response, and asks her/him to indicate if there are any changes to be made.

When Writing is Being Assessed

- The scribe will not punctuate, capitalize, or make any edits; the student will proofread to add punctuation, capitalization, capital letters, and other edits. The scribe will make student requested changes, even if incorrect.
- Students may punctuate as they dictate. For example, when stating the sentence, "The cat ran", the student can say, "The cat ran period."
- The scribe reads every word that is three or more letters long and has the student dictate precise word spelling, recording exactly as the student dictates. The scribe spells all one or two words as pronounced by the student and does not probe these words.

When Writing is Not Being Assessed

- The scribe will use correct spelling and add punctuation and capital letters.

For Mathematics

- The student must indicate operational signs (e.g., addition, subtraction).
- The student must be specific in terms of what numbers to write down with regard to position. The scribe will ask the student to indicate exactly where the numbers need to be placed. For example, when adding 37 and 8, the student can indicate 7 plus 8 is 15 by stating "put down the five and carry the 1".

Note: For ELL students whose Learning Plan indicates scribing, the scribe should be familiar with the student's 1st language math conventions. Different countries have different conventions for mathematical operations. One example is the European convention of using commas where we use periods and vice versa:

$$14.000 = 14,000$$

$$14,0 = 14.0$$

Upon Completion of Scribing Activities the Scribe will:

- Allow student to review responses and indicate needed changes or revisions.
- Update and provide for a final review by the student.

Thank you to the State of Washington, Office of the State Superintendent for allowing us to utilize many aspects of their adopted Scribing Protocol.

Approved NH DOE 4.2015

Performance Assessment for Competency Education Accommodation Guidelines for English Language Learners

PACE has established the following accommodation guidelines for English Language Learners, excerpted and adapted from Smarter Balanced Assessment Consortium.

Construction of Performance Tasks

For English language learner students (ELLs) who take large-scale content assessments, the most significant accessibility concern is associated with the nature of the language used in the assessments. Because ELLs have not yet acquired complete proficiency in English, the use of language that is not fully accessible to them in assessments will degrade the validity of the test score interpretations that can be inferred from their results. The following guidelines should be considered when designing performance tasks:

- Design test directions to maximize clarity and minimize the potential for confusion.
- Use vocabulary in test items that is widely accessible to all students; avoid unfamiliar vocabulary that is not directly related to the construct (August, Carlo, & Snow, 2005; Bailey et al, 2007).
- Avoid the use of syntax or vocabulary that is above the test’s target grade level (Borgioli, 2008). The test item should be written at a vocabulary level no higher than the target grade level, and preferably at a slightly lower grade level, to ensure that all students understand the task presented (Young, 2008).
- Keep sentence structures as simple as possible while expressing the intended meaning. ELLs will find a series of simpler, shorter sentences to be more accessible than longer, more complex sentences (Pitoniak, Young, Martiniello, King, Buteux, & Ginsburgh, 2009).
- Avoid false cognates, which are word pairs or phrases that appear to have the same meaning in two or more languages, but in fact, do not. Examples of false cognates include: billion (the correct Spanish word is mil millones; not billón, which means *trillion*); deception (engaño; not decepción, which means disappointment).
- Do not use cultural references or idiomatic expressions (such as “being on the ball”) that are not equally familiar to all students (Bernhardt, 2005). This includes questions related to sports (yards, quarterback, etc.) which could be considered culturally biased questions for ELL students.
- Avoid sentence structures that may be confusing or difficult to follow, such as the use of passive voice or sentences with multiple clauses (Abedi & Lord, 2001; Forster & Olbrei, 1973; Schachter, 1983).
- Do not use syntax that may be confusing or ambiguous, such as using negation or double negatives in constructing test items (Abedi, 2006; Cummins, Kintsch, Reusser, & Weimer, 1988).
- Minimize use of low-frequency, long, or morphologically complex words and long sentences (Abedi, 2006; Abedi, Lord & Plummer, 1995).

Excerpted from: Young, J.; Pitoniak, M.; King, T.; & Ayad, E. (2012) *Smarter Balanced Assessment Consortium: Guidelines for Accessibility for English Language Learners. Measured Progress/ETS Collaborative.*

Examples of effective instructional strategies for ELL students preparing for the PACE Assessments include:

- Teaching word learning strategies, especially the use of cognates.
- Providing sentence and paragraph frames with word banks.
- Teaching strategies to use visual cues in text to support meaning (e.g., pictures and diagrams, titles and subtitles)
- Allowing students to compose and discuss their initial ideas for writing in their first language; once they’ve figured out what they want to write, have them complete the



- finished product in English.
- Providing instruction in common assessment word and phrases (e.g., what best describes, select, mark, summarize, support with examples), and help students understand what types of responses will be expected for each.

Accommodations for English Language Learners during Assessment Administration

- *Read Aloud*
 - Read aloud of test directions in student’s native language
 - Read aloud of test questions (Math, Science, History/SS) to student by teacher or electronic media
- *Test Setting and Time*
 - Test in a familiar environment with other ELLs
 - Small group setting
 - Test Break
 - Extra time within the testing day
- *Use of Dictionaries and Other Resources*
 - Customized Dictionary/glossary in English (content-related terms removed) or Bilingual Dictionary
 - Picture Dictionary (alone, combined with oral reading of test items in English, and combined with bilingual glossary)
 - Traditional glossary with 1st Language translations (content-related terms removed)
 - Computer-based test (CBT)

Excerpted from: (Abedi, J & Ewers. (2013). N. Smarter Balanced Assessment Consortium: Accommodation

C. Percent Proficient Across All Grade Levels

Figures 20 and 21 show 2015-2016 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, indicating a high degree of comparability in the rigor of the standards between PACE and non-PACE assessment. But for the colors indicating the different assessments, student performance across the two systems would be indistinguishable.



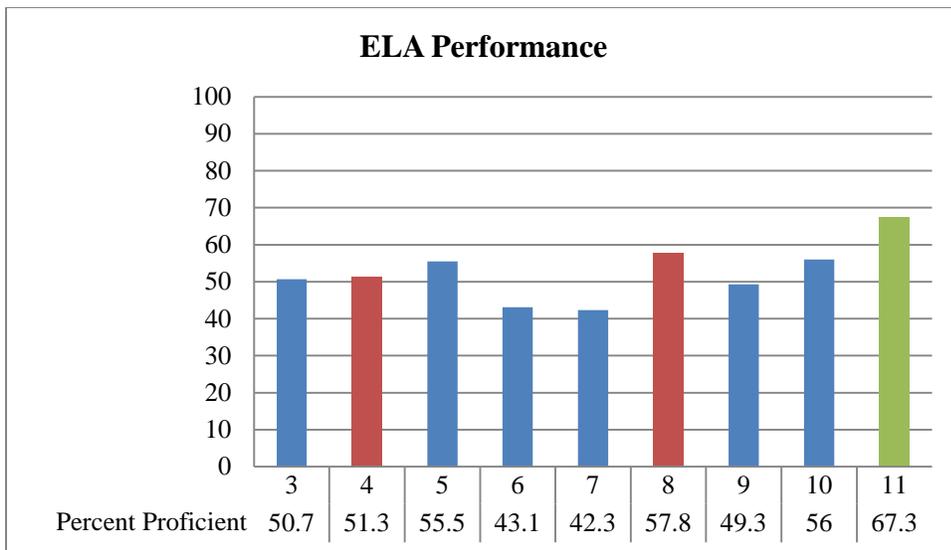


Figure 20. PACE District Performance in ELA across Assessment Systems

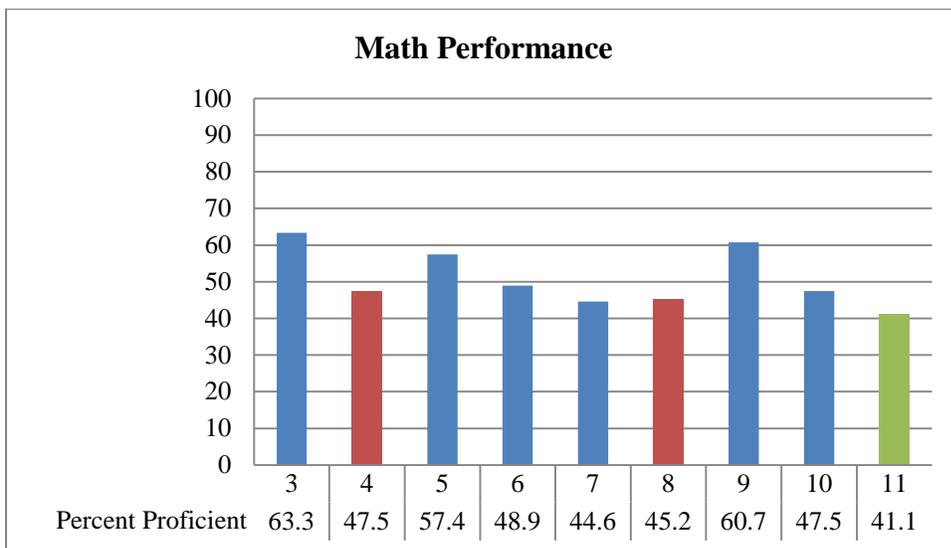


Figure 21. PACE District Performance in Math across Assessment Systems

D. Concurrent Comparability Evaluations

Importantly, the degree of comparability of the annual determinations across the two assessment systems within the state can be directly evaluated by administering an assessment that is common across the two programs to a sample of students. For example, since SBAC is administered once per grade span in grades 3-8 and SATs are administered in grade 11, the comparability of the annual determinations between pilot and non-pilot districts is evaluated by directly comparing annual determinations for the students that participated in both assessment systems. By calculating two sets of annual determinations for these students, the state has both traditional and innovative data points for some of the students in each pilot district. The degree of agreement between the two sets of annual determinations is then analyzed to provide further



evidence regarding the comparability of the interpretations of the reported achievement levels, or if systematic differences are detected, inform decisions about calibrating results to provide for comparability when appropriate.

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 22-27 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: Figures 26 and 27 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.

