2017 No. 017

Formative Evaluation of New Hampshire's Performance Assessment of Competency Education (PACE)

Summary Report

Prepared for: New Hampshire Department of Education
101 Pleasant Street
Concord, NH 03301-3860

Prepared under: Center for Innovation in Education
1648 McGrathiana Parkway, Suite 350
Lexington, KY 40511-0350
Contract Number UK-1652-16
(HumRRO No. S16-30)

Authors: Arthur Thacker
D. E. (Sunny) Becker

Date: March 10, 2017

Formative Evaluation of New Hampshire's Performance Assessment of Competency Education (PACE)

Summary Report

Table of Contents

Table of Contents (Continued)

List of Tables

List of Figures

Formative Evaluation of New Hampshire's Performance Assessment of Competency Education (PACE)
Summary Report

Executive Summary

New Hampshire's Performance Assessment of Competency Education (PACE) is an assessment and accountability strategy designed to reduce the amount of, and reliance on, standardized testing by supplanting much of the traditional end-of-year summative testing with teacher developed performance assessment tasks. PACE was created to support deeper learning through competency education, and to be more integrated into students' day-to-day work than current standardized tests. The PACE pilot program represents a fundamental qualitative shift in the way accountability assessments are developed, administered, and used to promote teaching and learning.

In spring 2015, the U.S. Department of Education granted New Hampshire (NH) a waiver from specific requirements of the No Child Left Behind Act and then the requirements of the Every Student Succeeds Act (ESSA) as part of a demonstration pilot program.[1] Participating NH districts administer Smarter Balanced assessments in grade 3 English Language Arts (ELA), grade 4 Mathematics, and grade 8 ELA and math, as well as the SAT to all grade 11 students. In addition to local performance assessment tasks (hereafter local tasks), a common performance assessment task (hereafter common task) is administered in each grade and subject (ELA, math, and science) without a state assessment.

The Human Resources Research Organization (HumRRO) was awarded a contract to conduct a formative evaluation of the PACE system in the Tier 1 districts between April 2016 and February 2017. The primary goal for this evaluation is to ensure that the PACE Leadership team has useful information to make decisions that advance the program's goals. This summary provides a brief description of data collection activities associated with the evaluation, an overview of the evaluation results, key findings, and recommendations. A full description of the research supporting this summary is available as a series of four technical reports.

A Brief Introduction to PACE

The PACE system relies upon locally developed, locally administered performance assessment tasks aligned with local district grade and course competencies. These local competencies and local performance assessments are aligned to the State Model Competencies, which, in turn, are aligned with national standards in each content area.

New Hampshire school districts must apply and demonstrate readiness and commitment before being allowed to participate in the PACE system. Districts enter via a three-tiered system, based on how fully they meet the requirements to implement PACE. Tier 1 districts have fully implemented PACE. Tier 2 districts implement competency-based education, but have not fully implemented PACE. Tier 3 districts are at a beginning stage. There are currently nine Tier 1 districts and they were the focus of the evaluation. Four districts joined PACE in 2014–15: Epping School District School Administrative Unit (SAU) 14, Rochester School District (SAU 54), Sanborn Regional District (SAU 17), and Souhegan School District (SAU 39). A second wave of districts became PACE Tier 1 districts in 2015–16: Concord School District

[1] ED granted NH DOE a waiver extension on October 6, 2016.

(SAU 8), Monroe School District (SAU 77), Pittsfield School District (SAU 51), and Seacoast Charter School (SAU 46). In addition, White Mountains (SAU 35) joined as a Tier 1 district in the 2016–17 school year. SAU 35 was included in limited fall/winter 2016 evaluation activities.

Because PACE replaces the Smarter Balanced assessments for several grade/subjects, the requirements for participation are rigorous. Districts must commit to administering a common task for every assessed grade/subject each year, plus they must agree to administer several local tasks. Students' scores on these tasks contribute to local student competency scores and feed into annual determinations. The tasks can often take several class periods to administer and a sample of papers must be double scored. Ensuring that the quality of all assessment stages, including developing, field testing, revising, administering, and scoring the performance tasks is sufficiently high requires a great deal of teacher professional development and a large time commitment for all participants.

Each common task undergoes a one-year pilot testing phase, with evidence-based revisions made after each round of pilot testing (the number of rounds determined based on the performance of the task), followed by an operational year. Administration of a pilot common task may occur in a subset of districts, but during the operational year, all Tier 1 districts administer the common task at the specified grade level. The common tasks must be administered in a standardized manner during the operational year to achieve comparability. After the pilot and operational years, these common tasks are available in a growing bank of tasks from which teachers can select to use as local tasks. Teachers may make modifications to the tasks at this time, including administering the task at a different grade level (Changing grade level would be primarily done in middle school science, where the curriculum is not consistent by grade across districts.).

The PACE common tasks and local tasks are intended to be closely linked to classroom instruction. All the tasks, local and common, are teacher-designed to assess the specific competency targeted by lessons within the curriculum. The tasks are not administered in a specific testing window, but instead come at the time during the year when it is most appropriate in the curriculum. Teachers know the content of the tasks well before administering them and the tasks are designed to test students' competency regarding specifically taught content topics. There is no guessing what the tasks will cover in a given year. PACE tasks are complex and require deep understanding of the content. There are no multiple choice-questions on PACE tasks. Students write and revise, perform real-world applications of mathematics, or conduct science experiments to demonstrate their competencies. And, while PACE likely requires more testing time than Smarter Balanced, because it is so integrated into the curriculum, students often do not realize they are taking a test. Instead, they consider the PACE tasks to be another part of their daily classwork[2].

Framing the Evaluation

HumRRO was tasked with three evaluation goals:

- **Evaluation Goal 1:** Refine and validate the PACE Accountability program's theory of change/theory of action
- **Evaluation Goal 2:** Provide formative feedback loops on key success criteria

[2] For a complete overview of PACE, see http://www.education.nh.gov/assessment-systems/pace.htm.

- **Evaluation Goal 3:** Capture the "Story" of PACE

Goal 2 included nine success criteria:

- **Success Criterion 1:** Gaining clear commitment from local leadership
- **Success Criterion 2:** Building cross-district leadership and cross-district collaboration
- **Success Criterion 3:** Developing high-quality performance assessments
- **Success Criterion 4:** Successfully implementing common performance assessments
- **Success Criterion 5:** Providing training and calibration
- **Success Criterion 6:** Reaching successful rates of inter-rater agreement
- **Success Criterion 7:** Producing "comparable" annual determinations
- **Success Criterion 8:** "No harm" on the Smarter Balanced Assessments
- **Success Criterion 9:** Ensuring equity

HumRRO provided interim reports, as well as informal feedback, organized around the goals and these success criteria to quickly provide ongoing feedback to the New Hampshire leadership during the course of the evaluation. The goals and criteria also served as major areas of inquiry for the final evaluation report.

At the onset of the evaluation, the theory of action/change was captured by three bullet points. They included:

- "If we believe that all students must be college- and career-ready . . .
- then, our system must advance students as they demonstrate mastery of knowledge, skills, and work study practices, . . .
- which requires a comprehensive system of educator and school supports."

The bullet points are compelling, but do not lead directly to claims that can be investigated in a traditional validity argument. Our first task was to capture the goals of PACE and to map the success criteria onto a framework that could be used to organize and structure evidence collected regarding PACE's quality and validity. That framework is presented in Figure 1 as the theory of action/change for PACE.

