New Hampshire’s Innovative Assessment System

PERFORMANCE ASSESSMENT OF COMPETENCY EDUCATION (PACE)

Evaluating Technical Quality (Volume 1): Manual

Last Updated May 21, 2020

National Center for the Improvement of Educational Assessment

Center for Assessment

www.nciea.org

TABLE OF CONTENTS

Overview of Technical Quality Manual (Volume 2): Purpose, Content, and Structure

Purpose

This technical quality manual provides comprehensive and detailed procedural evidence in support of the validity of the NH PACE Innovative 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, mathematics, and science in grades 3-8 when those determinations are made using local, standards-based assessments rather than a standardized achievement test. High school is not included because federally-required high school annual determinations in New Hampshire are supplied by students' scores on the SAT (English language arts and mathematics) or NH SAS (science).

Each year the NH DOE and its technical partner, the Center for Assessment, gathers evidence to demonstrate and evaluate the validity of the NH PACE system using the evaluation framework explained in this manual. The annual analyses, documentation, and resources are reported each fall to the U.S. Department of Education as part of the required Innovative Assessment Demonstration Authority annual report. Those reports can be found on the U.S. Department of Education website and on the NH DOE's PACE landing page.

Annual results of these evaluations are also located in a companion document—*Evaluating Technical Quality (Volume 2): Results*. Readers are referred to Volume 2 for specific results of the analyses, documentation, and resources particular to any given school year discussed in general in this technical quality manual.

Content

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. 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.
- ✓ **Reliability** is the consistency and the generalizability of the inferences about students' knowledge and skills over varying conditions and contexts such as raters, tasks, and occasions.

[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.

- ✓ **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 balanced 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, continuous, efficient,* and *useful*. These concepts of a high quality assessment system are not new, but are drawn from the National Research Council's *Knowing What Students Know*[2] and more recent discussion of balanced assessment system design[3].

- ✓ **Comprehensive** –The NH PACE system includes a range of measurement approaches "to provide a variety of evidence to support educational decision making"[4]. 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 system 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 –** 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.
- ✓ **Efficient –** The NH PACE system is efficient in that the information collected as part of the classroom assessment system is used to fulfill the state assessment system requirements and unnecessary and/or redundant assessments are minimized. There is no extra testing required as the PACE system is curriculum-embedded.
- ✓ **Useful –** The NH PACE system provides teachers, students, and parents with a continuous stream of performance information throughout the year. This actionable, real-time data can be used to make better instructional decisions and understand student progress towards proficiency on the state's academic content standards when adjustments can still be made.

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

[3] 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.

[4] Pellegrino, et al., 2001, p. 253.

Structure

This manual explains the procedures used to evaluate the technical quality of the NH PACE Innovative Assessment and Accountability System. Annual results of these evaluations are located in the companion document—*Evaluating Technical Quality (Volume 2): Results*. This manual starts with a brief overview of the PACE system, including: history, background, key design characteristics and theory of action. A comprehensive framework for evaluating the technical quality of the PACE system is then presented and used as an organizing structure for this manual. The manual ends with a section on reporting and interpretation of PACE system results.

Overview of the NH PACE Innovative Assessment and Accountability System

History

New Hampshire was awarded a waiver from federal statutory requirements under the *No Child Left Behind Act* related to state annual achievement testing by the U.S. Department of Education in March 2015. The proof of concept pilot, referred to as Performance Assessment of Competency Education or PACE, operated under NCLB federal waivers from the 2014-15 school year through the 2017-18 school year.

The PACE pilot then transitioned under new federal education legislation. The theory of action and design of the PACE system did not change. Instead, New Hampshire applied and was awarded the Innovative Assessment Demonstration Authority under Section 1204 of the *Every Student Succeeds Act* in the summer of 2018. The PACE innovative assessment and accountability system has operated under this demonstration authority since the 2018-19 school year.

Background

The PACE system was designed to support deeper learning for students and powerful organization change for schools and districts. PACE is grounded in a competency-based educational approach designed to ensure that students have meaningful opportunities to achieve critical knowledge and skills.

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).

Design Features

The PACE system is designed using a combination of local, common, and state level assessments (see Table 1). The core of the PACE innovative assessment system is locally-developed, locally-administered performance assessments tied to grade and course competencies determined by local school districts. In each PACE grade and subject (see orange boxes in Table

1), one, common complex performance task called the PACE Common Task is collaboratively developed and administered by all participating schools and districts. The PACE Common Tasks are NOT a state test. Rather, the PACE Common Tasks are designed to serve as calibration tools, providing evidence that each teacher's evaluation of student performance is comparable to the evaluations made by other teachers. The PACE Common Task is administered at any point in the school year at the discretion of each district and in ways that support its intended curriculum-embedded nature.

Table 1

PACE System Overview by Grade and Subject

Grade	ELA	Math	Science
3	Statewide assessment system (NH SAS)	Performance assessment system (PACE)	Local Performance Assessments
4	Performance assessment system (PACE)	Statewide assessment system (NH SAS)	Local Performance Assessments
5	Performance assessment system (PACE)	Performance assessment system (PACE)	Performance assessment system (PACE)
6	Performance assessment system (PACE)	Performance assessment system (PACE)	Local Performance Assessments
7	Performance assessment system (PACE)	Performance assessment system (PACE)	Local Performance Assessments
8	Statewide assessment system (NH SAS)	Statewide assessment system (NH SAS)	Performance assessment system (PACE)
9	Course-specific common performance assessments	Course-specific common performance assessments	Course-specific common performance assessments
10	Course-specific common performance assessments	Course-specific common performance assessments	Course-specific common performance assessments
11	Statewide assessment system (SAT)	Statewide assessment system (SAT)	Statewide assessment system (NH SAS)

Determinations of student proficiency in the PACE grades/subjects required under federal law are produce using: (1) teacher judgments at the end of the school year regarding which achievement level best describes each of their students; and (2) end of year competency scores for each student. PACE is designed so that the resulting levels are comparable in rigor and substance to the statewide academic assessment (NH SAS) by using achievement level descriptors that are aligned across the two systems.