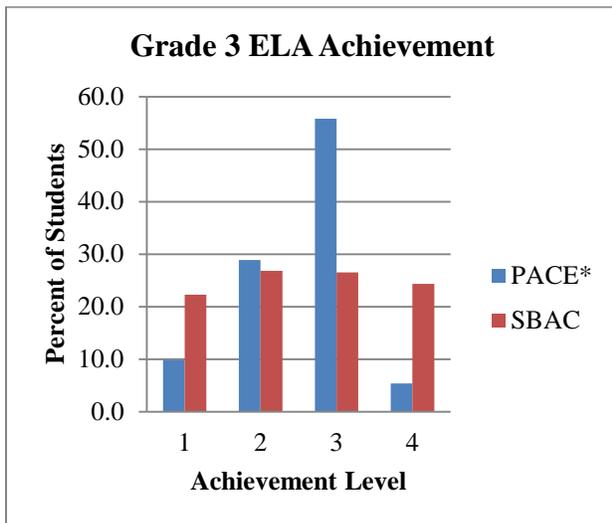


Figure 22. G3 ELA

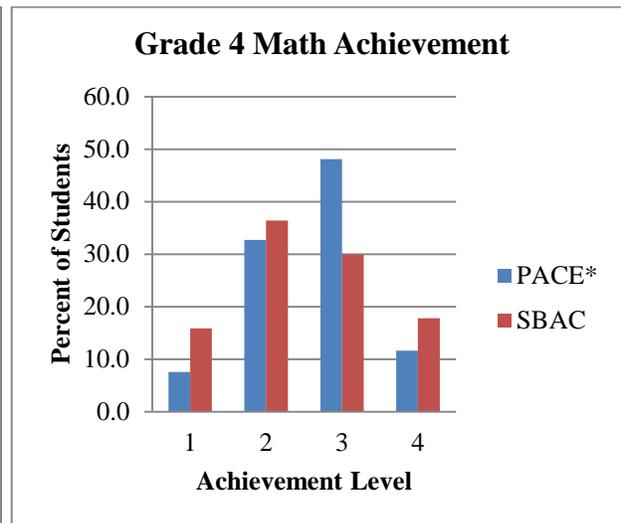


Figure 23. G4 Math



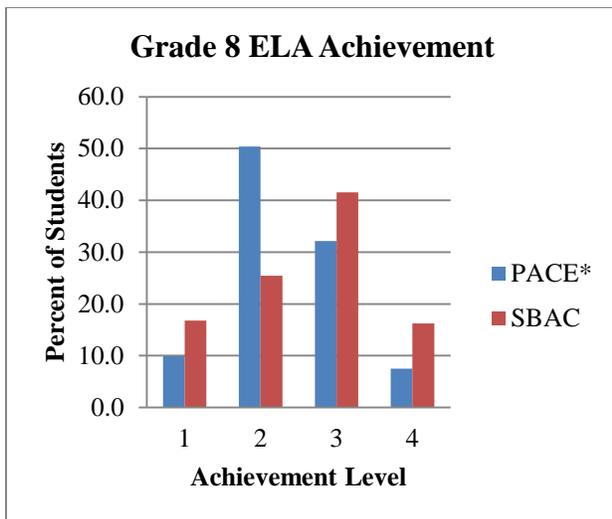


Figure 24. G8 ELA

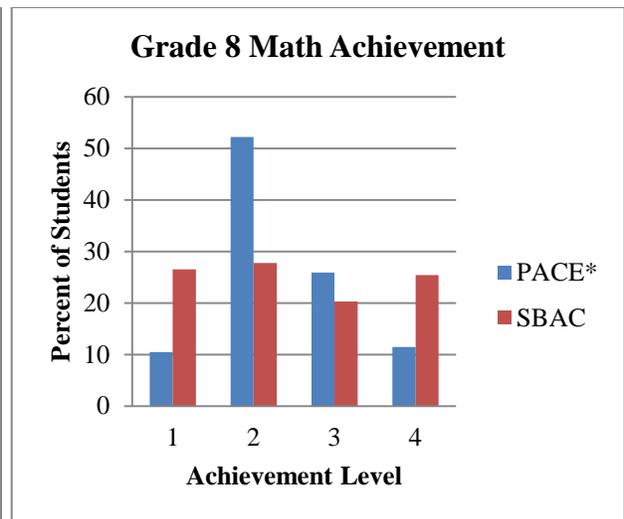


Figure 25. G8 Math

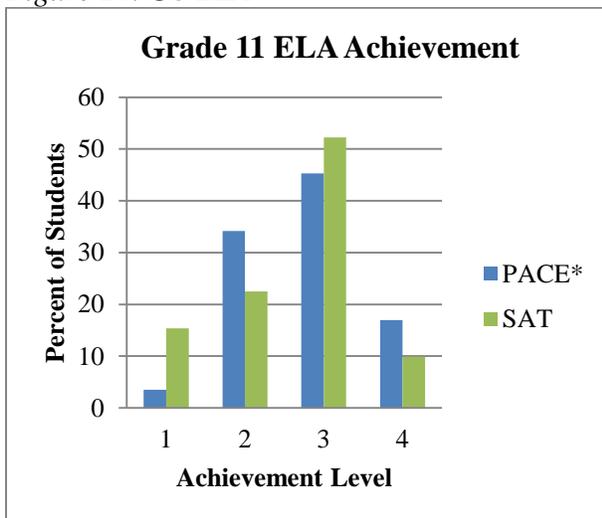


Figure 26. G11 ELA

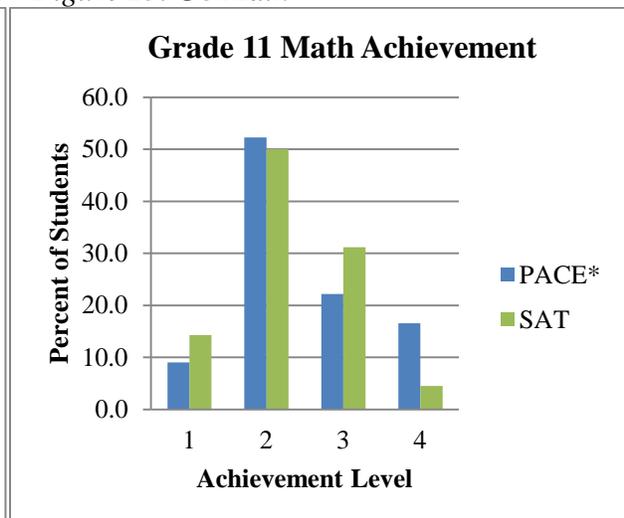


Figure 27. G11 Math

While the figures shown above are compelling, Tables 25-28 provide additional information regarding the classification accuracy by matching students across the assessment systems.

Table 25.
Classification Accuracy for SBAC ELA

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

Table 27.
Classification Accuracy for SBAC Math

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



Table 26.
Classification Accuracy for SAT ELA

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

Table 28.
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 25-28, 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 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 29 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 29.

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



E. Non-concurrent Evaluation of Comparability

I. 2015 SBAC to 2016 PACE. Since students participate in SBAC once per grade span, we have compared 2014-2015 performance on SBAC with 2015-2016 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 2015 and a PACE achievement level in 2016 are used for these analyses. Figure 28 shows the percent proficient for the matched cohort of students across years. In three out of the four grades and subject areas, the percent proficient rose from 2015 to 2016. Additionally, it seems that the percent proficient in ELA is more stable across the two years and systems of assessments than in math. This finding is discussed in more detail below.

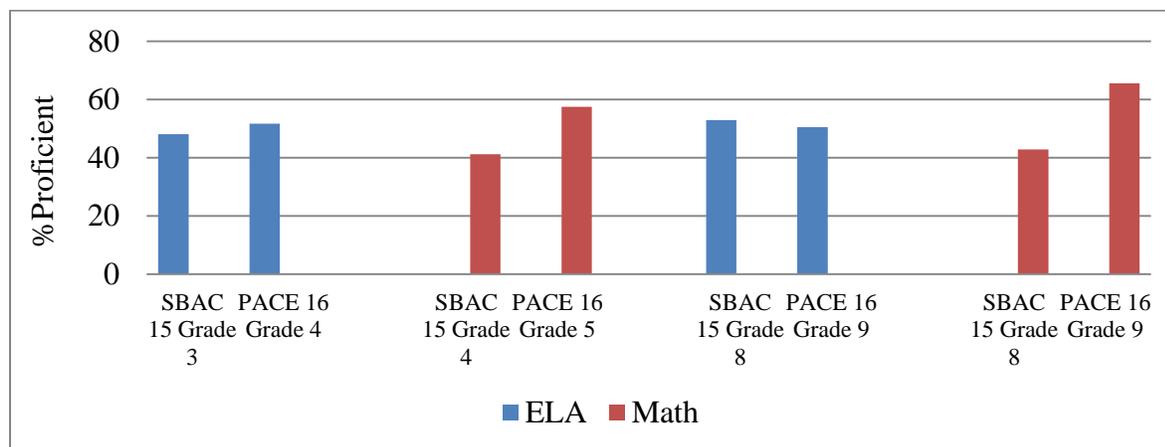


Figure 28. Cohort %Proficient across years and assessment systems

As was done with the concurrent comparability analyses, the 2x2 classification tables are provided in Tables 30-33. “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 30.

Classification Accuracy for G4 ELA

		Proficient on PACE	
		No	Yes
Proficient on SBAC	No	36.2%	15.6%
	Yes	11.5%	36.6%

Table 32.

Classification Accuracy for G5 Math

		Proficient on SBAC	
		No	Yes
Proficient on SBAC	No	36.2%	22.6%
	Yes	6.3%	34.9%

Table 31.

Classification Accuracy for Grade 9 ELA

		Proficient on PACE	
		No	Yes
Proficient on SBAC	No	31.4%	14.8%
	Yes	18.0%	35.8%

Table 33.

Classification Accuracy for G9 Math

		Proficient on PACE	
		No	Yes
Proficient on SBAC	No	27.7%	24.3%
	Yes	6.5%	41.5%



As would be expected, the classification accuracies across years are slightly lower than the classification accuracies observed for the concurrent year comparisons, ranging from 67.2% for Grade 9 ELA, to 72.8% for Grade 4 ELA. The pattern in the change in achievement does not seem to be consistent across the subject areas. While the observed differences in proficiency classifications for ELA is fairly evenly distributed between students moving from proficient to non-proficient and students moving from non-proficient to proficient, the same does not seem to hold for math. More students are moving from non-proficient on the 2015 Smarter Balanced assessment to proficient on the 2016 PACE assessment than in the other direction for math. Since this pattern was not observed in the concurrent analyses, it could be that this observed change is more reflective of true changes in achievement the assessment system. However, this pattern is certainly something we will continue to closely monitor in the coming years.

Table 34 shows the proficiency classification accuracies for the waiver-reported subgroups for the cross-year analysis. These statistics are disaggregated by subject but not by grade level in order to increase the cell sample sizes. 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 English learners in ELA. We will pay particular attention to this subgroup in next year's analyses to ensure this is not indicative of something systemic.

Table 34.