Figure 1 includes four interim goals and a set of underlying claims that must be substantiated to attain each interim goal. Lack of support for any one interim goal may undermine subsequent goals. For example, if the tasks are not administered as intended (i.e., Interim Goal 3), then the validity of the scores is called into question (i.e., Interim Goal 4), regardless of how high inter-rater agreement and inter-rater reliability are among the scorers. While the interim goals are not entirely linear and dependent on each other (as they might be in a stricter interpretive argument for validation of an interpretation of assessment scores), this framework illustrates potential threats to the intended outcomes of the program. It also provides a common way of understanding how any

potential threat within one of the interim goals might interact with others. The final evaluation report describes the various data collection activities and summarizes the evidence for each goal and its underlying claims and assumptions, thereby creating a validity argument for the PACE pilot program. In addition, the final report summarizes the successes PACE has achieved at this stage of implementation, concerns or issues that should be addressed, and conclusions and recommendations. Data collection was designed to include both qualitative and quantitative information from multiple stakeholders to triangulate and bolster the accuracy of the findings. Data collection methods included

- **observations at major PACE events (e.g., task development and scoring sessions);**
- **classroom observations;**
- **interviews with students, parents, teachers, principals, and district leaders;**
- **surveys of teachers; and**
- **analyses of score data.**

Collecting data from multiple stakeholders using multiple methods bolstered the accuracy of the inferences about PACE. It allowed us to capture the perceptions of the majority of PACE participants, and it allowed us to hear important minority opinions. Perhaps, most importantly, it allowed us to differentiate between the two.

NH Pace Chart: rev: 02.01.17

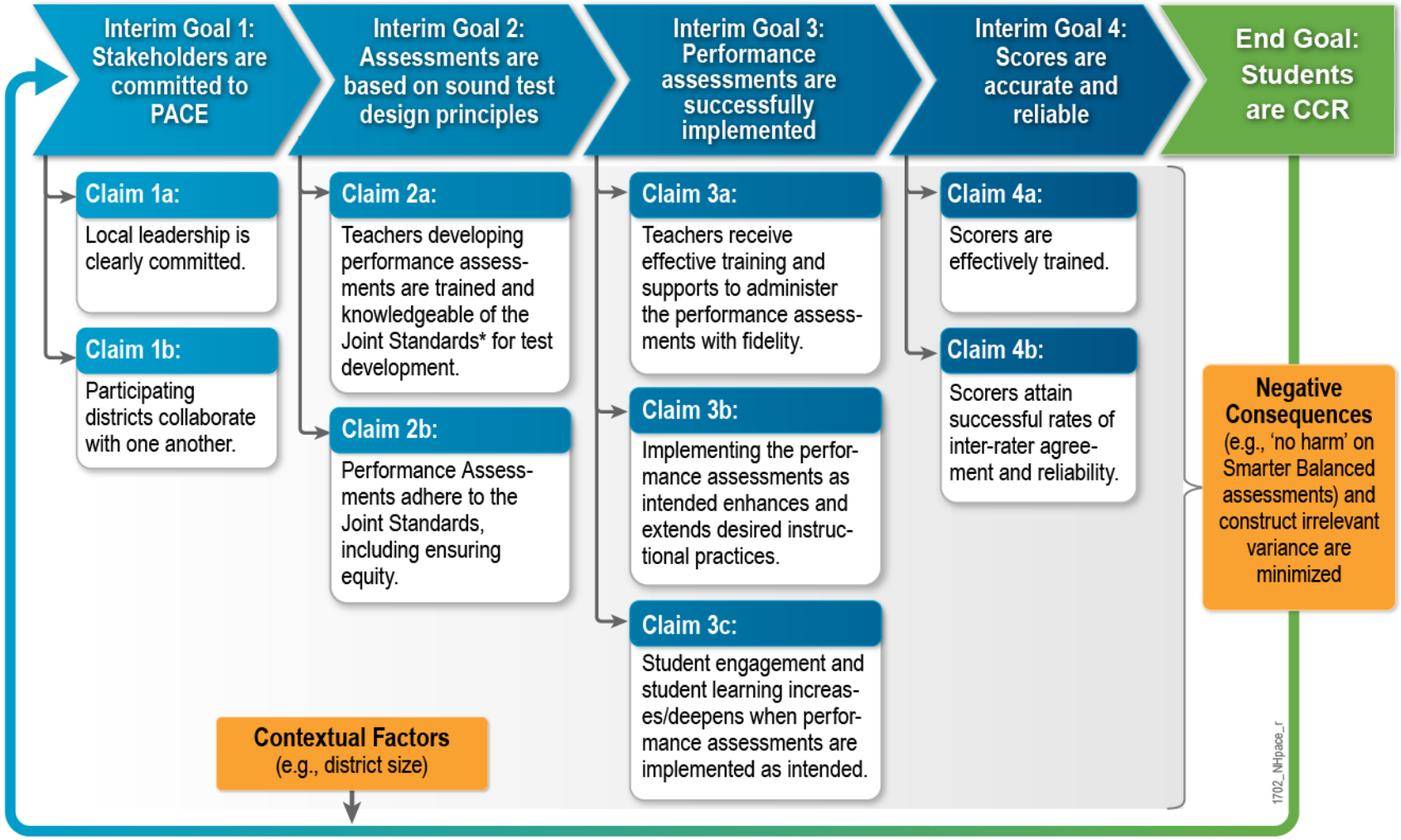

Figure 1. PACE theory of action/change.

** We understand that the PACE stakeholders are not test design experts and, therefore, that the AERA, APA, & NCME Standards are not firsthand knowledge for this audience. Consequently, our discussion with these stakeholders referred more generally to "high-quality assessment."*

Evaluation Activities

HumRRO conducted several data collection activities over the course of the evaluation, from April 2016 through February 2017. 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.

Evaluation

Interim Goal 1. Stakeholders are committed to PACE

We found strong evidence supporting Interim Goal 1. PACE participants overwhelmingly indicated that local leadership was highly supportive of the PACE initiative. There are several methods by which districts collaborate with one another, and participants report that collaboration is a major benefit of PACE membership. New collaboration mechanisms have recently been put in place to account for PACE expansion, but these have not yet been evaluated for effectiveness.

Claim 1a. Local leadership is clearly committed

The first testable claim from the Theory of Action is “local leadership is clearly committed” to PACE. For this claim, we gathered data from PACE District Leads, school administrators, and teachers. We include teachers because they are truly the most influential local leaders in the program. They develop the tasks and decide what is to be assessed. They also take the tasks back to their schools where they influence other staff members. The overwhelming majority of PACE participants reported high levels of commitment.

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 they even design the scoring rubrics. 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. Educators at all levels described ownership of the system as a major contributor to buy-in.

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. As such, PACE provides an alternative to 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. PACE proponents describe the tasks as authentic and important. They often describe the benefits of PACE in terms of better preparing students for life beyond school. It is relatively easy to buy in to a program if you believe its methods and outcomes are better than what came before.

Claim 1b. Participating districts collaborate with one another

The second testable claim from the Theory of Action is "participating districts collaborate with one another." This claim is also supported in a number of ways. 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. These cross-district meetings require that personnel from different schools work together to accomplish common goals. The meeting participants then implement the things they learn in their classrooms and share what they have learned with other educators within their school/district. A theme that emerged across the data collection activities is that teachers value and enjoy the opportunity for cross-district collaboration. They often refer to it as beneficial for their professional growth. They also describe it as useful for developing high quality common tasks and for calibrating the scoring of student work.