The statewide academic assessments (NH SAS) are administered in a few grades and subjects in the PACE system according to when the results will be most useful for informing programs and auditing the innovative assessment system—grade 3 ELA, grade 4 math, grade 8 ELA and math, and grade 11 ELA and math and high school science (see green boxes in Table 1).

Theory of Action

The NH PACE theory of action is grounded in the latest advances regarding how students learn[5], how to assess what students know[6], and how to foster positive organizational learning and change[7]. Figure 1 illustrates a version of the PACE theory of action with system design features on the left and intended outcomes on the right. The purpose of this theory of action is to illustrate broadly how implementation of the PACE system is intended to impact the instructional core of classroom practices, thereby advancing college and career readiness. The instructional core is the intersection of meaningful content, high-quality teaching, and engaged students[8].

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 an internal accountability system. This fosters positive organizational learning and change by supporting the internal motivation of educators.

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 facilitates transfer to new real-world situations. These approaches also 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

[5] 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.

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

[7] 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.

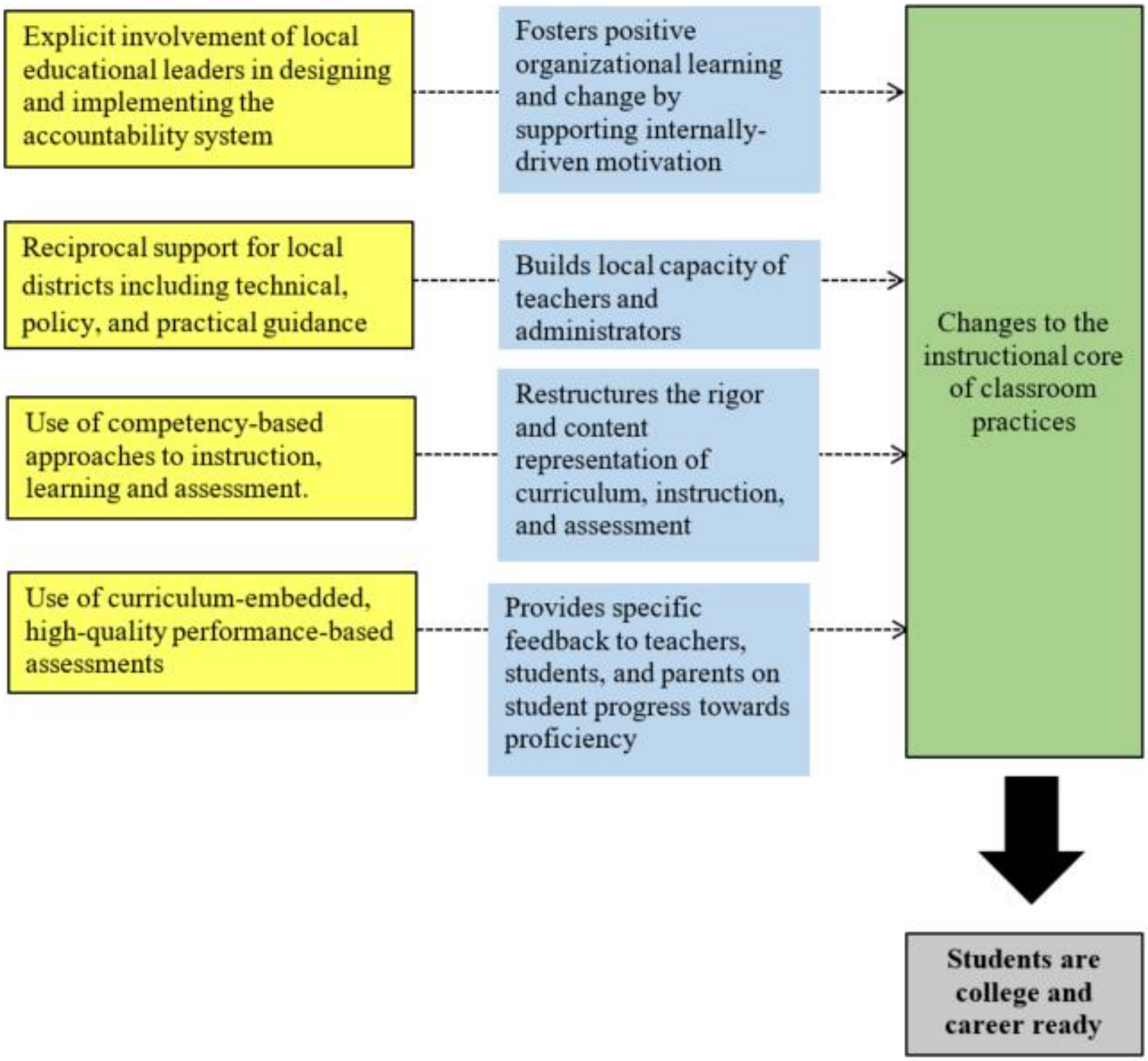
Pink, D. H. (2009). *Drive: The surprising truth about what motivates us*. New York, NY: Riverhead Books.

[8] City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). *Instructional rounds in education: A network approach to improving teaching and learning*. Cambridge, MA: Harvard Education Press.