2015 SBAC to 2016 PACE Proficiency Classification Accuracies for Subgroups

	ELA	Math
American Indian or Alaskan Native	88.2%	68.4%
Asian	73.5%	68.8%
Black or African American	69.4%	74.8%
Hispanic or Latino	68.3%	67.3%
Native Hawaiian or Pacific Islander	**	**
Two or more races (non-Hispanic)	70.0%	69.2%
White	71.2%	72.9%
WaiverSubgroup - SWD and EL - Not EconDis	**	
WaiverSubgroup - SWD and EconDis - Not EL	88.3%	83.6%
WaiverSubgroup - SWD and EconDis and EL	**	**
WaiverSubgroup - Students With Disability(SWD) only - Not EconDis, Not EL	73.6%	71.1%
WaiverSubgroup - Eng Learner (EL) only - Not EconDis, Not SWD	57.7%	67.9%
WaiverSubgroup - EconDis and EL - Not SWD	73.3%	76.9%
WaiverSubgroup - Economically Disadv (EconDis) only - Not SWD, Not EL	69.0%	66.1%

2. 2015 PACE to 2016 SBAC. Since students participate in SBAC once per grade span, we have compared 2014-2015 performance on PACE with 2015-2016 performance on SBAC for students in grade 8 in ELA, and in grades 4 and 8 in Math. Figure 29 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.

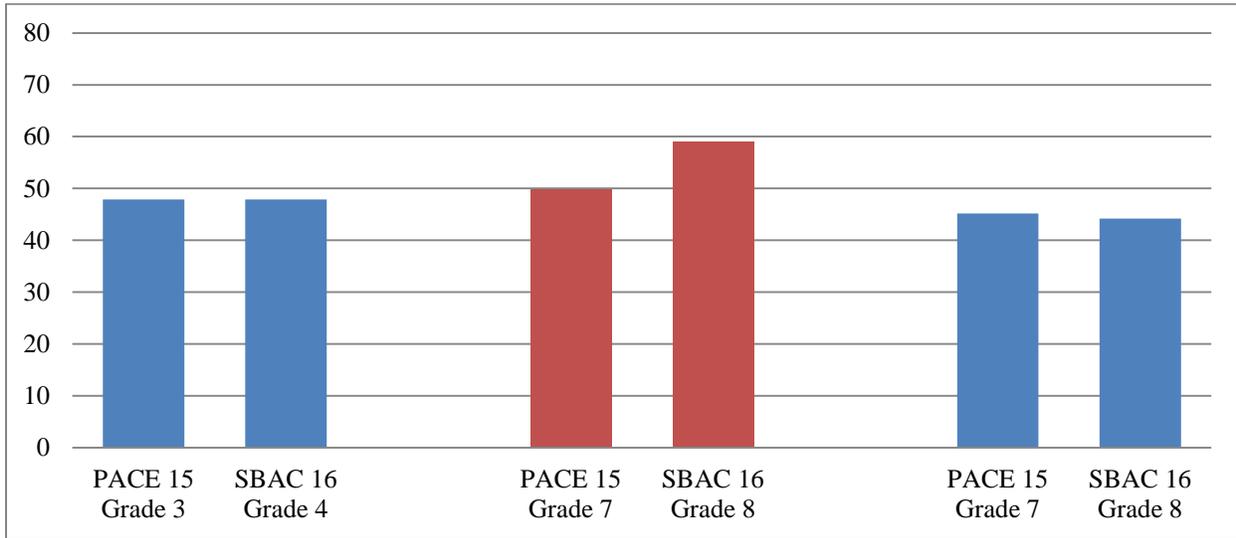


Figure 29. Cohort %Proficient across years and assessment systems

Classification tables are provided in Tables 35-37. “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 35.
Classification Accuracy for G4 Math

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

Table 36.
Classification Accuracy for G8 ELA

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



Table 37.
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 38 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 38.

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 39-41 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 39.

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 40.

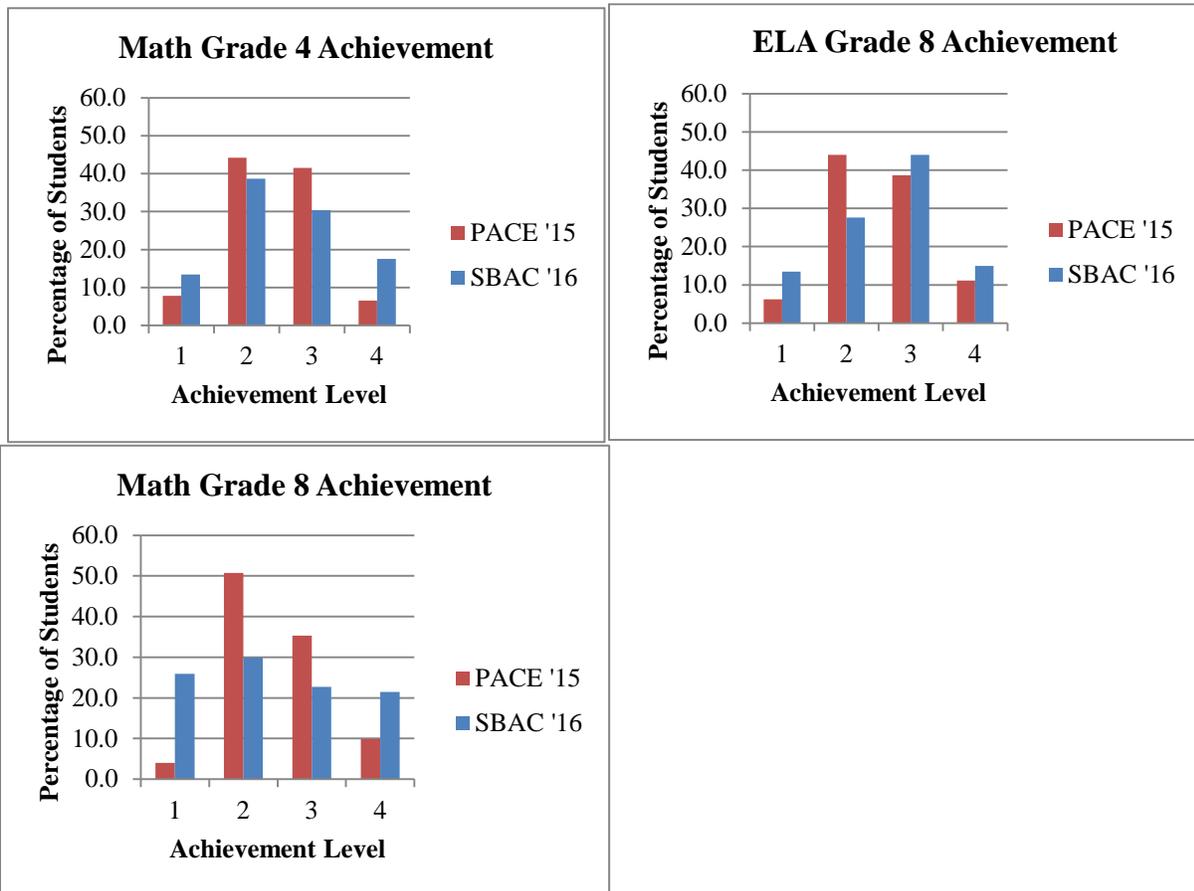
Crosstabs (n counts) for 2015 PACE and 2016 SBAC Math

		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 41.

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 $r = 0.630$.¹⁴ 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.

Summary

The intended uses and interpretations of PACE assessment system results are supported based on all the evidence presented on the comparability of accountability determinations within districts, among PACE districts, and between the two state assessment systems. There is also additional evidence that supports the validity of the PACE assessment system results—two external evaluations of the NH PACE pilot. The first was conducted by HumRRO starting in 2016 and the other is currently underway and examines the first two years of the PACE pilot (2014-2016). The next section details those two external evaluations and how their findings support the validity of the PACE system.

¹⁴ See the HUMRRO evaluation report (included later within Technical Manual) for these analyses.



External Evaluation of System Success

HumRRO Executive Summary Report (2016-2017)

HumRRO conducted several data collection activities over the course of the evaluation. These included interviews with nine PACE District Leads; visits to schools in eight PACE districts to conduct interviews or focus groups with administrators, teachers, parents, and students, as well as classroom observations; observation of cross-district meetings including task development sessions and scoring and calibration sessions; participation in monthly PACE Leads Meetings; and review and analysis of scoring and calibration data. In addition, we administered a teacher survey to all teachers in Tier 1 districts, in part to help determine the generalizability of our findings from the teacher focus groups.

Snapshot of Key Findings

Buy-in

One of the most challenging requirements for the success of any educational intervention is securing buy-in from the major participants and leadership of classrooms, schools, and districts. PACE addresses this challenge in several ways. First, educators are in charge of nearly all aspects of the program. Teachers decide what is assessed, how it is assessed, and how the tasks are scored. By placing the responsibility for creating the tasks on the primary users of the assessment data, PACE gives teachers more say in how their students will be assessed than in more traditional testing systems.

The second way PACE gains buy-in is by emphasizing the integrated nature of the assessments. Unlike end-of-year comprehensive statewide assessments, which sample from the past year's curriculum, PACE is targeted to the learning that is occurring at the time of administration. Since there is no specific testing window for PACE, and since the tasks are targeted to one broad curricular topic, teachers can administer the tasks when it makes the most sense. There is no need for intensive review during the weeks leading up to the testing window and no post-test slump between the end of the testing window and the end of the school year.

A third reason PACE participants are committed is that PACE replaces the Smarter Balanced assessments in the grade/subjects for which it is administered— an assessment that many New Hampshire educators regard as an interruption of their instruction that provides little useful information. PACE tasks require deep knowledge on the part of students. There is no chance of getting an answer correct by guessing. Students actually perform the tasks on which they are assessed, rather than answer questions about those tasks.

Collaboration

Participating districts reported a high degree of collaboration. First, educators from all Tier 1 districts meet regularly throughout the year. They participate in task development sessions, professional development, scoring sessions, standard-setting, and other meetings.

Districts also interact through the “LibGuide” system. This system is a repository for “all things PACE.” It is a web-based repository for PACE tasks, rubrics, and shared resources. Teachers who implement common tasks early share their lessons and provide tips for smoother implementation among their colleagues. The teachers share book lists that are suitable for use in English language arts tasks. They share equipment lists for science labs, including locally available inexpensive options for commonly needed equipment.



Over the course of the evaluation period, PACE implemented three key new collaboration measures:

- Naming an overall curriculum coordinator to assist with PACE task development activities.
- Naming of multiple Content Leads (about 30 total) for each grade level and content area combination. These teachers were identified as leaders in PACE and were recommended by peers and ultimately selected by the PACE District Leads to help coordinate subject/grade-specific activities.
- The third new innovation is the “buddy district.” Districts are now paired with other districts to promote collaboration. Districts with Content Leads are often paired with districts that do not have them. Newer PACE districts are typically paired with experienced districts.

These new collaboration initiatives help PACE cope with expansion. As the program expands, these efforts become increasingly necessary to maintain the requisite levels of participation and ownership among PACE educators.