In-person meetings are just the beginning. The second way educators across districts interact is 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. They also share guidance on the administration of the common tasks. Some commonly used documents include a guide for educational scaffolding, student-friendly rubrics, and principles of scoring student work.

Collaboration across districts is also accomplished by emailing the PACE coordinators and leadership. PACE teachers ask direct questions, some of which are answered individually, and some of which become group emails to eliminate potential common misunderstandings or misconceptions. If questions become common or concerning, they are addressed during in-person meetings and with guidance on the LibGuides.

Prior to the start of the evaluation, each district identified a PACE Lead to coordinate activities in the district and to communicate with PACE Leadership. Participation in monthly PACE Leads meetings is one venue for collaboration.

Over the course of the evaluation period, PACE implemented four new collaboration measures. The first was to name an overall curriculum coordinator to assist with PACE task development activities. This step was taken to (a) improve communication, (b) ensure common understanding of goals, instructions, and deadlines, (c) provide a master schedule earlier in the year, and (d) provide an additional resource for PACE participants.

Another new collaboration mechanism was the 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. Most have been PACE participants and task developers since the beginning of the PACE pilot program. The Content Leads program allows PACE to build deep expertise among local educators without requiring all educators to attend every meeting and activity. The Content Leads helped PACE address the expansion of the program. They act as liaisons to the educators in their districts and also in a "buddy district," which might not have a Content Lead. This allows PACE to field smaller groups of teachers when a very large group would be unwieldy, such as during task revision workshops. Wordsmithing a common task can benefit from multiple voices, but there is a point of diminishing returns when too many group members provide input. The Content Leads help keep these kinds of in-person interactions small. This approach has the added benefit of reducing time that some teachers must spend outside the classroom in collaborative activities. The Content Leads take the information from workshops and other activities back to the districts. Educators who are not Content Leads can still provide information, including suggested revisions to common tasks and rubrics, via the LibGuide. In the districts without Content Leads, Teacher Representatives were identified to coordinate among local teachers.

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. The Content Lead provides an opportunity for all participants to contribute to all aspects of PACE, in addition to the local tasks that all teachers develop. Buddy districts, as well as the other 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.

Finally, as of February 2017, PACE Leadership began inviting Tier 2 districts to attend monthly PACE Leads meetings. Observing these meetings will afford an opportunity to become more familiar with the PACE Tier 1 experience.

Interim Goal 2. Assessments are based on sound test design principles

We found strong evidence to support Interim Goal 2. The task developers (teachers) are well trained and thoughtful in the development of the tasks and scoring rubrics. They adhere to the central themes and major principles of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), even if they do not specifically reference them.

Claim 2a. Teachers developing performance tasks are trained and knowledgeable of the Joint Standards[3] for test development

In order for assessments to achieve their goals, providing valid and reliable inferences from students' scores, they must adhere to sound design principles. To build a sound assessment, we must understand the inferences that will be made from the scores, or put more simply, the ways we would like to use the assessment results. PACE scores serve three main purposes: (a) they provide progress checks on student learning at various points in the school year;

[3] We understand that the PACE stakeholders are not test design experts and, therefore, that the AERA, APA, & NCME Standards are not firsthand knowledge for this audience. Consequently, our discussion with these stakeholders referred more generally to "high-quality assessment."

(b) they inform teachers about students' competency related to specific knowledge, skills, and abilities; and (c) they are aggregated to provide an indication of school/district performance. Teachers routinely design assessments to check progress on the content they teach, and they did so prior to the PACE program. PACE adds the competency aspect, though many schools had implemented some form of competency education previously, placing the focus of the assessment on competency rather than progress or performance relative to peers. PACE is also used for accountability, as aggregate scores on PACE tasks replace Smarter Balanced scores for gauging school and district level performance. These new uses for the assessments require that the developers, in this case teachers, think through aspects of testing that they have not historically needed to consider. To do so, they must be effectively trained.

PACE teachers demonstrated high levels of assessment literacy during training sessions, scoring, and standards setting meetings. A large number of teachers who were trained and began the process of developing the common tasks are now highly knowledgeable about task design and provide a strong core of leadership for the program. While they do not typically reference the Joint Standards, they routinely discuss complex aspects of assessment design. They recognize and struggle with the dilemma between standardization and instructional and curricular freedom. They routinely discuss scaffolding as a method to give students access to assessment content, especially scaffolding for SWD. They recognize that scaffolding can represent both a benefit and a potential danger to the validity of student scores. They strive to ensure that the rubrics they design are well documented and can be consistently and accurately scored. When they design rubrics, they are careful to include only components for which students will provide appropriate evidence during the performance tasks. They routinely discuss the pros and cons of aspects of test design. For example, a common discussion point is whether to expand the content of a draft task. Broad tasks can address more content, but more discrete tasks may be better indicators of specific competencies.

The training model employed by PACE has allowed teachers to learn by doing, albeit with some assistance provided by assessment experts and personnel from the state department of education. They gain assessment literacy by encountering and dealing with assessment issues as they design, pilot, revise, administer, and score the tasks. When they have questions, they have expert help available, but they hold most of the decision-making power. This learn-doing model has been very effective for the teachers who started with PACE, but it takes a lot of time now to bring new participants up to speed. The experienced teachers have obtained considerable important knowledge, and orienting newcomers can be challenging. This will be an ongoing concern as PACE expands to more districts.

Entailed in the theory of action for PACE is that teachers apply what they learn from developing high quality *common* tasks to the development of high quality *local* tasks. A survey of PACE teachers reveals the majority of teachers report that they have been able to apply what they have learned from their experiences developing *common* tasks to developing higher quality *local* performance tasks. This lends further support for claim 2a.

Claim 2b. Performance assessments must adhere to the Joint Standards, including ensuring equity

PACE teachers do not routinely reference the Joint Standards. They focus on solid assessment design and the tools they have been given. The common tasks used across all districts are highly scrutinized for potential biases or sensitive content. They undergo extensive editing and revision before and after they are field tested. Teachers follow detailed guidelines for administration, including guidelines for ensuring standardization and for promoting accessibility.

They follow guidelines for providing accommodations. Students' work is double scored and scorers' consistency is verified. The common task is centrally rescored and used to adjust for any systematic scoring differences on local tasks by district. Well-established standard setting methods are used to classify students by performance categories. PACE results are compared with an external reference assessment (Smarter Balanced). These efforts largely parallel the processes of large-scale testing companies that adhere to the Joint Standards and they contribute to a high quality assessment system.

The local tasks do not undergo the same levels of scrutiny as the common tasks. There were substantial differences in the quality and depth of the local tasks, by district, discovered during standard setting. According to PACE teachers, there were several factors that might account for the differences. Teachers learned late in the 2015-16 school year[4] that they were expected to keep documentation for a sample of students of local tasks conducted throughout the year. This may have caused some teachers to provide less than optimal student work samples. Teachers also have potentially differing levels of experience and expertise with competency based education, depending on when their district joined and how involved they were in PACE. Improvements in communication and efforts to extend participation are expected to result in improved and more consistent local tasks. PACE teachers have indicated that they believe the common tasks and the local tasks are authentic measures of their students' achievement.