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 Tasks are exemplar high-quality performance assessments with associated rubrics and scoring protocols and procedures. There is now a growing bank of high-quality performance tasks (retired Common 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

PACE System Theory of Action


Explicit involvement of local educational leaders in designing and implementing the accountability system
Fosters positive organizational learning and change by supporting internally-driven motivation
Reciprocal support for local districts including technical, policy, and practical guidance
Builds local capacity of teachers and administrators
Use of competency-based approaches to instruction, learning and assessment.
Restructures the rigor and content representation of curriculum, instruction, and assessment
Use of curriculum-embedded, high-quality performance-based assessments
Provides specific feedback to teachers, students, and parents on student progress towards proficiency
Changes to the instructional core of classroom practices
Students are college and career ready

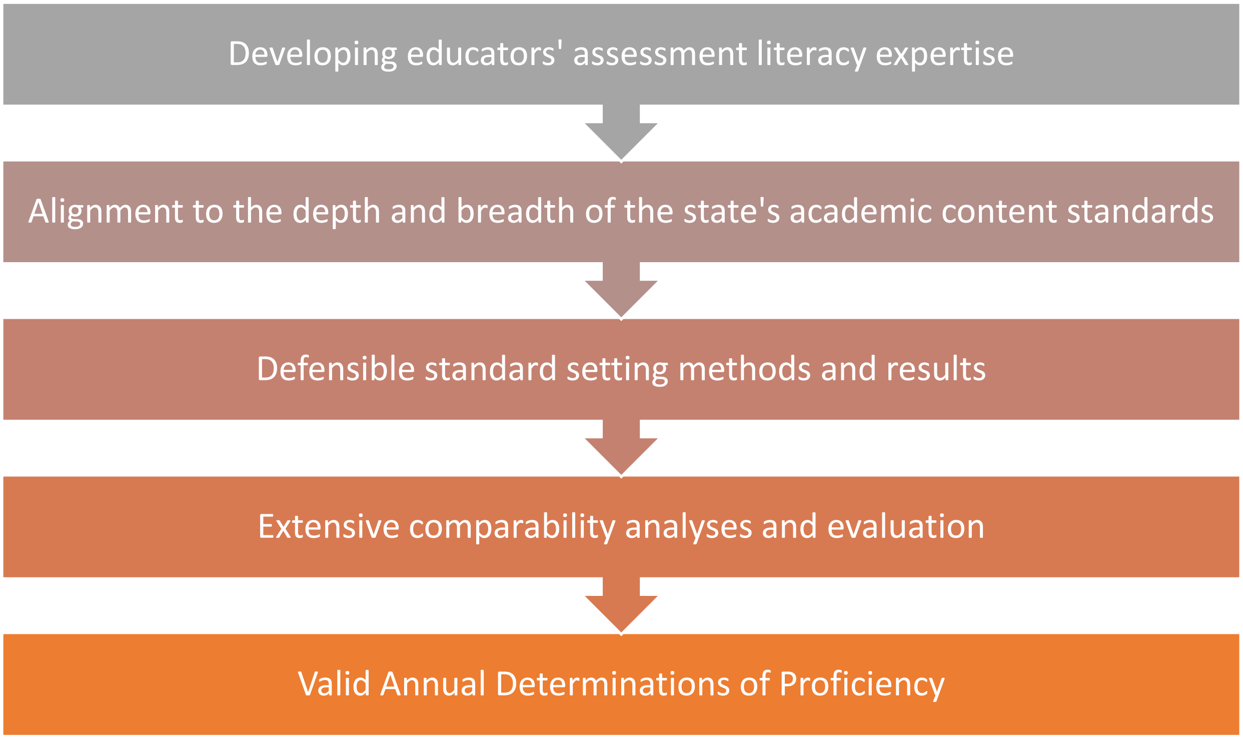
Framework for Evaluating the Technical Quality of the PACE System

This technical quality manual provides comprehensive and detailed evidence in support of the validity of the NH PACE Innovative Assessment and Accountability System. 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 supports the validity of PACE system results for use in reporting on student achievement and inclusion in the state's school accountability system. Like a good argument, the evidence presentation follows a story from the theory of action or the logic model guiding PACE to the ultimate claims supporting the defensibility of the PACE annual determinations of proficiency.

The following graphic highlights the chain of reasoning supported by evidence presented in this report that supports PACE system results and uses. The analyses and documentation presented in this report supports the validity of inferences from PACE assessments for the intended uses in the PACE system.

Figure 2

Framework for Evaluating the Technical Quality of the PACE System


Developing educators' assessment literacy expertise
Alignment to the depth and breadth of the state's academic content standards
Defensible standard setting methods and results
Extensive comparability analyses and evaluation
Valid Annual Determinations of Proficiency

Section 1: Developing Educators' Assessment Literacy Expertise

Developing the assessment literacy of school/district leaders and teachers is a key components of supporting balanced assessments systems[9]. Assessment literacy refers to the knowledge and skills associated with designing, using, interpreting, and/or selecting high-quality assessments to improve student learning and to serve other important educational and policy purposes. Given the central role of high-quality local assessment systems to the validity of PACE system results, there are multiple approaches to professional development embedded within the system that are tailored to the unique (though related) assessment literacy needs of school/district leaders and teachers[10].

School and District Leaders

School and district leaders receive professional development targeted to their assessment literacy needs at monthly PACE leadership meetings. Topics are chosen based on expressed needs of PACE school and district leaders and may vary from year-to-year. Some topics are repeated over time; others are one-time conversations based on events. Examples of topics include: design of district assessment systems, auditing the quality of district assessment systems, understanding assessment results, making sense of multiple measures, interpreting and using within-district and cross-district analyses related to the reliability of teacher scoring within their district, benefits and limitations of interim assessments and state assessment results, and so on.