Teaching & Learning

Teachers across districts expressed that PACE has had a positive impact on increasing the depth of knowledge at which they teach and gives them real-time feedback that they can use to make “on-the-spot” adjustments to their instruction to better meet the needs of their students.

Unlike most large-scale assessment systems, which are focused on the estimation of student and/or school performance, PACE is also intended to influence instructional practices. PACE leadership is not overly concerned about teachers “teaching to the test.” PACE, ideally, supports “testing to what is taught.”

PACE also represents a shift for students. Typically, students learn content prior to the tests and then demonstrate their learning through their performance on the tests. PACE certainly has similar aspects, but because of the integrated nature of the assessments, students learn while testing as well. PACE tasks often require multiple classes to complete and might involve several steps (e.g., reading a novel, discussing the characters and their motivations, then writing a response to a prompt related to the novel). Because of the integrated nature of PACE, testing and learning are not entirely separate components of a student’s day.

Context

While there are several contextual factors influencing the quality of PACE implementation worth mentioning, the largest stems from implementing PACE at the district level. Districts vary in their capacity, student populations, and in the expertise and experience of their staff members. Early adopters of competency-based education had a significant advantage in implementing PACE. They already had a collection of locally developed tasks from which to start and were familiar with the design of competency-based rubrics. In many cases, their students had largely become accustomed to the kinds of tasks PACE requires.

District size plays an important role in PACE implementation as well. Smaller districts typically have only one teacher per grade/subject. In some cases, there may be only one teacher per grade; in elementary school this teacher is responsible for ELA, mathematics, and science tasks. This means that all of the work associated with developing and administering the local tasks is



concentrated among very few people. Smaller districts often have to solicit help from outside the district to conduct double scoring.

Larger districts have more support staff and typically have same-grade/subject teachers who can work as teams within districts, or even within the same school. This does not always mean that the teachers in larger districts have less work, however. The more students in a school who take a PACE assessment, the larger the effort required for scoring. A very small district might only have 10 students who complete a task. A larger district could have a few hundred students completing a task.

PACE was implemented, in part, to reduce perceived negative consequences associated with large-scale, end-of-year standardized testing. PACE was designed to stave off reductions in the depth of learning of students, to promote critical thinking, and to integrate curriculum, instruction, and assessment into a cohesive system of education.

But PACE requires a tremendous amount of work on the part of teachers. While most teachers were very supportive of PACE, it was not uncommon for them to comment on the time and effort required to implement the program, including development of tasks and rubrics as well as task administration and scoring. Survey results indicate that approximately one fourth of respondents did not think that the time and effort required by the PACE initiative was worth the benefits.

Recommendations

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.

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.



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.

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.

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.

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.

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.

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.



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.

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.

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.

The PACE Story

PACE has lofty ambitions. Ideally, PACE will lead to an integrated competency based education system that is unbound by time in class, age, location where learning takes place, and other artificial methods of categorizing students. Instead, the system would focus on a core set of competencies and move students to the next phase of their education irrespective of when, where, or how the student achieves those competencies. The system will incorporate a large number of ways for students to demonstrate the competencies, and demonstration will take place in an on-demand way, where students can choose to complete a performance event (not necessarily limited to the current task format) when they are ready, rather than on a school calendar. Instruction would be more individualized and targeted toward the next competency the student needs to master. Such a system would represent a dramatic shift from the traditional system of schooling.

PACE, as it is implemented currently, has taken steps toward this ideal. The PACE districts have begun identifying important competencies and they have designed performance tasks to measure those competencies. They have begun to build a bank of high-quality performance tasks that can be drawn on throughout a student’s academic preparation. They have moved toward a more integrated system of curriculum, instruction, and assessment. Assessment is being woven into all aspects of teaching and learning, and the consideration of assessment when planning curricular sequence and planning lessons have increased among teachers since joining PACE. Students, even those who don’t like PACE, describe the tasks as complex and difficult, but as strong measures of their knowledge, skills, and abilities.



But there is still a long road ahead if PACE is to realize all of its bold goals. First, PACE has to prove to be sustainable. The program is relatively new and a few highly-motivated districts have been instrumental in implementing the system. As new districts join PACE, there will be challenges. Getting new staff members oriented to such a complex new way of educating students takes considerable time and effort. If the experienced teachers train the new ones, they will need time to do so.

The sustainability of PACE will rely on demonstrating that the benefits of PACE continue to outweigh the challenges. For this to happen, PACE will require continuous feedback and improvement as the system expands.

In addition to sustainability, PACE must also prove that it is scalable. New districts are joining PACE, but NH DOE recognizes the considerable challenges involved in scaling PACE statewide as it is currently conceived. PACE is currently adopted at the district level. This is, in part, because New Hampshire districts are extremely autonomous. It is, after all, the “Live Free or Die” state.

In New Hampshire, PACE began with a few highly motivated districts and is expanding carefully. This model seems to be effective for a system like PACE, and if the system is transported outside New Hampshire, other states may want to adopt a similar implementation plan.



Effects of PACE on 8th Grade Student Achievement Outcomes (2014-2016)

A Ph.D. candidate at the University of New Hampshire is currently investigating the effects of the NH PACE pilot on 8th grade student achievement outcomes in English language arts (ELA) and mathematics for her dissertation. The dissertation compares PACE student achievement with demographically similar non-PACE comparison student achievement using Smarter Balanced achievement test results from the first two years of the PACE pilot (2014-15 and 2015-16 school years). Findings will provide empirical evidence of the average effect of the PACE pilot and the extent to which those effects vary according to student-level characteristics such as gender, disability status, free- and reduced-price lunch status, and prior achievement. The dissertation is not yet complete, however, preliminary findings suggest that there are positive effects on 8th grade student achievement outcomes in ELA and math starting in Year 2 of the pilot (2015-16 school year) and that students with disabilities attending PACE schools tend to perform significantly higher than their IEP counterparts in non-PACE schools. These initial findings suggest that PACE students are provided an equitable opportunity to learn and are benefiting from the assessment system. The full paper will be provided to the NH DOE when complete.



Appendix A: NH PACE Readiness Tool

NH PACE READINESS TOOL

			STUDENT WORK.	NITIES ESTABLISHED FOR REVIEW OF
--	--	--	---------------	----------------------------------

NHPACEJUNE2016 *EHH*

Appendix B: NH PACE Task Development Template

 <h1 style="margin: 0; font-size: 2em;">NH PACE</h1> <h2 style="margin: 0; font-size: 1.5em;">Performance Assessment of Competency Education</h2> <h3 style="margin: 0; font-size: 1.2em;">Performance Task Framework 2016-2017</h3>					
<input type="checkbox"/> LOCAL TASK	<input type="checkbox"/> COMMON TASK	<input type="checkbox"/> In Development	<input type="checkbox"/> Reviewed #1	<input type="checkbox"/> Reviewed #2 (NCIEA)	<input type="checkbox"/> FINAL APPROVAL
Performance Task Name <i>Unique name given to this performance task</i>					
Content Area <i>For example: ELA, Science, Math, Social Studies, etc.</i>					
Grade-Level <i>If this is a middle or high school task, indicate grade level and course name if applicable</i>					
NH State Model Competencies: Task Targets <i>List each NH State Model Competency that will be assessed through this task; these are one or two <u>primary</u> task targets</i>					
Contributing Author(s) <i>List the names, emails, and schools or agencies of ALL contributing authors in the task.</i>					
Citations/Attributions <i>If this task is an adaptation of work published elsewhere, list all citations/attribution. Permission to include copyrighted work must be obtained by the author(s) listed above from the originator of the adapted work and documented here.</i>					
Performance Task Description <i>Describe the performance task in detail, specifying the context for the task, the anticipated student activities, products and/or presentation and resources, texts, scaffolding, and materials needed. What will the students be asked to do, to produce, and through what actions will they demonstrate mastery of the target competencies? Refer to the NH PACE Accommodations and ELL Guidelines in ensuring that the construction of the task leads to activities that are accessible to all students.</i>					
Standards Addressed in the Performance Task					
Source of Standards: <i>List the</i>			Standards: <i>List the complete wording of the target standards associated with the key</i>		

<p>document(s) from which the standards are drawn i.e. CCSS, NH State Frameworks, NGSS, etc., including any locally developed competencies or standards.</p>	<p>competencies included above (may copy & paste). There should be a direct and obvious alignment between the standards and the competencies.</p>
<p style="text-align: center;">Rubric(s) Used in Assessing this Task</p> <p><i>Include all rubrics to be used in the assessment of students' proficiency with this performance task. Be specific in the description of the student product(s) and activities to which the rubric will be applied. Cut and paste or upload the rubric document here. Annotate the rubric to make clear which standards and competencies are aligned with each scoring dimension. Rubrics adapted to student-friendly language should be included in the student instructions section. However, they should align with teacher-use rubrics included here.</i></p> <p><i>Listing which part (activity and/or product) of the task is used for assessment through the rubric assists in comparable administration across districts and replication of the task by various educators.</i></p>	
<p>Student Activities/Product(s) to be scored using this rubric:</p>	<p>Rubric: (copy or upload the entire annotated rubric to this section)</p>
<p style="text-align: center;">Teacher Directions</p> <p><i>In this section, describe all directions that the teacher needs to use in the administration of all aspects of the performance task, including lesson focus and formative assessment tasks. Bear in mind that teachers other than the original author(s) will need these directions in order to administer the task. Include hyperlinks for online resources.</i></p>	
<p style="text-align: center;">Student Instructions</p> <p><i>Describe clearly and in detail all student instructions used in the administration of this performance task. Attach or upload aligned rubrics that have been adapted to student-friendly language.</i></p>	
<p style="text-align: center;">Artifacts</p> <p><i>Optional: In this section, include links to artifacts depicting student products that may be useful in gaining greater clarity of this performance task. These may be digital pictures, podcasts, websites, etc.</i></p>	



Appendix C: Think Aloud Protocol

[This think aloud protocol was written as a guide for teachers to understand the purpose and processes involved in soliciting student feedback about the quality and understandability of the performance tasks].

Cognitive laboratories, also known as “think alouds,” are a valuable and efficient way to gather feedback from students about the quality and understandability of the tasks and items we create. They are surprisingly under-utilized outside of research settings. Even large-scale assessment programs do not use think alouds enough.

The basics of think alouds are quite simple. **As the name implies, students are encouraged to verbalize their thinking while they are solving tasks.** This information produced can help us understand whether the directions to the task are clear, students are calling on the knowledge and skills we thought necessary to approach the task, and students were calling on the cognitive processes that we thought the task would require. **More formally, cognitive laboratories provide evidence related to “response processes,”** which is a one key source of evidence necessary to validate the inferences we make from task/test scores.