Ensuring equity for all students is a challenge for any assessment system. When asked during interviews and focus groups whether PACE tasks were more or less accessible compared to Smarter Balanced, most teachers indicated that PACE tasks are more accessible. They described the embedded nature of PACE and the availability of the same accommodations students routinely received for classroom work as justification. Because PACE is embedded in instruction, students often do not realize that the PACE tasks are different from their regular class work. This helps make the tasks more accessible than an "assessment event." A few teachers, however, expressed concern that they might inadvertently impact the standardization, and consequently the measured construct, of the tasks (both common and local tasks) by providing too much accommodation or too much scaffolding. This is a common challenge for standardized tests and the concern led to the creation of an accommodations guide and a scaffolding guide to help teachers make informed and sound decisions about accommodations and scaffolding. Some teachers also expressed concern on the PACE teachers survey about the accessibility of PACE tasks. Contextual information gathered from teachers' open-ended comments on the survey, and also from focus groups, indicates that some teachers believe the reading and writing demands of PACE tasks are quite high and could limit accessibility for some students.

Interim Goal 3. Performance Assessments Are Successfully Implemented

We found strong evidence to support Interim Goal 3 for the common tasks. For the system overall, we found considerable evidence that the training and support are adequate and that PACE has had a substantial positive impact on both teaching practice and student learning. There was insufficient evidence to fully evaluate the quality and implementation of the local tasks at this time.

[4] This notification was accelerated in the 2016-17 school year so teachers were notified of this expectation from the outset.

Claim 3a. Teachers receive effective training and supports to administer the performance assessments with fidelity

Most teachers report that their training is adequate for administering the PACE tasks. Most teachers report that their school's administration provides them with the resources and supports they need to effectively implement the common tasks. And most report that they received effective training to effectively implement common tasks.

Because the tasks are developed by teachers, their familiarity with the assessment is better than it could be for a less familiar testing event. Teachers who do not participate in collaborative task development sessions have access to the task materials on the LibGuide and can consult with their Lead Teachers or Content Leads. Teacher support for administering the PACE assessment includes the online LibGuides, where tasks, rubrics, the Implementation Guidelines Manual, and other materials reside.

Teachers also support each other within schools and districts and outside their districts as well. They meet both formally and informally, in person and through shared working documents. They are also supported by their school administration, PACE District Leads, PACE leadership, and several expert consultants.

It is important to examine both the common tasks and the local tasks when considering the fidelity with which tasks are implemented. The common tasks are collaboratively developed and have a suite of task-specific student-friendly instructions and more generic supports (e.g., administration guide). Differences in implementation during the pilot phase of each task are reviewed and the task materials are revised to clarify, as needed.

Local tasks are exactly that, and could be used by only a single teacher. There is little worry that the local tasks are not administered as intended, but they may not be as fully developed or as in-depth as the common tasks. They are almost certainly not as well documented. This does not mean they are not effective performance assessment tasks. During interviews, teachers reported that their locally developed tasks have improved with every year of implementation. Starting with the 2015-16 school year, PACE will be auditing one local task per competency for every course in each Tier 1 district. These will be used to document local task quality and to provide feedback to teachers. When we think about implementation fidelity, we must consider the information provided by the local tasks. Scores from the local tasks are combined with scores from the common tasks to determine students' overall score and achievement level. So, training and support must be sufficient to allow teachers to create their own local tasks and administer them in a manner that supports the validity of the inferences made from annual test results. During the course of the evaluation we observed several local tasks being administered and it was clear that most PACE teachers understood how to design local tasks to support those inferences and how to assess their students in a way that elicits reliable performance data.

Claim 3b. Implementing the performance assessments as intended enhances and extends desired instructional practices

Teachers across districts expressed that implementing performance tasks has had a positive impact on their instruction. They commonly mentioned that PACE has had a positive impact on increasing the depth of knowledge (DOK) 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. Preparing students for the PACE assessments requires high DOK

lessons and opportunities for students to apply and extend the content they've learned independently.

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. Impact on instruction for most assessments would fall under the heading of unintended negative consequences. PACE leadership is not overly concerned about teachers "teaching to the test." PACE, ideally, supports "testing to what is taught." While most accountability assessments drive instruction to at least some extent, their influence on instruction would not be viewed as positive by many educators. The high stakes and comprehensive nature of end-of-year tests may cause teachers to superficially teach many topics, spend days or weeks reviewing previously taught content, and spend instructional time on testing strategies.

PACE represents a significant step toward true integration of curriculum, instruction, and assessment. Several teachers reported that participation in PACE led to a qualitative shift in the way they approached assessment in their everyday work. Where assessment was previously used to differentiate superficial levels of performance and tests were a combination of items with varying degrees of difficulty and obscurity, they now focus on providing evidence that students have or have not achieved competency within a given content topic. This leads to more effective critical thinking about best practice, remediation and extension activities, and more productive reflection on and revision of day-to-day lessons.

Claim 3c. Student engagement and student learning increases/deepens when performance assessments are implemented as intended

Much like the shift in focus for teachers, PACE also represents a shift for students. Typical assessments are primarily focused on estimating achievement. 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.

Teachers report higher engagement for their students and deeper learning of the content, during PACE assessments and as a result of improvements in their instructional practice that they attribute to participating in PACE. The majority of students report that they would rather take a PACE assessment than an end-of-year comprehensive test like Smarter Balanced or the New England Comprehensive Assessment Program (NECAP) test. When the students who indicated they would rather take a mostly multiple-choice assessment were asked why, they typically responded that they liked having some chance of getting the answer correct, even if they did not know the content very well. They commented on guessing and using test-taking strategies that were of no help on the PACE tasks. Others indicated that they preferred not to write their answers, as that was more difficult for them than a more traditional multiple-choice test. The students endorsing PACE discussed how closely the tasks were linked to their curriculum and how strong a measure of their abilities the tasks were.

Interim Goal 4. Scores are Accurate and Reliable

We found considerable evidence that students' scores and annual determinations are accurate and reliable. Scorers were effectively trained and PACE tasks were double scored. The common task was used to calibrate among the districts and to evaluate scorer accuracy.

Claim 4a. Scorers are effectively trained

PACE tasks, local and common, are scored using teacher-developed rubrics. These rubrics describe student work at four levels of competency. The teachers strive to make the distinctions as clear and concrete as possible. Adjectival scales (e.g., poor, acceptable, good, very good) are not acceptable. When teachers discuss the rubrics during the development process, they focus on the distinctions between the score levels and how to judge when students' work would represent one level versus the other.

After field testing of the common tasks, teachers come together to discuss and revise the tasks and scoring rubrics. During this process, the teachers score students' work. If there are inconsistencies or if the rubric is too vague to categorize students reliably, they adjust the rubrics. They also discuss the effectiveness and accuracy of the rubrics. For example, if the teachers agree that a students' work is exemplary, but some idiosyncrasy of the rubric forces them to give a lower-than-deserved score, they can adjust the rubric to deal with the issue. Once the rubric has been finalized, all the districts can score the task consistently.

During scoring, scorers begin with calibration sessions. These occur within districts and allow scorers to come to a common understanding of the application of the rubrics. They select and use anchor papers, or papers with known scores, to help calibrate and as a reference during the scoring process. While many teachers reported that the scoring process was time consuming, they were confident in their ability to score accurately and consistently.

The majority of teachers reported on the survey that the scoring rubrics for the common tasks are sufficiently clear and detailed to ensure that separate scorers scoring the same student work arrive at the same score, and that the scoring resources available on the LibGuide effectively explain how to score student work on the common tasks. Local tasks do not receive as much scrutiny. Training for developing and scoring rubrics for local tasks comes largely from teachers' prior experiences and from their work with the common tasks. They build in good scoring practice for the local tasks based on their experience and training.