Teachers

The approach to teacher professional development has rested largely on the cross-district task development sessions in which teachers are trained and coached in a sustained, on-going way on the development and use of performance assessments in their classrooms. The professional development partners at the New Hampshire Learning Initiative (NHLI) and assessment experts at the Center for Assessment support these capacity building efforts.

However, these meetings are by no means the only opportunities for professional development offered to teachers. All teachers implementing PACE undergo within-district training on task implementation and scoring—including calibration sessions. In addition to cross-district task development and within-district implementation and calibration training, the following professional development is offered to PACE teachers. Each is discussed in more detail below.

1. Content Leaders: Advanced training is provided to select PACE teachers in assessment design and development.

[9] Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2019). *The challenges and opportunities of balanced systems of assessment: A policy brief.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from https://www.nciea.org/sites/default/files/publications/Assessment%20Systems%20Policy%20Brief.pdf

[10] Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2019). *A tricky balance: The challenges and opportunities of balanced systems of assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education (Toronto, Ontario). Retrieved from https://www.nciea.org/sites/default/files/publications/A%20Tricky%20Balance_031319.pdf

2. PACE Summer Institute: This opportunity is open to all PACE teachers with multiple strands of professional development including training in cross-district calibration, reviewing of student bodies of work, introductory and advanced task development, and leadership training.

Content Leads. Content leads receive advanced performance assessment training, including discussions of how to apply principled assessment design processes to performance assessment development and scoring. Additionally, content leads receive support, tools, and resources relating to 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 learners to help them work with their colleagues to support the development of high-quality common performance tasks. Content leaders are responsible for the following duties:

- ✓ Support their colleagues in the development of the local and common performance tasks,
- ✓ Facilitate the performance task development process,
- ✓ Review the online task bank 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, and
- ✓ Lead the review of student work from the pilot to improve the task.

PACE Summer Institute. Teachers from implementing PACE 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. Typically, over 80% of teachers agree each year that the calibration activities positively impact them professionally.

Section 2: Alignment to Depth and Breadth of State Academic Content Standards

The PACE system is aligned to the depth and breadth of the state's academic content standards. There are four main sources of evidence demonstrating alignment: (1) comprehensive local assessment system peer reviews; (2) high-quality performance task development process; (3) rigorous PACE common task expert reviews; and (4) administration of extended, high-quality, and complex performance assessments throughout the year to measure the depth and breadth of the state's academic content standards.

Comprehensive Local Assessment System Peer Reviews

The NH DOE and the Center for Assessment collect and review local summative assessment maps and aligned summative assessments from all participating PACE schools and districts as part of the annual Data Collection Protocols. The purpose of reviewing the assessment maps and aligned summative assessments is to ensure all students are provided with a meaningful and multiple opportunities to demonstrate proficiency on required grade level content standards and competencies at the appropriate level of cognitive complexity. The review of a sample of summative assessments evaluates alignment to the state content standards and examines the quality of local summative assessments used to inform competency determinations throughout the year. Participating PACE schools and districts submit materials in all grades and subjects where annual determinations of student proficiency are produced in the PACE system. Other grades and subjects may be submitted by a school or district looking for feedback on the coherence of their K-12 local assessment system.

The assessment maps and aligned summative assessments provide the base level of assurance and documentation that all state academic content standards are addressed in the local assessment system and that students are assessed at the depth and breadth of knowledge appropriate for the state academic content standards. The assessment maps and aligned summative assessments document:

- ✓ The competencies assessed in each course
- ✓ The alignment of the state academic standards to the competencies
- ✓ The alignment of the local summative assessments to the State academic standards
- ✓ The number, type, and timing of the summative assessments administered for each competency
- ✓ The quality of local summative assessments

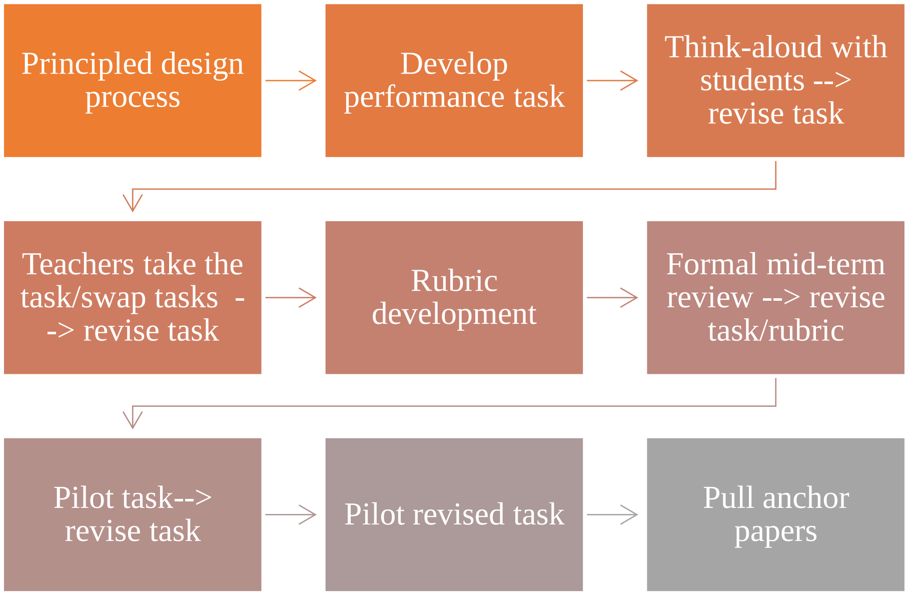
The assessment maps and aligned summative assessments are peer reviewed each year by the NH DOE, PACE districts, and Center for Assessment experts. Districts are provided formative feedback from the peer reviewers on the quality of their local assessment system as represented in the assessment maps and aligned summative assessments.