Think alouds should be a regular part of the PACE task development process, ideally used between the first two meetings (assuming a draft task has been produced) so that when participating teachers bring information back from their school colleagues’ reviews, they can also share information from the students’ perspective.

PACE Think Aloud Sample

1. Each participating teacher should select 2-3 students to participate. Since the tasks being designed are generally design for the end-of-course, it will be important to interview students who have had an opportunity to learn the necessary content and skills. This likely means relying on students who are in the next grade (e.g., current 4th graders to review 3rd grade tasks).
2. Importantly, **we do not want this to be “one more burden,”** so if the grade-level group feels like having each teacher conduct 2-3 think alouds is too much, the group can have each teacher agree to just one student, but to make sure that the group covers the range of student performance. That said, we think that it will take at least one think aloud to get the hang of it, so we urge everyone to try to conduct two think alouds.

PACE Think Aloud Protocol

1. The think aloud is a **one-on-one activity**, so find a **quiet place to conduct the protocol** where the student will feel comfortable working (e.g., a classroom during lunch).
2. Have **two printed copies of the task**—one for the student and one for you. The teacher copy of the task will be used for taking notes. Ideally, you would audio or video each interview, but we do not want you to have to worry about parent permissions at this time. However, if you are able to record the interview, just for your purposes without permission, we urge you to do so.
3. **Welcome the student and put them at ease** by saying something like: “Thank you so much for coming to help me today. We are really happy that you are here, and I know



you will be a big help to me.” Emphasize to the student that you are not “testing” them, but that you are trying out a task and need their help to do so.

4. **Say something like the following to the student:** *We’re going to be doing something called “think alouds.” Think-alouds involve a lot of talking, because we ask you to say out loud everything you are thinking. It feels a little silly at first to say everything you’re thinking out loud, but it will really help us. See, when we give a task to students, we don’t know what they are thinking when they see the questions, and we really want to learn. It will help us make better tasks and activities. The more you tell us about what you are thinking, the more we will understand. So, it’s important for this activity that you think out loud.*
5. We know that everyone is pressed for time, but we think having students (and you) **go through this example and practice activity** will help:
 - a. Let me give you an example of how a think-aloud works. Let’s say someone asked me how many windows are in my home. Here’s how I would answer while thinking out loud:
 - b. Let’s see...when I walk in the front door, I’m in the hallway. There are no windows in the hallway. But, there are three little windows at the top of the front door. Should I count those? I think I should. So, that’s 3 (*write down the number 3 on a piece of paper*).
 - c. Next, the kitchen is on my right. There is one big window in the kitchen plus two little windows. So, I’ll write down 3 for the kitchen (*write down the number 3*).
 - d. Then, the kitchen connects to the dining room. Hmm...there aren’t any real windows in the dining room, but there is a big sliding glass door that is sort of like a window. Should I count that? Hmm...no, I don’t think I should count a glass door as a window. So there are no windows in the dining room. Then, I move into the living room. There are two windows in the family room (*write down 2*).
 - e. Then, I go down the hallway into the bedroom, and there are two windows in the bedroom (*write down 2*). Then, there is one window in the bathroom (*write down 1*). The last room is an office, and there is one window in the office (*write down 1*).
 - f. So, all together there are $3+3+2+2+1+1 = 12$ (*show them that you are referring to the paper where you wrote down the numbers to do this*) windows in my house. So, I would tell the person who asked me the question that the answer is 12.
 - g. Finish your example by saying something like: “Do you see how a think aloud works? Now you try it. Tell me how many windows are in your house.” *Give the child time to answer. Prompt them to tell you what they are thinking if there is too much silence.*
 - h. Finish by saying, “That was great. Do you understand how to think out loud now? Do you think you can do this for me with the question I’m going to show you?”
6. **Working with the task:** Think of the protocol taking place in two phases:
 - a. Phase 1, the child thinks out loud and the interviewer uses only passive prompts to encourage the child to think out loud.
 - b. Phase 2, the interviewer asks the child specific questions to probe their understanding of the child’s cognitive process.
 - c. Phase 1 should be allowed to finish before Phase 2 starts. Phase 1 finishes with the child writing down the answer.



7. Ask the child to read the passage (for ELA and perhaps science) and directions to each part of the task aloud (all subjects). The child should read each question or part of the task aloud as she reaches it. **Note on the teacher copy of the task where the student is either struggling with the directions or interpreting them differently than intended.**
8. After they finish reading the task, **ask the student to work through the problem** while talking about their thinking like they did in the window example. **Again, try to record notes as completely as possible.** What strategies are they using? What knowledge and skills are they using? Where are they getting stuck?
9. **Here are some potential Phase 1 prompts** (*you don't need to say all of these, but reinforce good think alouds, and prompt the child to think aloud when you there is more than 5 seconds of silence*):
 - a. What are you thinking?
 - b. Don't forget to tell me what you're thinking.
 - c. You look like you're thinking hard. Can you tell me what you're thinking?
 - d. Keep going.
 - e. Now what are you thinking?
10. **Before moving onto Phase 2, make sure you praise the student for doing a great job, such as:**
 - a. Thank you so much for saying all of that.
 - b. Your explanations are really helping me understand these questions better.
11. **Here are some potential Phase 2 probes**
 - a. How did you get that answer?
 - b. What makes you believe that answer is the right one?
 - c. Was there anything that seemed tricky about this question?
 - d. Was there anything that confused you about this question?
 - e. Were there any words in this question that you did not know?
 - f. Could we do anything, change the item in any way, to make it clearer to you?
12. **Passage probes for ELA (and perhaps science):**
 - a. Did you think this passage was easy or hard to read?
 - b. Were there any words you did not understand?
 - c. Was any part of it confusing to you?
 - d. Could you find the answers to the question in the passage?
13. As you conclude, don't forget to thank the student for their help and insight.
14. **Take a few moments to review your notes to make sure you've accurately recorded important observations regarding how students performed on the task.**



Appendix D: High Quality Assessment Review Tool

Part 1: Assessment Profile
<p>Items Submitted – check all that is submitted and <u>fully</u> completed:</p> <p><input type="checkbox"/> NH PACE Performance Task Template</p> <p><input type="checkbox"/> Teacher Instructions: materials needed, time required for administration, procedure</p> <p><input type="checkbox"/> Student Performance Task: what the student is required to do and produce (prompt, directions, materials, checklists, etc.)?</p> <p><input type="checkbox"/> Scoring Rubric</p> <p><input type="checkbox"/> Answer Key or Guidelines: <u>Please circle if Not Applicable</u></p> <p><input type="checkbox"/> Actual Texts or links to texts, videos, data charts, etc. (provides materials)</p>
<p>Performance Task Description:</p> <p><input type="checkbox"/> Fully describes the context, the anticipated activities, products and/or presentations, resources, texts, and materials needed, and what students are expected to demonstrate.</p> <p><input type="checkbox"/> Partially describes the context, the anticipated activities, products and/or presentations, resources, texts, and materials needed, and what students are expected to demonstrate.</p> <p><input type="checkbox"/> Minimally describes the context, the anticipated activities, products and/or presentations, resources, texts, and materials needed, and what students are expected to demonstrate.</p>
<p>Teacher Directions:</p> <p><input type="checkbox"/> Fully describes all aspects of the administration of the task including pre-requisite learning, lessons for scaffolding, what the students will do independently. These directions follow the guidance outlined in the document entitled “Guidelines for Independent Student Work Products for NH PACE Assessments: Implications for instructional scaffolding.”</p> <p><input type="checkbox"/> Partially describes the aspects of the administration of the task including pre-requisite learning, lessons for scaffolding, what the students will do independently. These directions partially follow the guidance outlined in the document entitled “Guidelines for Independent Student Work Products for NH PACE Assessments: Implications for instructional scaffolding.”</p> <p><input type="checkbox"/> Minimally describes aspects of the administration of the task including pre-requisite learning, lessons for scaffolding, what the students will do independently. These directions minimally follow the guidance outlined in the document entitled “Guidelines for Independent Student Work Products for NH PACE Assessments: Implications for instructional scaffolding.”</p>
<p>To what extent is scaffolding provided?</p> <p><input type="checkbox"/> No scaffolding is provided for aspects of the task that are being scored with the rubric</p> <p><input type="checkbox"/> Low level of scaffolding is provided for aspects of the task that are being scored with the rubric</p> <p><input type="checkbox"/> Some scaffolding is provided for aspects of the task that are being scored with the rubric</p> <p><input type="checkbox"/> High level of scaffolding (teaching, modeling, think-alouds, conferences, and/or organizers) is provided for aspects of the task that are being scored with the rubric</p>



Student Instructions:

- Fully** describes all student expectations.
- Partially** describes student expectations.
- Minimally** describes student expectations.

Comments:

A high quality teacher-created assessment should be ... Aligned

Part 2: Alignment

The standards evaluated by the assessment are identified and are aligned to the expectations of the task:

- Yes**
- Partial/Unclear**
- No**

The standards and objectives are appropriate for the intended grade level for which the assessment is being used?

- Yes**
- Partial/Unclear**
- No**

The skills and knowledge assessed are grade level appropriate:

- Yes**
- Partial/Unclear**
- No**

To what extent do you see a content match between the prompt on the task and the corresponding Standards?

- Full match** – all aspects of the task or items fully address or exceed the relevant skills and knowledge described in the corresponding standard(s)
- Close match** – most aspects of the task or items address the relevant skills and knowledge described in the corresponding state standard(s)
- Partial match** – Some aspects of the task or items address or partially address the skills and knowledge described in the corresponding state standard(s)
- Minimal match** – Few aspects of the task or items match some relevant skills and knowledge described in the corresponding state standard(s)
- No match** – No aspects of the task or items are related to the skills and knowledge described in the corresponding state standard(s)



Identify the Depth-of-Knowledge range of the Standards measured by the assessment (see Webb’s DOK charts):

- DOK 1:** recall and reproduction
- DOK 2:** skills and concepts
- DOK 3:** strategic thinking/reasoning; requires deeper cognitive processing.
- DOK 4:** extended thinking; requires higher-order thinking including complex reasoning, planning, and developing of concepts.

Are the set of items or tasks reviewed as cognitively challenging as the standards? In other words, the student performance task elicits sufficient evidence for judging the level of student understanding related to the competencies and standards identified. Use the definitions below to select your rating:

- More rigor** – most items or the tasks reviewed are at a higher DOK level than the range indicated for the state standard(s)
- Similar rigor** – most items or the task reviewed are similar to the DOK range indicated for the state standard(s)
- Less rigor** – most items or the task reviewed are lower than the DOK range indicated for the state standard(s)

Comments/Suggestions for Improving Alignment (if any)

Relevant evidence to justify ratings:

A high quality assessment should be ... Scored using Clear Guidelines and Criteria

Part 3: Rubric

Is the rubric are aligned to the assessment task?