Electronic score files are sent for generalizability analyses to the National Center for the Improvement of Educational Assessments (Center for Assessment). The Center for Assessment conducts cross-district comparability analyses and uses a standards setting procedure to establish district-level cut scores. Individual task scores are not adjusted during this process. The cut scores impact the overall determinations, and consequently the proportion of students in each of the achievement levels (1-4). Results provided by the Center for Assessment indicate that the overall scoring consistency is quite high and that few adjustments are necessary to the initially set cut scores due to inconsistent scoring (either too lenient or too strict) within the districts, indicating effective training for the scoring of PACE tasks. This process ensures consistency of scoring across districts. It is also the way that scores are made comparable across years.

Claim 4b. Scorers attain successful rates of interrater agreement and reliability

A sample of student responses to the common tasks are drawn for consensus scoring. Scorers work with a partner to rescore several students' work. Scorers may not be from the same district as the students whose work they score. Subsequent to the consensus scoring meeting, the scores from the central scoring group are compared with the scores from the district. If there is poor agreement between the district results compared to the consensus scored results, the scores on the common tasks are adjusted to account for the discrepancy. If the differences between adjusted common task performance is substantially different from local task performance, it may also signal a district level scoring bias. If such a difference is discovered, scorers can be retrained on a district by district basis.

Within-district inter-rater reliability is computed by the Center for Assessment. They determine whether a teacher scores more leniently or strictly by comparing the teachers' scores on the common task to the consensus scores on that task. The index they use for this purpose is a "deviance" index, which describes how far from the consensus scored papers an individual teacher scores (averaged across students). Several flags for potentially inconsistent scoring have been established, but scoring for 2015–16 was quite consistent. While there were minor differences between subjects and by district, scoring for PACE common tasks by teachers was largely verified as accurate and consistent during consensus scoring.

The Center for Assessment also computes within-district rater agreement statistics (e.g. % exact agreement, % adjacent agreement) and Cohen's Kappa statistics for a sample of the double-scored common tasks (Evans & Lyons, 2016). Pairs of raters had exact agreement rates of between approximately 60 and 85%. There were substantial differences by grade, subject, dimension, and by district, but nearly all districts achieved greater than 60% exact agreement rates across all grade subjects. Kappa statistics indicate moderate to substantial agreement of ratings across all grades and subjects as well.

Samples of local tasks are also double scored. Teachers examine the results, but formal reliability statistics are not monitored during active scoring. Students' scores on the local tasks represent their work over the course of the year. They might be compared with more typical end-of-unit test scores. Unlike typical end-of-unit tests, students receive rubrics along with their PACE task instructions. This allows them to self-monitor as they work. If, at the end of the task the teacher score is different from the students' expectations, they can discuss the differences with the teacher. This provides the teacher with a quality check on the rubric and gives the student an opportunity to understand how to interpret and use the rubric to achieve the score they desire. Also, parents noted that the rubrics provide information to facilitate a discussion with their children about their performance on the task. The rubrics provide clear expectations for the students who use them, which improves the validity of the scores. The feedback teachers receive from the double scoring and from their interactions with students helps improve their locally developed tasks and rubrics to achieve better reliability.

Contextual Factors

While there are several contextual factors influencing the quality of PACE implementation worth mentioning, the largest stems from implementing PACE at the district level. Interviews with teachers and administrators in multiple districts yielded several district-specific positive and negative experiences with PACE. 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 bank of

mostly suitable locally developed tasks and were familiar with the design of competency-based rubrics. Their students had largely become accustomed to the kinds of tasks PACE requires. Districts that joined later had to build the infrastructure necessary to implement PACE.

District size plays an important role in PACE implementation as well. There are distinct advantages and disadvantages associated with being in a larger or smaller district, and it is not clear which is better. 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. In addition, the requirement to submit copies of sample student work can be challenging because the smaller districts have very few support staff.

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.

District size can also influence teacher buy-in. Some of the smaller districts are close knit teams of educators, all of whom are supportive of both PACE and one another. In some districts, there are a few educators who are resistant to the implementation of PACE. This can cause strife for those who are committed to implementing the program with fidelity. It can be especially challenging when teachers bring PACE training or information back to their schools. These resistant teachers can have a larger impact in a small district, but larger districts are more likely to encounter them.

District size also impacts teacher expertise in PACE tasks. In larger districts, a subset of teachers participates in task development, necessitating that they keep their colleagues informed of the rich discussions from which they benefited. In small districts, a lone teacher could conceivably participate in task development for ELA, mathematics, and science—requiring substantial time out of the classroom.

In addition to district size, there are other contextual factors that may influence the implementation of PACE. Previous experience with competency based education and development of performance tasks made the transition to PACE easier for some teachers. Most PACE districts had previously developed Quality Performance Assessments (QPAs) that are similar to the PACE tasks. Those QPAs are often the basis for the local tasks for PACE, and participation in the QPA Institutes is an expectation for TIER 2 PACE member districts.

Another important contextual factor is the perceived value of participation in PACE leadership roles versus the time requirements, especially time out of class. At least one district decided not to have any teachers serve as Content Leads. This district was heavily involved in developing the early common tasks. Many of the teachers from this district drafted the initial text for common tasks. The district decided to pull back from their leadership role to preserve the teachers' time in the classroom. Some teachers in the district were pleased with the decision, while others were disappointed that their role in PACE had been reduced.

Some parents, teachers, and students commented on the way that PACE tasks are scored. PACE tasks typically use "conjunctive" scoring, where multiple components of a task are scored separately, and the lowest of those scores becomes the final score. While conjunctive scoring is not a PACE-wide policy, common tasks are scored this way and teachers emulate the common tasks when they develop the local tasks. There was concern among multiple stakeholders that this method of scoring could result in lower than expected task-level scores. It was also not clear at what level the stakeholders were describing when discussing conjunctive scoring. Tasks might assess two or three dimensions within a subject area (e.g., mathematical modeling, computational accuracy, communication). The scoring rubric might contain several specific bullets describing students' performance under each dimension, each bullet scored 1-4. It is possible that conjunctive scoring could be done within a dimension, where the lowest bullet score determined the dimension score. This is the way the common tasks are scored. The dimension scores might be aggregated in some other way (e.g., by averaging). It is also possible that the lowest dimension-level score could be used to determine the overall task score. Parent focus groups in two districts referenced the scoring method as one reason that scores were lower than expected for otherwise high performing students, but they were not sufficiently specific to allow for fine distinctions regarding the scoring mechanism. Parents' concerns regarding task-level scoring was emphasized, in part, because a single task may be a large part of a student's 6-weeks grade (tasks are used for in-class grading as well as for the annual determination), and because the student might only have a few task scores to contribute to the annual determination.

Scoring at the dimension level using conjunctive rules should result in better scorer consistency. If the scorers examine the evidence for each bullet, which can be very specific in terms of what the student is expected to produce, and then take the lowest scored bullet as the dimension score, we would expect a high degree of comparability at the dimension level (scorers only rate discrete well-defined evidence). Scorers might vary on other bullets, but if they agree on the lowest bullet, the overall task score would be the same. If the scorers use a compensatory approach, they must contend with defining "good enough" across the bullets.

Negative Consequences Minimized

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. We have discussed some of the ways that PACE has succeeded in reducing the negative consequences that already existed in New Hampshire schools, but it is also important to recognize potential negative consequences of PACE and to guard against them.