High-Quality Performance Task Development Process

The development of high-quality PACE Common Tasks is grounded in the training of local educators. Teams of teacher content leaders from all PACE districts who receive advanced assessment coaching are responsible for leading much of the task development work with their fellow teachers. Teachers from all PACE districts collaborate in grade and subject area teams and follow a disciplined process of task development. Figure 3 illustrates the PACE Common Task development and pilot-testing process.

Figure 3

PACE Common Task Development and Pilot-Testing Process


Principled design process
Develop performance task
Think-aloud with students --> revise task
Teachers take the task/swap tasks --> revise task
Rubric development
Formal mid-term review --> revise task/rubric
Pilot task--> revise task
Pilot revised task
Pull anchor papers

The process begins with a principled assessment design process that incorporates the principles of Universal Design for Learning (UDL). PACE teachers are trained through the process of PACE Common Task development to consider UDL in their design of local performance tasks and assessments. This 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" model[11] performance task template is used to provide guidance on the characteristics of a high-quality task and PACE expectations. 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 state content standards and competencies for English language arts, math, and science. In line with principles of UDL, PACE Common Task developers consider during the design phase the extent to which the

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

performance task provides students with (1) *multiple means of representation* to give learners various ways of acquiring information and knowledge, (2) *multiple means of expression* to provide learners alternatives for demonstrating what they know, and (3) *multiple means of engagement* to tap into learners' interests, challenge them appropriately, and motivate them to learn.

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*. 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. 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 PACE High-Quality Assessment Review Tool. 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.

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 anchor papers are identified 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.

Purposes of PACE common tasks. 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 section 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 for operational use by the NH DOE and Center for Assessment. The result is the set of operational tasks that provide models for designing rich, authentic assessment experiences that measure deep learning. The tasks are reviewed specifically to evaluate the degree of independent student inquiry, determine the extent of multi-step problem solving and argument building required, and examine the ability to employ 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 system 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: maintaining 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. Previous 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.

Rigorous PACE Common Task Expert Reviews

PACE Common Tasks go through a rigorous technical review by the Center for Assessment experts prior to operational use. Reviewers evaluate the extent to which the PACE Common Task is aligned to the state’s academic content standards and competencies, the quality of the scoring guidelines and criteria, use of fair and unbiased presentation and response availability, and use of appropriate text/visual resources. Particular attention is paid to the appropriate specifications around accommodations for students with disabilities and English language learners using the principles of UDL. Specifically, PACE Common Tasks are reviewed based on whether they measure student skills that are outside the intended construct, use extraneous words that potentially distract students from the main learning target of the task, use idioms, or culturally-specific language, crowd text and/or graphics too closely on the page, and/or use graphics that require certain levels of visual acuity to understand.

The PACE Common Tasks are reviewed in an on-going, formative way where specific and meaningful feedback is provided to the teachers involved in task development during the design and piloting phase, which takes places in the year prior to operational use. Task developers used the feedback to revise/edit the PACE Common Tasks until they are ready for final approval by the Center for Assessment and NH DOE.

Administration of High-Quality Performance Tasks Throughout the Year

One of the most compelling sources of evidence for alignment, particularly the depth of knowledge criterion, is the use of the performance assessments to measure high-order thinking skills and understanding. PACE relies on curriculum-embedded, extended, high-quality, and complex performance-based assessments to assess deeper learning. The use of local and common extended performance tasks allows the PACE system to validly measure the true depth of the state's academic content standards and competencies.

Section 3: Defensible Standard Setting Methods and Results

The purpose of standard setting is to designate cut scores that define the four levels of performance for the PACE Annual Determinations. Standard setting plays a central role in the validity of the interpretations drawn from the scores in essentially all standards-based assessment systems. This is especially true for PACE due to three main reasons:

1. PACE does not report out any individual-level scale scores. This places extra burden on the validity of the interpretations drawn from the achievement level designations.
2. Each PACE district has a unique scale associated with their competency scores. Even if the scales are nominally the same (e.g., 1-4) the interpretations associated with the score points will differ across districts due to differences in scoring practices. Therefore, PACE standard setting is used as a critical aspect of comparability for the PACE assessment system.
3. The PACE innovative assessment system is required to produce annual determinations that are comparable to the statewide assessment system. Therefore, the standard setting methodology is grounded in achievement level descriptors that are aligned across systems. Each of the achievement levels is intended to carry the same interpretations about what students know and can do whether they participate in PACE or NH SAS.

Over the years, the PACE assessment system has achieved a strong record of creating comparable annual determinations. This has required leveraging multiple methods (e.g., see Performance Standards Validation) and refining our psychometric processes to continuously improve as we scale. We have relied primarily on a contrasting groups standard setting methodology described in more detail below.

Contrasting Groups Standard Setting Method

The PACE standard setting method involves two primary steps: 1) collecting teacher judgments regarding student placement into achievement levels using the PACE achievement level descriptors and 2) setting cut scores on each districts' competency score scale (scale refers to each district, grade, and subject combination) using the teacher judgements in a contrasting groups methodology.