- Fully aligned**
- Partially aligned**
- Not aligned**

Are the score categories clearly defined and coherent across performance levels?

- Yes**
- Partial**
- No**

Is it clear which aspects of the task this rubric will be used to evaluate?

- Yes**
- Partial/Unclear**
- No**



Based on your review of the rubric would the scoring rubric most likely lead different raters to arrive at the same score for a given response?

- Yes
- Partial/Unclear
- No

A high quality performance assessment should be...Fair and Unbiased

Part 4: Fair and Unbiased

(the areas below should be discussed relative to the needs of ELLs, gifted and talented students, and students with disabilities)

To what extent are the tasks visually clear and uncluttered (e.g., appropriate white space and/or lines for student responses, graphics and/or illustrations are clear and support the test content, the font size seems appropriate for the students)?

- Formatting is visually clear and uncluttered
- Formatting is somewhat confusing or distracting
- Formatting is unclear, cluttered, and inappropriate for students

Are the directions and the task presented in as straightforward a way as possible for a range of learners?

- Yes
- Partial/Unclear
- No

Is the vocabulary and context(s) presented by the task free from cultural or other unintended bias?

- Yes
- Partial/Unclear
- No

Comments/Suggestions for Improvement for Fair and Unbiased (if any)

Relevant evidence to justify ratings:

A high quality performance assessment includes appropriate reading and visual materials

Part 5: Appropriateness of Text/Visual Resources

The texts and visual resources support the topic and prompt:

- Yes
- Partial/Unclear
- No
- N/A



The texts have characteristics of a:

- Simple Text**
- Somewhat Complex Texts**
- Complex Texts**
- Very Complex Texts**
- N/A**

Note: Refer to the *Text Complexity Rubric for Literary Texts or Informational Texts*

The amount of texts and visual resources are:

- Appropriate for the grade level and the time allotted for the task**
- Appropriate for the grade level, but may exceed the time allotted for the task**
- Burdensome for the grade level and the time allotted for the task**
- No texts and/or resources are included**
- N/A**

Comments/Suggestions for Improvement for Fair and Unbiased (if any)

Relevant evidence to justify ratings:

Recommendation for this assessment:

- No changes needed**
- Minor changes recommended**
- Some changes required, please address and resubmit**
- Substantial changes needed, please address and resubmit**
- Task rejected—new task needed**

Discussion:



Appendix E: Summary of 2016 PACE Common Task Review

English Language Arts (ELA)

Literary writing tasks for English Language Arts (ELA) were developed for grades 4, 5, 6, 7, 9, and 10. These six tasks were reviewed in July 2016 by the Center for Assessment and all necessary revisions were completed based on the review. These tasks received final approval by the NH DOE in October 2016 and are currently considered to be operational tasks for the 2016-17 school year. The review process included an examination of the NH PACE Task Development Template, student instructions, teacher instructions, as well as any other ancillary documents, such as graphic organizers. This review was conducted using the NH PACE High-Quality Assessment Review Tool as a guide. This process ensures that the expectations and directions in all submitted documents are coherent and cohesive, as well as demonstrate alignment, appropriate depth-of-knowledge, and fairness. The literary writing rubrics that accompany these PACE Common Tasks were developed by the Center for Assessment.

The focus of these tasks is for students to demonstrate their ability to write in response to prompt about a text. The tasks did not expect students to demonstrate the ability to read on-grade level texts or to demonstrate comprehension independently. For example, in some grades, especially the elementary grades, teachers read the text aloud to students and discussed the literary elements, such as theme and the characters. Students were expected to use information from these discussions along with evidence from the text to construct a text-dependent essay. Consequently, the complexity of the texts selected varied for each grade level.

All six tasks required some revision, with grades 5 and 9 requiring minor changes due to incomplete information provided. All other grades required some changes primarily due to inconsistencies across documents. For example, the completed NH PACE Task Development Template specified that the texts should be read aloud, while the student directions required students to read independently. Grade 6 was the only task that provided excessive scaffolding for writing. The inconsistencies described above were corrected and resubmitted for review by the Center for Assessment. The tasks were approved and submitted to the NH DOE for final approval.

Math

Tasks in grades 3, 5, 6, 7, Geometry and Algebra were reviewed by the Center for Assessment in August 2016. All revisions made on the basis of the feedback for the math tasks were completed and submitted to the NH DOE for final approval by early October 2016. The comprehensive task review included a careful evaluation of the NH PACE Task Development Template, the teacher instructions, the student instructions, the scoring rubric, and the answer keys. The key components of quality reviewed include: coherence across the documents and resources, clarity, alignment, depth of knowledge, and fairness.



Of the six tasks reviewed, five of the tasks required only minor changes, while one of the tasks necessitated more substantial revision. The most common minor modifications included: simplifying the language and layouts of the performance tasks to improve clarity and reduce construct-irrelevant cognitive load for students, adding notes to the teacher directions regarding administration conditions to improve standardization, and adding or updating the language of the rubrics to further distinguish among the points and/or dimensions to improve inter-rater reliability. The Grade 7 task required more than just minor revision upon its formal review from the Center for Assessment. Based on the feedback provided, the task was modified to remove inconsistencies between the rubric and the task, remove unnecessary scaffolding of the task, and streamline and improve the authenticity of the task so the students are not repeatedly asked to low-level questions of identifying the constant of proportionality but are instead required to apply their knowledge of constants of proportionality to create and interpret graphs and equations representing plant growth.

In order to prevent the need for these more substantial revisions in the future, the development and review process for the 2017-2018 operational tasks has been improved. First, content leads are now responsible for leading task development, tracking progress, and organizing the pilot. These content leads are in close communication with the PACE leadership team and the Center for Assessment to ensure they have all the latest and relevant information they need to share with the participating teachers and access to assessment expertise when needed. Additionally, the Center for Assessment will begin reviewing the 2017-2018 tasks earlier in the development process in the spring of 2017 to ensure that task development can be re-directed if necessary. The final tasks will then be reviewed again before final approval as they have in the past.

Science

Tasks in grades 4, 8, Life Science, Physical Science and Chemistry were reviewed by the Center for Assessment in September 2016. The minor revisions required for the science tasks were completed and submitted to the NH DOE for final approval by October 2016. The comprehensive task review included a careful evaluation of the NH PACE Task Development Template, the teacher instructions, the student instructions, the scoring rubric, and the answer keys. The key components of quality reviewed included coherence across the documents and resources, clarity, alignment, depth of knowledge, and fairness.

In general, all seven of the tasks required minor changes. The most common minor modifications included: adding notes to the teacher directions regarding administration conditions to improve standardization and adding or updating the language of the rubrics to further distinguish among the points and/or dimensions to improve inter-rater reliability. However, almost all tasks revealed a struggle between the group work common to extended science laboratories and the individual work necessary for documenting each student's level of competency. In almost all cases, the science teachers handled this quite well, but comments were provided in several cases to help the task development teams that might have been struggling with this tension. Finally, in their efforts



to assess the three-dimensional structure of the Next Generation Science Standards, task development teams occasionally specified certain standards that they intended the task to assess, but such content (usually) dimensions were not often found in the rubric.

As noted for mathematics above, the science task development process now includes having content leads remain in close communication with the PACE leadership team and the Center for Assessment to ensure they have all the latest and relevant information they need to share with the participating teachers and access to assessment expertise when needed. Additionally, the Center for Assessment will begin reviewing the 2017-2018 tasks earlier in the development process in the spring of 2017 to ensure that task development can be re-directed if necessary. The final tasks will then be reviewed again before final approval as they have in the past.



Appendix F: Scaffolding Brief

[The following brief was written for teachers to support high-quality PACE Common Task implementation].

Delineating Instructional and PACE Performance Tasks

Instructional Tasks: Part of the theory of action of the NH PACE pilot is that through improvements in teaching practices, in part by focusing on deeper and more authentic instructional experiences, student outcomes will improve. Often times this modification in teaching will occur through developing local performance tasks to use as instructional aids in the classroom. Throughout these performance tasks, teachers are typically instructing students on key concepts and guiding students on how to apply their new knowledge and skills. In other words, in the midst of these performance tasks, teachers are “scaffolding” student learning. In the field of education, the term **scaffolding** refers to a process in which a teacher models or demonstrates the key skills necessary for understanding and applying concepts, and then steps back to offer support as needed. Psychologist and instructional designer Jerome Bruner introduced this term in the 1960s using the theory that when students are given the support they need early in the process of learning something new, they stand a better chance of using that material independently. In other words scaffolding student learning is a “temporary” structure that builds student capacity towards independence.

Example of Instructional Task:

There are many different types and reasons for providing scaffolding structures for students during instruction. Consider the following instructional task for seventh-grade students:

After reading The Circuit by Francisco Jimenez, you will write a literary analysis that answers the following prompt: What is the theme of The Circuit? Use evidence from the text to explain how story elements work together to reveal the theme.

- *Using the outline provided, decide on the main story elements you will write about in your analysis and what you want to say about them. Using your notes, decide and sort your text evidence in to appropriate paragraphs. Plan to share during conferencing prior to writing your essay.*
- *Write an opening paragraph that includes an objective summary of the literature and a thesis about the theme of the story. Be prepared to share and receive “warm” and “cool” feedback from the class.*
- *Write an initial draft complete with introduction, body, and conclusion; insert and cite textual evidence. Conference with your partner and using the rubric receive feedback for improvement. Consider which areas are strong, what needs to be added, moved, changed or deleted. Plan for your next steps. Be prepared to review during a teacher-student conference.*
- *Revise your rough draft according to the feedback from your peers.*
- *Reread your draft and revise for spelling, capitalization, punctuation, and grammar. Adjust the formatting as needed to provide clear, appealing text.*

This instructional task expects students to be able to write a literary essay; however, there are several supports in it which include: 1) an outline, 2) conferencing with the teacher, 3) breaking a complex task into smaller “doable” steps, and 4) peer-editing. We consider all of these scaffolds as appropriate during instruction to provide supports for students as they are learning the process of writing a literary analysis.



PACE Common Tasks: The PACE Common Tasks are intended to emulate the design of local classroom tasks, but are also intended to provide evidence of student performance on key competencies. Because one purpose of the PACE Common Tasks is to assess student performance against key competencies, and contribute to the annual determination of student “proficiency,” student performance on the PACE Common Tasks must be reflective of what students can do **independently**, without instructional strategies and scaffolding.

Example PACE Common Task:

Below is an example of one way that the instructional task above may look different as a PACE Common Task.

This performance task assesses your knowledge, skills and abilities related of the following New Hampshire ELA & Literacy Competency:

Students will demonstrate the ability to effectively write informative texts to examine and convey complex ideas for variety of purposes and audiences.