PACE tasks, especially science and English language arts tasks, can take a long time to implement. PACE tasks are designed to measure big, but reasonably discrete, ideas from the content standards. The developers must constantly ask themselves if the time investment to implement the performance assessments generates sufficient information to justify that time. Some of the science tasks can take more than a week's worth of classes to complete. Some of the English language arts tasks, because they may require that students read an entire novel in class, can take even longer. PACE task developers must guard against the tasks becoming so long that they unintentionally narrow the curriculum.

The PACE common task, in most ways, counts no more than any local task. It is used as an instrument to ensure scoring accuracy and reliability and as an equating tool to guard against

cross-district incomparability. These additional uses, however, can cause teachers to give it much more attention than the local tasks. This added attention can be positive or negative. For example, one high school English language arts common task required students to respond to a text. One school chose a Shakespearean play as their text. The school then chose that same play as the winter drama production and staged the play with student actors for all the school prior to the administration of the task. While this is certainly not a prohibited activity, it may have given the school and its students an advantage over other schools that were not so savvy. Emphasizing the common task may limit the available time for other content. It may also create unintended differences between the way that the common versus local tasks are treated, which could, in turn, make attributions about scoring quality or other aspects of the PACE assessments less certain. Some teachers described spending a month or more in preparation for the common task (often including reading a novel aloud in class). If the task promotes strong lessons and broad and deep learning of the content, this level of effort may be entirely justified. If, however, the task represents a relatively discrete aspect of the overall curriculum, that time may be better spent. Interviews with PACE teachers revealed that most were very positive about the tasks and considered the preparation of students for them a major benefit. A small minority, however, indicated that the work on the part of teachers and students was disproportional to the benefit the students received.

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 Also, a few outlier responses obtained during interviews and focus groups suggested going back to Smarter Balanced. One goal of PACE is to generate enough tasks that development can become a more reasonable ongoing process of replenishing or revising only a few new tasks per year. Until that goal is reached, there is the potential for over-burdening teachers.

Once teachers develop units of study and associated performance tasks, they tend to use them for several years. The nature of PACE promotes this practice and, because of the complex nature of the tasks, we are not overly concerned with test security. There may be concerns, however, when the common task addresses the same, or closely related, content. Some teachers described having to abandon very strong units of study and local tasks because they were required to use the common tasks for that content. Using both would be time prohibitive and redundant, so they used the common task only. In one example, a teacher typically taught a life sciences unit on oceanography. The unit took advantage of the teacher's major area of study from college and was a highly-developed set of interconnected lessons for the students. However, because not all PACE districts have easy access to an ocean, the same content from her oceanography unit was tied to rivers and streams in the common task. She will likely teach the oceanography unit next year, when the common task changes, but she was disappointed that she had to replace it this year with a task she did not feel did as good a job of teaching the related content.

New Hampshire does not currently have a grade-by-grade curriculum for middle school science, but the common science tasks are grade specific. There is, therefore, some concern among educators that the tasks do not always match their curriculum. If, for example, one district teaches life sciences in grade 8, while another teaches physical sciences in grade 8, a common task in grade 8 related to life sciences could potentially disadvantage the latter district. The science tasks for middle school have been designed to address science and engineering principles and cross-cutting concepts, but these do not come content free. This issue may be

resolved in one of two ways, based on current curriculum plans in New Hampshire. The New Hampshire Board of Education recently adopted the Next Generation Science Standards (NGSS). It is possible that this will lead to the adoption of a more consistent curriculum (at least by major content topic by year) in middle schools. PACE is also planning to allow districts to use matrixed/crossed designs with the common tasks. Alternatively, once a sufficient number of common tasks are developed, they can be administered based on the content of the course, rather than the grade of the student. So an eighth-grader taking physics might complete a physics task, while a sixth-grader taking a physics course might complete the same task. An eighth-grade in a different district might take a life sciences task. This would allow district-level control of curriculum, but may introduce new challenges for maintaining district-to-district comparability.

We conducted focus groups with a small number of parents in each Tier 1 district. While most of these parents were very supportive of PACE, there were a few who questioned the reliance of the program on performance tasks. This was especially true if the school or district adopted alternative reporting methods (e.g., changed report cards from traditional ABCDF grading to the 1-4 ratings for the PACE tasks). A few parents were concerned that colleges might not understand how the PACE tasks were scored and might inadvertently penalize their child because the grading system was so different from a traditional system.

Data Analyses

Comparison of Aggregate Data

Because PACE results are used in place of Smarter Balanced scores, it is important to consider the validity of PACE as an overall indicator of students' achievement in ELA, mathematics, and science at a specific grade level. This is the primary use of Smarter Balanced mathematics and ELA scores and we would expect PACE to provide similar results. We would not expect the results to be interchangeable. All of the differences in the design, purpose, development, administration, and scoring described earlier are expected to make PACE unique from Smarter Balanced. If the final results were the same, it would call into question if PACE truly represented a major shift in instruction and assessment.

New Hampshire does not require students to take both the Smarter Balanced and PACE assessments during the same year, so we can't directly compare assessment results for individual students. We can, however, compare the PACE results in aggregate to the Smarter Balanced results for the state. We can also compare the results from 2015 to those from 2016. Figure 2 presents this information for mathematics.

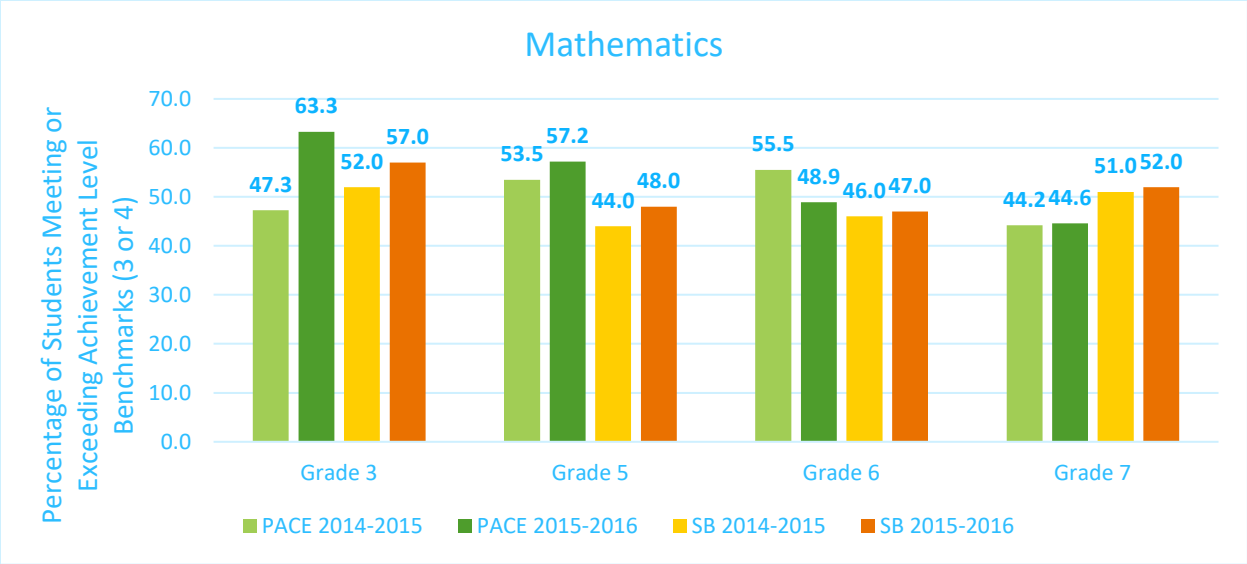

Figure 2. Comparisons of PACE and Smarter Balanced Mathematics Results for 2015 and 2016.[5]

Figure 2 shows us that the PACE results tended to be somewhat higher than Smarter Balanced for grades 3, 5, and 6, but somewhat lower for grade 7 (except for grade 3, for which PACE scores were lower in 2015 and higher in 2016). If we look across years, we see that the Smarter Balanced results improved from 2015 to 2016 in all grades, while PACE improved in grades 3, 5, and 7, but declined in grade 6. PACE results also tended to be more variable from year to year. The results are similar and indicate that PACE and Smarter Balanced tended to classify reasonably close to the same proportions of students as Level 3 or above.