This standard setting method involves asking teachers to make judgments about the achievement level of the students based on their professional judgment and knowledge of the student. The teachers are provided with rich, narrative descriptions of each of the achievement levels called Achievement Level Descriptors (ALD). Every PACE teacher completes a teacher judgment survey at the end of the school year to make gestalt judgments about which achievement level best describes each of their students. The subject and grade specific PACE ALDs are entered into an online survey where teachers can easily read the descriptions and match their students to the appropriate achievement level. This process relies heavily teacher knowledge of each of their students and on a common understanding and interpretation of the ALDs. As mentioned previously, the PACE ALDs are aligned to the NH SAS ALDs and intended to carry the same interpretations about what students know and can do at each achievement level.

The contrasting groups standard setting methodology involves comparing the average PACE competency scores with the teacher judgment scores in order to determine the cut scores that most accurately classify the students into the achievement levels. Logistic regression is used to determine the point in the score distribution where examinees have a 50% chance of being classified in the next performance level or above (e.g., the probability that a student with a score of X has a 50% or greater probability of being classified in Level 3 or higher). A logistic regression analysis is run separately for each cut point—Level 2, Level 3, and Level 4—in each district, subject, and grade.

Quality Control Processes and Procedures

Data quality control checks and district flagging business rules are used to ensure the quality of factors related to producing cut scores and are completed prior to calculating PACE cut sores.

Data quality control checks. The data quality control checks include a systematic process for ensuring the data quality prior to running the logistic regression. The data quality control checks include the following:

- ✓ Flag out of bound values (e.g., 0.75 on a scale of 1.00 - 4.00).
- ✓ View raw data by scale (district, grade, and subject) to complete human reasonableness checks. We use scatterplots of end of year competency scores by teacher judgment survey ratings for each district, grade, and subject combination.
- ✓ Verify the number of student records received matches the expected enrollment by scale.
- ✓ Replicate end of year competency score averages provided by state using disaggregated competency score data.

District flagging business rules. Submitted teacher judgment survey ratings are analyzed by district, grade, and subject in order to identify unexpected distributions of teacher judgment prior to calculating PACE cut scores. The flagging rules evaluate variability in the teacher judgment survey ratings by district, grade, and subject in three ways:

1. Identify instances where there is *no variance* in teacher judgment survey ratings (i.e., all 1s, all 2s, all 3s, or all 4s);
2. Identify instances where there is *reduced variance* in teacher judgment survey ratings (i.e., all 1s and 2s, all 2s and 3s, or all 3s and 4s); and
3. Identify instances where there is *bimodal distribution* of teacher judgment survey ratings (i.e., all 1s and 3s, all 1s and 4s, or all 2s and 4s).

Instances where teacher judgment survey ratings show evidence of no variance, reduced variance, or bimodal distribution are then analyzed using the Table 2 decision matrix below. The decision matrix guides follow-up decisions with districts and was created to balance the need for district follow-up with the realities of data issues that result from very small sample sizes. Step 1 is a simple examination of the sample size in the district, grade, and subject combination. Step 2 is an examination of the percent of students proficient or above from prior state standardized assessment results for the district and subject in the grade level closest to the grade level under investigation. Given the design of the PACE assessment system and based on the number of years the district has been involved in PACE, the available state assessment data may be limited to grade 3 ELA, grade 4 Math, or grade 8 ELA and math.

Table 2

PACE Flagging Rules for Variability in TJS Ratings Decision Matrix

Flag for TJS Ratings	Step 1: Examine Sample Size	Step 2: Examine Prior State Standardized Assessment Results
No variance	<=5 students→no follow-up >5 students→go to Step 2	Percent of students proficient is within ± 5% of the prior state standardized assessment results→no follow-up
Reduced variance	<=15 students→no follow-up >15 students→go to Step 2	
Bimodal distribution	<=15 students→no follow-up >15 students→go to Step 2	Otherwise the district will be contacted by the NH DOE or the Center for Assessment to verify the teacher judgment survey results.

The complete district flagging business rules analysis along with the subsequent decisions related to each flag based on the decision matrix is reported to the NH DOE by the Center for Assessment each year. It is atypical to contact districts for follow-up based on no variance, reduced variance, or bimodal distributions in the teacher judgment survey ratings. In most years, teacher judgment survey ratings tend to concentrate in Levels 2 and 3 (about 75% of the time), the other 25% of judgments are distributed between Levels 1 and 4.

If follow-up with districts on the distribution of their teacher judgment survey ratings is deemed necessary, the business rules specify that the Center for Assessment will not calculate cut scores until teacher judgment survey results can be verified with the district. If the teacher judgment survey results cannot be verified with the district then the district will be notified that they will receive PACE determinations for the year, but the district will need to take NH SAS along with submitting PACE data in the following year. Results from NH SAS in the following year will be compared to PACE standard setting results and if within ± 5% on percent proficient or above in the same grade and subject area then the district will not need to administer the NH SAS the following year. Otherwise the process will continue until the district meets the ± 5% on the proficiency threshold.

Cut Score Calculation Business Rules

Cut score calculation rules are used to ensure consistency in setting standards by delineating rules for the following:

- ✓ Addressing every possible pattern of presence/absence of teacher judgments placing student achievement in each achievement level,
- ✓ Describing the statistical process (dichotomous logistic regression) used for estimating cut scores where there are sufficient data, and
- ✓ Ensuring consistency in calculating cut scores when there are problems with estimating a cut score using the logistic regression.

There are two major parts in cut score calculation: (1) initial cut score calculations, including logistic regression of teacher judgments of students' achievement being at or above a given achievement level on students' mean competency scores to estimate cut scores for a given scale (a scale is a district, grade, and subject combination); and (2) alternate cut score calculations for situations in which the logistic regression does not converge or in which the logistic regression found a lower probability of students being at or above a specific achievement level associated with increases in mean competency scores.