After reading The Color Purple by Alice Walker, write a literary analysis that answers the following prompt: How does the setting of the book support the author’s purpose? Use evidence from the text to explain how elements of time and place are used together to reinforce the author’s purpose.

Suggestions:

- *Select a pre-writing strategy (e.g., graphic organizer, outline) to get started.*
- *Begin your first draft by the start of next week.*
- *Before turning in your final product, check for the quality of your evidence, the structure of your argument, grammar, and spelling.*

The implication here is that students are able to write their literary analysis independently, transferring the process knowledge to a different text and different prompt. The scaffolds that were appropriate during instruction, such as peer-reviewing and conferencing about an outline are no longer appropriate. Because the student will be scored on the quality of the final draft, it is expected that the teacher has not provided any support or guidance that would affect the quality of the final draft.

Drawing the line between where good teaching ends and valid assessment begins can be tricky, which is why we have attempted to outline some assumptions and general principles in this brief.

Assumptions of PACE Performance Task Scores

Before administering a PACE Common Task, there are a number of instructional activities that must have already taken place within the classroom. Because the PACE Common Task is a summative assessment, the results and score interpretations carry with them the following assumptions:

- 1) Students have had an opportunity to learn, practice and master the tested content. The PACE Common Task does not introduce knowledge or skills with which students are unfamiliar.
- 2) The work produced and submitted for the PACE Common Task is a result of the students’ efforts and knowledge alone. Student performance is representative of what the student could be expected to reproduce independently.



Considerations When Making Scaffolding Decisions

Scored Work: It is important to consider which criteria are going to be evaluated and scored on the PACE Common Task rubric. This guiding principle requires seriously considering the expectations of the standards and the competencies. For example, the task may expect students to complete an outline or graphic organizer, write a rough draft, revise based on a self-assessment, and complete a final draft. Although the only part of this PACE Common Task that is being scored is the final draft, providing feedback to a student about the content, details, evidence, computations, selected strategy, analysis of data, etc. from the organizers and drafts greatly influences the criteria used to evaluate the quality of the final piece of student work. Consequently, the assessment will not accurately reflect what the student is able to do without feedback and support.

Developmental Appropriateness: Considering the grade level of students and what is expected in the standards will help to guide acceptable scaffolding on the PACE assessments. The Common Core State Standard for Production and Distribution in grades 3-8 states: *With some guidance and support from peers and adults, develop and strengthen writing as needed by planning, revising, editing, rewriting, or trying a new approach, focusing on how well purpose and audience have been addressed.* We suggest that students in these grades can be provided with organizers and engage in a more scripted process for writing or problem solving. Because we want to set students up for success, asking students to demonstrate that they have worked through the process (completing an organizer, writing a rough draft, etc.) before moving to the next phase, is also acceptable. However, to reiterate, commenting on the quality of the information will inevitably impact the final product, and therefore change the inferences we can make about what students can do independently. Additionally, it is more appropriate for students in higher grade levels (e.g., 9 and 10) to be asked to “select” an organizer or strategy independently rather than providing one for them.

Students with Disabilities and English Language Learners: All students need to be given the opportunity to demonstrate their understanding of concepts. For students with disabilities or other special considerations such as English Language Learners, individualized educational plans should specify what types of accommodations are appropriate. These accommodations are legally mandated in order to provide a more level playing field for students with disabilities or non-English speaking students. Accommodations are commonly categorized in five ways: presentation, response, setting, timing, scheduling, and linguistics. Though accommodations come in a variety of forms, what they all have in common is that *they do not alter what is being measured.* Students are expected to demonstrate the same understandings, even if, for example, the response mode or timing has been modified.

Final Thoughts

Scaffolding describes teaching strategies geared to support learning when students are introduced to and learning new subject matter. It gives students context and a foundation in which to understand new information and how to integrate it with prior learning. Scaffolding techniques, such as the examples above, are considered fundamental to high-quality teaching for all students. In order for learning to progress, scaffolds should be gradually removed as instruction continues, so that students will eventually be able to demonstrate proficiency without these supports.



Appendix G: NH PACE 2016-2017 Data Collection Protocols

#1: Assessment Map
Due November 1, 2016

Process:

- Provide an assessment map that can serve as a key to all of the **summative** assessments for grades 3-11¹⁵ (Math and ELA) and grades 4, 8-10 (Science) that will factor into the competency scores.
- All of the state standards should be mapped to at least one competency. Karen Matso will use the competencies listed to create score frameworks in the Learning Management System.
- The summative assessments for each competency should be labeled by type and mapped by time of administration. Anything included in the assessment map may be subject to a state audit to ensure assessments are aligned to intended standards and are high quality.

Assessment Map: Example for Grade 3 Math

Competenc	Standards	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
1. Algebraic Thinking	CC.3.OA.1	Short Summa- tive		PBA	Unit Test						
	CC.3.OA.2										
	CC.3.OA.3										
	CC.3.OA.4										
	CC.3.OA.5										
	CC.3.OA.6										
	CC.3.OA.7										
	CC.3.OA.8										
	CC.3.OA.9										
2. Number Operations	CC.3.NBT.1		Short Summa- tive		Unit Test	Short Summa- tive	Unit Test	PBA	Short Summ- a-tive	PBA	
	CC.3.NBT.2										
	CC.3.NBT.3										
3. Fractions and Proportiona I Reasoning	CC.3.NF.1			Short Summa- tive		Short Summa- tive	Unit Test	PBA			
	CC.3.NF.2										
	CC.3.NF.2a										
	CC.3.NF.2b										
	CC.3.NF.3										
	CC.3.NF.3a										
	CC.3.NF.3b										
	CC.3.NF.3c										
	CC.3.NF.3d										
4. Data	CC.3.MD.3				Short Summa- tive					Short Summa- tive	Unit Test
	CC.3.MD.4										
5. Geometry and Measurement	CC.3.MD.1				Short Summa- tive			Short Summa- tive	PACE Common Task		
	CC.3.MD.2										
	CC.3.MD.5										
	CC.3.MD.6										
	CC.3.MD.7										
	CC.3.MD.7a										
	CC.3.MD.7b										
	CC.3.MD.7c										
	CC.3.MD.7d										
	CC.3.MD.8										
	CC.3.G.1										
	CC.3.G.2										

¹⁵ For grade 11, only submit the course map for the ELA course and Math course in which the majority/plurality of eleventh grade students are enrolled.



#2: Local Assessment Quality Review - State
Due January 16, 2017

To monitor the quality of the local assessments, the NH DOE will now be conducting a quality assurance audit in which a sample of local assessments are reviewed. Formative feedback from the reviews will be provided to districts. Additionally, if systematic problems in the quality of assessments are detected for any district, additional state support for improving the quality of local assessments will be offered.

Process:

- Select one major summative assessment from each competency for each of the 16 PACE grade/subject combinations.
- Label each assessment with the grade level, subject area, and name/number of competency in such a way as to easily correspond with the assessment map provided on or before November 1st.

Submission:

- Upon collecting the sample of assessments for each course, please email copies of those assessment packages to Mariane Gfroerer. If you would prefer to submit these assessments by mail, please email Mariane to make arrangements.

#3: Performance Task Feedback Review - SCALE
Due January 16, 2017

To provide feedback on locally developed performance assessments that are designed using the PACE template, the NH DOE has contracted with the Stanford Center for Assessment, Learning, and Equity (SCALE) to provide feedback reviews to districts.

Process:

- Submit all locally developed performance assessments that are designed using the PACE template for feedback from SCALE.

Submission:

- Email copies of the PACE templates and supplementary materials to Mariane Gfroerer.
- The contract with SCALE does not end on the 16th, as more local tasks are developed with the PACE template, please continue to submit these assessments in an on-going fashion.



#4: PACE Common Task Student Work Samples for Cross-District Calibration
Mail to: Center for Assessment, Attn: PACE Scanning, 31 Mount Vernon, Dover, NH 03820
Due May 26, 2017

The student work samples will be used in the PACE Summer Institute to provide evidence of comparability in the evaluation of student work across districts.

Process:

- Select eighteen (18)¹⁶ final student work samples for each PACE Common Task (no names, drafts, comments, or scored rubrics). This sample should span all score points and should be representative of the distribution of achievement in the district.
- Student ID#s (SASIDs) should be written on the top of each student work sample.

Submission:

- **Please attach a cover page to the top of each grouping of PACE Common Tasks so we know the district, grade, subject area, and number of student work samples submitted.**
- Mail to the Center for Assessment anytime during the 2016-2017 school year prior to May 26, 2017.

#5: Body of Work Samples
Mail to: Center for Assessment, Attn: PACE Scanning, 31 Mount Vernon, Dover, NH 03820
Due May 26, 2017

The main purpose of collecting student work samples throughout the year is to help document and evaluate student performance through the year along with the PACE Common Tasks. This collection will help support standard setting (cut scores) activities and cross-district comparability activities during the PACE Summer Institute.

Process:

- Districts are asked to submit 5-7 samples of student work for a minimum of nine (9) students from each subject area and grade level specified in the table below. The nine students should be selected to represent a range of achievement. For example, three generally low-performing students, three high-performing students, and three students who perform at about an average level. Student work of the same 9 students should be used throughout the year so districts may want to select one or two additional students in case a student moves.

¹⁶ For districts with fewer than 18 students in a given grade, the district should submit all available papers.



Subject Area	Grade Levels
Math	Grades 3, 6, and Algebra
Science	Grades 4, 8, and High School Life Science
ELA	Grades 5, 7, and 10

- The student work samples should come from major summative assessments throughout the year (e.g., unit tests, and performance based assessments) and demonstrate student achievement across the breadth and depth of the course content. The samples will be used to provide evidence of student achievement relative to the achievement level descriptors (see the content area ALDs).
- The PACE Common Task can serve as one of the assessments submitted for each student. It is critical that enough of the context of the assessment is included so that an outside teacher would know that a student was responding to a particular problem, prompt, exercise, reading, etc. Therefore, including the student instructions and specific questions asked along with student responses is critical. **Please remove students' names, as well as any comments, grades, rubrics, and score marks prior to submission. Label each student work sample with the student's ID# (SASID) on the top right-hand side of the page.**

Resources:

- Short instructional video on the administrative libguide.
- PACE Body of Work Explanation & Examples are provided on the administrative libguide.
- Content area ALDs on the administrative libguide.

Submission:

- **Please attach a cover page to the top of each grouping of Body of Work samples so we know the district, grade, subject area, and number of student Body of Work samples submitted.**
- Mail to the Center for Assessment anytime during the 2016-2017 school year prior to May 26, 2017.