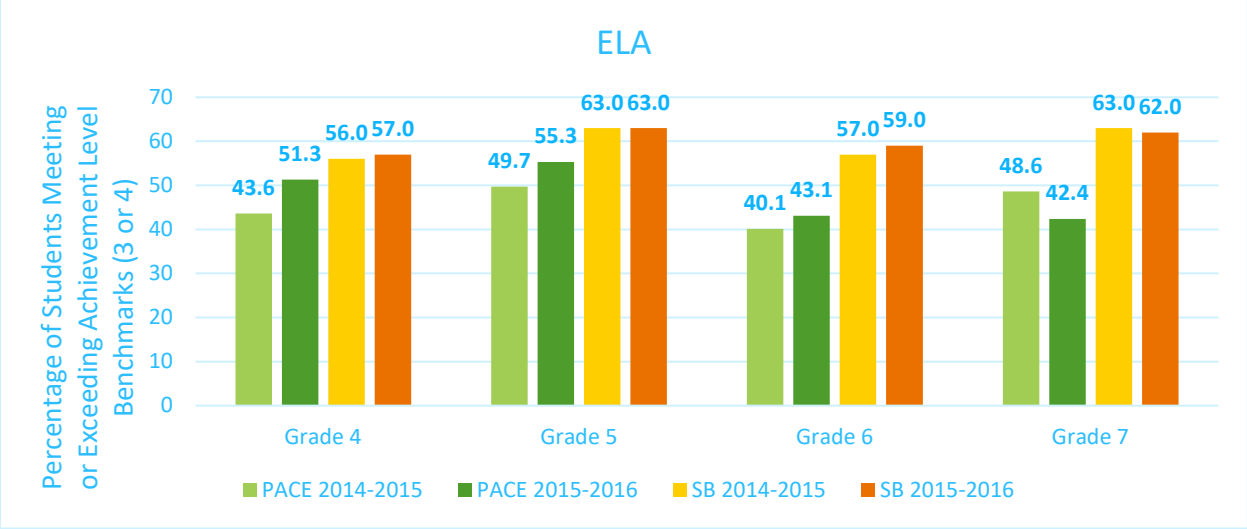

Figure 3. Comparisons of PACE and Smarter Balanced ELA Results for 2015 and 2016.[6]

[5] PACE results were not available for high school for 2015. High school results are presented in the full technical report.
[6] PACE results were not available for high school for 2015. High school results are presented in the full technical report.

Figure 3 provides the same information for ELA. The results were even more consistent for ELA. In grades 4, 5, 6, and 7 PACE classified fewer students at Level 3 of higher than Smarter Balanced. Both assessments showed improved performance from 2015 to 2016 for grades 4, 5, and 6, while both assessments showed a decline in performance at grade 7. This demonstrates that PACE and Smarter Balanced were likely classifying students similarly for ELA. Taken together, Figures 2 and 3 indicate that PACE and Smarter Balanced yielded differing results for classifying students as Proficient/Not Proficient, but those results were not so large or so variable as to call into question the similarity of the measured construct.

Student Level Correlation Results

In addition to examining scoring patterns across the PACE districts, we were also able to match a substantial portion of students' PACE scores from 2015 to their scores from 2016. PACE districts use differing scale scores, but use a common score level system (Levels 1-4), that has the same meaning for all PACE districts. We were able to correlate the scores across PACE assessments and across years to examine scoring patterns. Much like the comparisons of PACE and Smarter Balanced, we would not expect the correlations to be perfect, even for the same subject across years. If the correlations were perfect, we would not need to administer the assessments every year. Similarly, we expect scores across subjects to be correlated. Students who perform well in math tend to perform well in science and in ELA as well. So, we expect correlations that are strong and positive, but not perfect. This "Goldilocks" range of correlations that are neither too high nor too low indicate that the assessment system is functioning as expected.

We are also interested in patterns of correlations. Convergent validity coefficients (correlations between same subjects across years) should be higher than discriminant validity coefficients (correlations between differing subjects across years). We limit these comparisons to correlations across years because the time and instruction between assessments can attenuate correlations and we want to make the comparisons as similar as possible. Tables 1-5 present the correlations among the PACE scores that were available for this evaluation by grade pairs. Each correlation represents between 386 and 455 students (the number that could be matched from the 2015 and 2016 data per grade pair). All correlations are for Achievement Levels (1-4) due to differences in scale scores by district. All reported correlations for all grade pairs were statistically significant ($p < 0.01$).

Table 1 presents the correlations for grade 3 in 2015 and grade 4 in 2016. Third-grade students only had mathematics scores for 2015, but had ELA and science scores in 2016. This yielded 3 correlations, all of which were for differing subjects. Not surprisingly, the highest correlation (0.637) was between ELA and Science, both administered in 2016. Given the potential attenuating effect of using Achievement Level instead of raw or scale score, the correlations were strong and positive.

Table 1. Correlation Table Grade 3-4

	Math 2015	Science 2016
Science 2016	.487	
ELA 2016	.317	*.637*

Table 2 presents correlation results for grade 4 in 2015 matched to grade 5 in 2016. This table presents the first available convergent validity coefficient (ELA 2015 correlated with ELA 2016,

in bold), which is 0.630. This correlation is strong and positive and is higher than the two available discriminant validity coefficients (Science 2015 correlated with ELA 2016 (0.459), Science 2015 correlated with Math 2016 (0.440). This pattern of correlations represents strong validity evidence for PACE. This same pattern persists for all grade pairs except grades 7 and 8.

Table 2. Correlation Table Grade 4-5

	Science 2015	ELA 2015	ELA 2016
ELA 2015	*.555*		
ELA 2016	.459	**.630**	
Math 2016	.440	.603	.735

Table 3. Correlation Table Grade 5-6

	Math 2015	ELA 2015	ELA 2016
ELA 2015	*.635*		
ELA 2016	.520	**.619**	
Math 2016	**.625**	.590	*.648*

Table 4. Correlation Table Grade 6-7

	Math 2015	ELA 2015	ELA 2016
ELA 2015	*.470*		
ELA 2016	.482	**.586**	
Math 2016	**.558**	.513	*.531*

For grades 7 and 8 we see somewhat weaker convergent validity coefficients (0.483 for ELA, 0.488 for math). These are still strong validity coefficients, but are not as strong as for previous grade pairs. It is more telling that the discriminant validity coefficient for Math 2015 to ELA 2016 is higher than the convergent validity coefficients. This may indicate an update in task development, administration, or scoring that impacted the 2016 data and attenuated the correlations between like subjects. This analysis should be revisited next year to ensure that the correlations for this grade pair are similar to the other grade pairs and follows the expected pattern.