The business rules take the following form:

1. For each student, identify the scale on which the student's mean competency scores exist. Typically, each school administrative unit (SAU) has its own scale in each year, subject, and grade. However, there are some exceptions to this general rule in that in some districts within a SAU may also have separate scales. The scale for each student can be uniquely identified by doing the following:
    a. For each student, obtain in the standard setting data file the value of the following variables: *District_Name* and/or *District_ID*, *Scale_Year*, *Scale_Grade*, and *Scale_Subject*;
    b. Identifying the single row in the *PACE Entity Master* data file that has those same values for the same variables; and
    c. Extracting from that row the value of the variable/column labeled *Scale_ID*.
2. Saving the Scale_ID to the appropriate row of the standard setting data file.
3. For each scale, do the following:
    a. For each achievement level, identify whether the scale has at least one teacher judgment rating in that level (*1*) or not (*0*);
    b. Create a four-bit string (*HasX*) combining the *0/1* designations from the previous step with the left-most indicating presence/absence of a rating in level 1 and the right-most indicating presence/absence of a rating in level 4 (e.g., *0110* would indicate ratings in levels 2 and 3 but no ratings in levels 1 and 4);
    c. Using the four-bit string identified in the prior step, follow the rules for calculation given in Table 4 which shows three calculations in order (i.e., first calculation, second calculation, third calculation) covering three cut scores that correspond to the four-bit string. For this table, the names of variables are explained in Table 3 and *cut(…)* represents estimating the logistic regression described above and, if the results converge and do not predict higher achievement levels for lower scoring students, the mean competency score at

which the probability of being in a higher category passes 50 percent. The cut score is identified as the mean competency score with the lowest value from 10,000 equally separated values from the minimum possible competency score to the maximum possible competency score with a probability greater than or equal to 50%. The order of calculations prioritizes calculation of the cut score between levels 2 and 3, followed by the cut score between levels 1 and 2, followed by the cut score between levels 3 and 4. Where there are insufficient data to calculate a cut score, the others are calculated first, so there may be some different orderings to reflect this caveat.

   d. If any given cut score was problematic, it should remain uncalculated to wait for the next step.

4. For each scale with at least one cut score where the logistic regression was problematic, do the following:
   a. Create a three-bit string (*Needed*) identifying for each cut score whether the cut score calculation was problematic (for example, “011” indicates that the cut score between levels 1 and 2 was successfully calculated, but the cut scores between levels 2 and 3 and levels 3 and 4 were problematic).
   b. Using the three-bit string (*Needed*) identified in the prior step, follow the rules for calculation given in the corresponding row of Table 5 (which shows up to three ordered calculations; i.e., first calculation, second calculation, third calculation).

Table 3

Explanation of variables used in business rules.

Full	Description
Cut12	Scale-specific cut score between levels 1 and 2
Cut23	Scale-specific cut score between levels 1 and 3
Cut34	Scale-specific cut score between levels 3 and 4
MinPossCS	Scale-specific minimum possible competency score (or LOSS when LOSS = *Lowest Observ**able** Scale Score*)
MaxPossCS	Scale-specific maximum possible competency score (or HOSS when HOSS = *Highest Observ**able** Scale Score*)
MinObsMCS	Scale-specific minimum attained mean competency score (or LOSS when LOSS = *Lowest Observ**ed** Scale Score*)
MaxObsMCS	Scale-specific maximum attained mean competency score (or HOSS when HOSS = *Highest Observ**ed** Scale Score*)
Has1	Scale has at least one student in achievement level 1 as judged by teacher in the dummy-variable form [0 | 1]

Has2	Scale has at least one student in achievement level 2 as judged by teacher in the dummy-variable form [0 | 1]
Has3	Scale has at least one student in achievement level 3 as judged by teacher in the dummy-variable form [0 | 1]
Has4	Scale has at least one student in achievement level 4 as judged by teacher in the dummy-variable form [0 | 1]
HasX	As-character concatenation of Scale_HasAL1, Scale_HasAL2, Scale_HasAL3, and Scale_HasAL4
AL	Student achievement level as judged by teacher at the end of the year (1, 2, 3, or 4)
Met2	Student achievement is at the end of the year judged by the teacher to at or above achievement level 2 (1) or not (0)
Met3	Student achievement is at the end of the year judged by the teacher to be in achievement level 3 or 4 (1) versus achievement level 1 or 2 (0)
Met4	Student achievement is at the end of the year judged by the teacher to be in achievement level 4 (1) versus achievement level 1, 2, or 3 (0)
MCS	Student mean competency score at the end of the year
'12'	Parameter indicating that the cut score between achievement levels 1 and 2 should be calculated
'23'	Parameter indicating that the cut score between achievement levels 2 and 3 should be calculated
'34'	Parameter indicating that the cut score between achievement levels 3 and 4 should be calculated

Table 4

Business rules for calculating cut scores based on presence or absence of teacher judgments in each category (Step 1 level).