#6: PACE Common Task Scores
 Upload into the Learning Management System
Due June 16, 2017

This is a critical step for documenting that the scores that students receive are NOT contingent upon the district where the student goes to school. In other words, this step is designed to evaluate the extent to which teachers evaluate student work the same way (comparable) across districts. The PACE Common Task Scores will be reconciled with the consensus scores that are generated from the PACE Summer Institute to ensure the evaluation of student work is comparable across districts.



Process:

- Within district calibration sessions are highly encouraged to maximize the consistency and validity of scores.
- Upload PACE Common Task scores by rubric dimension into the Learning Management System for all students administered a PACE Common Task.

Resources:

- Recommended protocols for identifying anchor papers and individual teacher scoring are provided on the administrative libguide.

Submission:

- Score data (by rubric dimension) for each student who completed a PACE Common Task uploaded into the Learning Management System.
- Indicate if accommodations were used for the student.
- Indicate if the student has an IEP that modifies the instructed content standards to off grade level.

#7: Teacher Judgment Survey
Upload into the Learning Management System
Due June 16, 2017

All teachers in grades 3-11 (Math and ELA) and grades 4, 8-10 (Science) should complete a Teacher Judgment Survey for their students in the Learning Management System. The results of the Teacher Judgment Surveys will be one variable used to produce each student's "annual determination" of proficiency in ELA, math, and science in grades/subjects where the PACE Common Task is administered. The Teacher Judgment Survey asks teachers to classify their students based on PACE Achievement Level Descriptors (ALDs) for a given grade/subject. ALDs articulate the expected levels of performance related to the knowledge and skills described by the grade-level content standards.

Resources:

- Teacher Judgment Survey Instructions on the administrative libguide
- Content area ALDs on the administrative libguide

#8: Full Set of Student Competency Scores
Upload into the Learning Management System
Due June 16, 2017

In order to produce annual determinations based on multiple sources of evidence, we need to be able to collect consistent and accurate information for each student. These data will be used along with the data collected from the Teacher Judgment Surveys to produce annual determinations of student proficiency.



Process:

- All teachers in PACE districts should be keeping records of students’ progress on each of the course competencies.
- The competency scores that are submitted should be reflective of summative student achievement on each competency by the end of the year.
- The competency score scale (e.g., 1.00-4.00, 0-100) is district determined, but should be consistent within each grade level and content areas in each district. Work with teachers to ensure scores are not submitted that are out-of-range (e.g., 0.75 on a 1.00-4.00 scale).

Submission:

- Please ensure that all students in grades 3-11 (Math and ELA) and grades 4, 8-10 (Science) have scores entered into the Learning Management System for their work related to each competency.
 - For grade 11, only submit the competency scores for the ELA course and Math course in which the majority/plurality of eleventh grade students are enrolled.

#9: Competency Score Data
Due June 16, 2017

Per recommendations from our Technical Advisory Committee (TAC), we ask that those districts that have electronic grade books email competency score data to the Susan Lyons at the Center for Assessment (slyons@ncea.org) to support generalizability analyses this coming summer. By competency data, we are looking for all of the individual scores that go into the end of year competency (e.g., summative tests, quizzes, projects, performance tasks), see Appendix B for an example data sheet. We do not need the assignments themselves, but rather we are asking for the student score data, and information regarding which competency/competencies each score informs. As we have done with the student work sampling procedures, we are only asking for samples of this data according to the following table:

Grade	Subject Area
3	Math
4	Science
5	ELA
6	Math
7	ELA
8	Science



#10: Within-District Double Scoring of the PACE Common Tasks
Due June 16, 2017

Within-district double scoring is a critical step for documenting the quality of scoring for the PACE Common Tasks. As a result, we need every teacher administering a PACE Common Task to submit at least 3-4 student work samples for double scoring with a minimum of 20 student work samples double scored per PACE Common Task within each district. For smaller districts, this may mean that every PACE Common Task student work sample in elementary grades is double scored.

There are two potential options for conducting the inter-rater reliability analyses:

1. The “embedded” approach does not require a stand-alone step, but is embedded in individual scoring..
2. The second option would require a stand-alone event for approximately ½ day.

Option #1 (embedded):

- Each teacher submits 3-4 student work samples, depending upon the total number of teachers at the grade level, from a range of performance levels.
- These student work samples are embedded in the scoring packets of the other teachers either at their grade level or grade span such that each teacher will end up double scoring approximately 3-5 extra student work samples.
- Teachers score these embedded student work samples along with their regular student work and record the scores.

Option #2 (stand-alone):

- Each teacher submits 3-4 student work samples, depending upon the total number of teachers at the grade level, from a range of performance levels. For districts with multiple schools, the district leader can determine whether or not to do this within each school or across schools at the district level.
- These student work samples are distributed to a grade level or grade span cohort of teachers such that each paper is scored by at least one other teacher. As an example, if there are 4 teachers at a given grade/subject level and each teacher submits 3 student work samples, there would be a total pool of 12 student work samples to score among second readers. Since each of the 12 student work samples needs two scores, that means that there are 24 scored responses needed for each grade/subject. This means that each of the 4 teachers will have to score 6 other teachers’ student work samples.

Resources:

- Short instructional video on the administrative libguide.
- PACE Double Scoring Collection Spreadsheet (Excel file) on the administrative libguide.

Submission:

- Using the PACE Double Scoring Collection Spreadsheet, enter your district’s double scores for all courses with a PACE Common task. Leave the columns for the extra score dimensions blank for the tasks with rubrics that have fewer dimensions than the



spreadsheet allows. Save the file as: District_PACE Double Scoring_1617.xlsx and email to slyons@nciae.org

#11: Participants List for PACE Summer institute
Due June 16, 2017

The PACE Summer Institute is an event of critical importance to the success of the project. The goals of the last two days of the PACE Summer Institute involve operational outcomes such as the scoring and rating of student work samples, but also include professional capacity building objectives for educators involved in the PACE pilot.

Process:

- Each district is asked to send 32 teachers (2 per PACE Common Task for Consensus Scoring and 3-4 per Body of Work grade/subject area for Body of Work Rating) to the last two days of the PACE Summer Institute for calibration and body of work scoring purposes. Administrators or other educators may sub for teachers if necessary. We need participants to commit to attend as our randomized scoring procedures rely on pairing teachers together in specific ways.

Resources

- District Participant's Excel Sheet on the administrative libguide.

Submission:

- Prior to June 16, 2017 please send the completed participant's excel sheet to Mariane and Susan.

#12: Data Collection Requirements Checklist
First submission: January 16th
Second Submission: May 26th
FINAL SUBMISSION: June 16th

In an effort to improve organization and communication about district progress on meeting the data collection requirements presented in this document, we have provided a Data Collection Requirements Checklist to be completed by districts, signed, and **submitted three times this year**.

Resources:

- The template for the Data Collection Requirements Checklist is provided in Appendix C and also on the administrative libguide.

Submission:

- At each of the due dates, please fill the checklist for what has been submitted thus far and sign. Please scan and email to Mariane Gfroerer and Susan Lyons.



Data Collection Requirements Checklist

District: _____ Lead: _____ Lead Signature: _____ Date: _____

		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
Gr	Subject Area	Assessment Map Emailed (Y/N)	Local Assessments for Quality Review- State Emailed (#)	Performance Tasks for Feedback Review-SCALE Emailed (Y/N)	Common Task Work Samples Mailed (x/18)	Body of Work Samples Mailed (x/9)	Common Task Scores Uploaded (Y/N)	Teacher Judgment Surveys Completed (Y/N)	Full Set of Student Competency Scores Uploaded (Y/N)	Competency Score Data Emailed ¹⁷ (Y/N)	Within-District Double Scoring Emailed (Y/N)	Participants for Summer Institute Emailed (x/2)
		Nov 1, 2016	Jan 16, 2017	Jan 16, 2017	May 26, 2017	May 26, 2017	June 16, 2017	June 16, 2017	June 16, 2017	June 16, 2017	June 16, 2017	June 16, 2017
3	MATH				/18	/9						/2
4	ELA				/18							/2
4	SCI				/18	/9						/2
5	ELA				/18	/9						/2
5	MATH				/18							/2
6	ELA				/18							/2
6	MATH				/18	/9						/2
7	ELA				/18	/9						/2
7	MATH				/18							/2
8	SCI				/18	/9						/2
9	ELA				/18							/2
10	ELA				/18	/9						/2
HS	Algebra				/18	/9						/2
HS	Geometry				/18							/2
HS	Life Sci				/18	/9						/2
HS	Phys Sci				/18							/2

¹⁷ For districts with electronic grade books only



Achievement Level 3- PACE
Compose full compositions with grade-appropriate techniques, transitions, structure, organization, details, concluding statement, audience, purpose, and text features for narrative, informational, and opinion writing using the elements of the writing process and publishing with technology.
Conduct short research projects to answer a question or investigate a topic or concept and locate information from data, print, or non-print resources; select and use sufficient accurate text evidence for research and writing.
Use of grade-appropriate conventions of standard English grammar, usage, capitalization, punctuation and spelling when writing in all genres; errors may occur, but overall meaning is clear.

Achievement Level 3- SBAC	
Writing Targets 1-10	<p>The student who just enters Level 3 should be able to:</p> <ul style="list-style-type: none"> • Write or revise one paragraph, demonstrating narrative techniques, chronology, appropriate transitional strategies for coherence, or author’s craft appropriate to purpose. • Write full compositions, demonstrating narrative techniques: chronology, transitional strategies for coherence, or author’s craft with minimal demonstration of purpose. • Write or revise one or more informational/explanatory paragraphs, demonstrating ability to organize ideas by stating focus, including transitional strategies for coherence, supporting details, or a conclusion. • Use text features in information texts to enhance meaning without support. • Write or revise one or more paragraphs, demonstrating ability to state an opinion about a topic or source, set a context, organize ideas using linking words, develop supporting reasons, or provide an appropriate conclusion. • Write full opinion pieces, demonstrating ability to state opinions about topics or sources, attend to purpose and audience, organize ideas by stating a context and focus, include structures and transitional strategies for coherence, develop supporting reasons, and provide a conclusion. • Without support, use grade-level vocabulary appropriate to the purpose and audience when revising and composing text. • Apply or edit grade-appropriate grammar, usage, and mechanics to clarify a message and edit narrative, informational, and opinion texts. • Without support, use tools of technology to produce texts.
Listening Target 4	<p>The student who just enters Level 3 should be able to:</p> <ul style="list-style-type: none"> • Interpret and use information delivered orally or audio-visually without support.
Research Targets 1, 2, and 4	<p>The student who just enters Level 3 should be able to:</p> <ul style="list-style-type: none"> • Conduct short, limited research projects to answer a question or to investigate a topic or concept. • Locate information to support central ideas and key details; select information from data or print and non-print text sources without support. • Generate opinions with evidence to support the opinion based on prior knowledge and information collected.