Table 5. Correlation Table Grade 7-8

	Math 2015	ELA 2015	Science 2016	ELA 2016
ELA 2015	*.517*			
Science 2016	.523	.448		
ELA 2016	.557	**.483**	*.574*	
Math 2016	**.488**	.477	*.590*	*.541*

Taken together, the correlation results provide strong evidence that PACE is functioning as intended. The correlations among the PACE subject areas within and across grades are similar

to other statewide assessments (Dickinson & Thacker, 2009). Correlations within year among the PACE subjects were quite high, especially for elementary grades.

Recommendations

The recommendations in this section stem from the data collected during the course of the evaluation only. There is little literature that can be directly referenced and applied to a system like PACE. For that reason, there are no statements in the recommendations section that reference aspects of similar successful programs. We did not find systems that were both successful and sufficiently similar to PACE to make direct comparisons.

The recommendations also reflect that PACE is currently functioning largely as intended. The early success of PACE is well documented in this summary report and in the associated technical report. No broad or sweeping recommendations are indicated. 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.

Recommendations for Interim Goal 1: Stakeholders Are Committed to PACE

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. The more dispersed that input becomes, the less obvious individual teacher's input may be. If some teachers perceive the PACE program as coming from the outside rather than as a direct result of their own work, buy-in could suffer. Teachers regularly commented that the cross-district collaborations are a great source of professional development, and that they greatly value those opportunities. If the new collaboration methods reduce opportunities for cross-district collaborations, then teachers may perceive less personal value in PACE. Findings from the survey indicate that those teachers who had not participated in cross-district collaborations tended to have less favorable ratings of PACE. Regular monitoring and adjustments can help safeguard against this potential issue.

Recommendations for Interim Goal 2: Assessments Are Based on Sound Test Design Principles

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. Teachers routinely reported that the process of developing the common tasks greatly improved their own task development process and their approach to assessment. As the program expands, it will be important to maintain that benefit for all participants.

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 SWD. Scaffolding should be available to all students, including SWD, and is currently built into task development activities. The task instructions for teachers now include more information about appropriate scaffolding to ensure that all students can demonstrate their knowledge, skills, and abilities. These do not alter the nature or difficulty of the tasks, but provide entry into the various activities associated with it. Accommodations are provided to SWD as a means of improving access to the content of the task. They are based on students' needs and common supports identified for each student. They may alter the task, but should support measurement of the underlying construct. As the system expands and as attrition necessitates the inclusion of new teachers, it is important that these issues continue to be addressed to ensure both accessibility and validity.

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. Several teachers indicated concerns that the reading and writing requirements for PACE were much higher than for traditional assessments. This can potentially result in reduced test score validity, especially for SWD. This phenomenon occurs when the reading/writing load interferes with the measurement of the intended construct. If, for instance, we are interested in knowing whether student 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. Instituting a system of regular task review will help ensure that happens. Some reviews have been completed at this time (by the New Hampshire Department of Education or by Stanford University), but teachers often reported that there was no feedback, or that feedback came very late from these reviews. Review of local tasks would benefit from a regularly scheduled, timely process.

Starting in the 2016-17 school year, districts will be required to submit one major assessment per competency per course, in addition to all local performance tasks in a common task template. At this stage in the evaluation, it is unknown how the assessments/tasks collected during the coming year will be reviewed, what feedback will be available to teachers/schools, or when that feedback will be provided. As this data becomes available, it will be very important to monitor the ways that feedback to teachers/schools is interpreted and used. This process has the potential be very useful and positive for the PACE program, but it also has the potential to introduce unintended consequences.

Recommendations for Interim Goal 3: Performance Assessments Are Successfully Implemented

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. If there is a benefit to spending several years in the PACE program, then that may bolster district-level support for the program and promote fidelity of implementation by educators. Teachers described a long period of adjustment and evolution of their teaching and assessment practices. 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.

Recommendations for Interim Goal 4: Scores Are Accurate and Reliable

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. It is important to consider how changes in performance are reported and how they are characterized. Early reports to USED comparing student performance on PACE with performance on Smarter Balanced within and across years[7], 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. Examples from this report include the lower correlations and reversed convergent/discriminant validity coefficient pattern for grades 7 and 8, as well as larger than typical gains in math for grade 3. Conducting these analyses again next year will help PACE determine if these anomalies are random or if they represent some systematic difference in the way PACE is implemented by grade or subject.

We also recommend that PACE provide guidance for making valid inferences from annual performance information to schools, districts, and, if possible, the media.

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 both experienced 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.

Capturing the Story of PACE

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 reduce non-productive redundancy and allow students to learn at a much faster and more customized rate. 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. The 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

[7] See https://www.education.nh.gov/assessment-systems/documents/overview.pdf.

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.

The scores generated from the PACE tasks are sufficiently reliable for their intended use and they are valid for uses beyond those that can be gained from more traditional end-of-year tests. Students understand where they performed well and where they did not. Students can be given an opportunity to redo parts of their tasks once they have addressed the areas where they were not quite ready to demonstrate competency.

PACE has had a great deal of early success, 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. They will need time in addition to the time they spend implementing PACE in their own schools and classrooms. There may also be performance gaps between the experienced and newly joined districts. These issues, as well as potential changes in the political and economic climate in which PACE is being implemented will likely challenge PACE. 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.

The current PACE has been very responsive to challenges and has improved based on feedback. For example, task development and piloting have been accelerated to make sure every task is sufficiently piloted and revised before it is used operationally. Communication regarding data collection, in-person meetings, and other important calendar-specific activities has been improved and teachers have received this information earlier in the year. This helps teachers plan and makes the PACE system more readily implemented. PACE has begun to distribute minutes from Leads meetings as a means of ensuring common understanding of decisions and future plans. PACE has established Content Leads and Teacher Leads to limit the time teachers must spend outside their classrooms. All of these examples of program improvements resulted from PACE leadership responding to requests from teachers and/or feedback from this evaluation's interim reports.

In addition to the improvements PACE has already made, more enhancements are planned for the near future. PACE leadership plans to accelerate task development even more. The goal would be to allow pilot testing of the common tasks to begin in the fall semester if that is the most appropriate time in the curriculum to use them. This would allow a more genuine piloting of the tasks and provide data even earlier to facilitate review and revision of the tasks and rubrics. The PACE Content Leads are also discussing senior projects and senior exhibitions as a natural extension of this work. One of the monthly PACE Leads meetings was devoted to a presentation related to senior projects and exhibitions. The group decided that it was a useful idea to create a separate sub-group to explore ideas for implementing these new assessment components.

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[8]. However, if PACE proves to be a substantially better system for

[8] Indicated by NH DOE leadership and reiterated by district superintendents during interviews.

educating students than the system that currently exists, it stands to reason that PACE should expand. 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. Other states may not be structured similarly. Still, there is a great deal of preparation a district must do to become a Tier 1 PACE district. It would be difficult to suddenly implement PACE on a much broader scale because of the integrated nature of task development, teacher professional development, and collaboration. Getting a full state's population of teachers to suddenly begin to effectively collaborate seems unlikely. 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.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Dickinson, E. R., & Thacker, A. A. (2009). *Relations among Kentucky's Core Content Test ACT scores and students' self-reported high school grades* (FR-09-32). Alexandria, VA: Human Resources Research Organization.

Evans, C. & Lyons, S. (2016). *NH PACE: Inter-rater reliability analysis report.* Dover, NH: Center for Assessment.