HasX	First Calculation	Second Calculation	Third Calculation
0001	Cut23 <- (Cut12 + Cut34) / 2	Cut34 <- MinObsMCS	Cut12 <- MinPossCS + (Cut34 - MinPossCS) / 3
0010	Cut34 <- MaxObsMCS	Cut12 <- MinPossCS + (Cut23 - MinPossCS) / 2	Cut23 <- MinObsMCS
0100	Cut23 <- MaxObsMCS	Cut12 <- MinObsMCS	Cut34 <- (Cut23 + MaxPossCS) / 2
1000	Cut12 <- MaxObsMCS	Cut23 <- Cut12 + (MaxPossCS - Cut12) / 3	Cut34 <- Cut34 <- (Cut23 + MaxPossCS) / 2
0011	Cut23 <- (Cut12 + Cut34) / 2	Cut34 <- cut('34', Met4, Cut12, Cut23, Cut34, MCS)	Cut12 <- MinPossCS + (Cut34 - MinPossCS) / 3
0101	Cut23 <- (Cut12 + Cut34) / 2	Cut34 <- cut('34', Met4, Cut12, Cut23, Cut34, MCS)	Cut12 <- MinPossCS + (Cut34 - MinPossCS) / 3
0110	Cut23 <- cut('23', Met3, Cut12, Cut23, Cut34, MCS)	Cut12 <- MinPossCS + (Cut23 - MinPossCS) / 2	Cut34 <- (Cut23 + MaxPossCS) / 2
1001	Cut23 <- (Cut12 + Cut34) / 2	Cut34 <- cut('34', Met4, Cut12, Cut23, Cut34, MCS)	Cut12 <- MinPossCS + (Cut34 - MinPossCS) / 3
1010	Cut23 <- cut('23', Met3, Cut12, Cut23, Cut34, MCS)	Cut12 <- MinPossCS + (Cut23 - MinPossCS) / 2	Cut34 <- (Cut23 + MaxPossCS) / 2
1100	Cut12 <- cut('12', Met2, Cut12, Cut23, Cut34, MCS)	Cut23 <- MaxObsMCS	Cut34 <- (Cut23 + MaxPossCS) / 2
0111	Cut34 <- cut('34', Met4, Cut12, Cut23, Cut34, MCS)	Cut12 <- MinPossCS + (Cut23 - MinPossCS) / 2	Cut23 <- cut('23', Met3, Cut12, Cut23, Cut34, MCS)
1011	Cut34 <- cut('34', Met4, Cut12, Cut23, Cut34, MCS)	Cut12 <- MinPossCS + (Cut23 - MinPossCS) / 2	Cut23 <- cut('23', Met3, Cut12, Cut23, Cut34, MCS)
1101	Cut12 <- cut('12', Met2, Cut12, Cut23, Cut34, MCS)	Cut34 <- cut('34', Met4, Cut12, Cut23, Cut34, MCS)	Cut23 <- (Cut12 + Cut34) / 2
1110	Cut23 <- cut('23', Met3, Cut12, Cut23, Cut34, MCS)	Cut12 <- cut('12', Met2, Cut12, Cut23, Cut34, MCS)	Cut34 <- (Cut23 + MaxPossCS) / 2
1111	Cut23 <- cut('23', Met3, Cut12, Cut23, Cut34, MCS)	Cut12 <- cut('12', Met2, Cut12, Cut23, Cut34, MCS)	Cut34 <- cut('34', Met4, Cut12, Cut23, Cut34, MCS)

Table 5

Business rules for calculating cut scores based on whether each logistic regression had problematic results (Step 2 level).

Needed	Cut12	Cut23	Cut34
001			Cut34 <- MaxPossCS
010		Cut23 <- (Cut12 + Cut34) / 2	
011		Cut23 <- (Cut12 + MaxPossCS) / 3	Cut34 <- MaxPossCS
100	Cut12 <- (MinPossCS + Cut23) / 2		
101	Cut12 <- (MinPossCS + Cut23) / 2		Cut34 <- MaxPossCS
110	Cut12 <- (MinPossCS + MinPossCS + Cut34) / 3	Cut23 <- (MinPossCS + Cut34) / 2	
111	Cut12 <- (MinPossCS + Cut23) / 2	Cut23 <- (MinPossCS + MaxPossCS) / 2	Cut34 <- MaxPossCS

Application of Cut Score Calculation Business Rules

The results of the contrasting groups standard setting analyses with applied cut score calculation business rules is includes in the annual standard setting report created by the Center for Assessment. If a cut score calculation business rule was applied it can be found under "Result12", "Result23" or "Result34".

- "<Estimated successfully>" means that no business rule was applied to produce a cut score.
- "Set via step 1 rule>" means that the absence of a teacher judgment survey rating in a particular achievement level necessitated the application of the cut score calculation business rules found in Table 4 above.
- "<Set via step 2 rule after estimation failed to converge>" means that the logistic regression did not estimate successfully (due to small sample size, for example) and therefore the cut score calculation business rules found in Table 5 above were applied.

Quality Assurance Processes and Procedures

Prior to submitting the calculated cut scores as final to the NH DOE, the Center for Assessment conducts several impact analyses to evaluate the consistency and stability of the cut scores. The purpose of these quality assurance process and procedures is to review the outcome and reasonableness of the cut scores produced using historical data to flag results that seem unlikely or unreasonable given trends over time for each scale.

Historical data from previous years of the PACE and NH SAS system are used alongside the most recent year of data whenever possible. Impact analyses are run at both the system- and district-level. The impact analyses include:

- ✓ Cohort analysis: Examines how students in a given grade/subject perform in comparison to students in the same grade/subject for the previous year and any other years of data available using percent of students proficient or above;
- ✓ Longitudinal analysis: Compares how students in a given grade perform in the previous grades (same subject) for the previous year and any other years of data available using percent of students proficient or above; and
- ✓ State test analysis: Compares proficiency rates between PACE and NH SAS in grades 3-8 using percent of students proficient or above by subject.
- ✓ Performance level analysis: Compares the percent of students in each performance level (1, 2, 3, or 4).

Performance Standards Validation

We employ a “body of evidence” approach to help evaluate the PACE performance standards each year. This analysis is an additional source of validity evidence to support the PACE innovative assessment system. This approach involves collecting a Body of Work (BOW) for a small sample of students from a sample of courses in each participating district. Districts are instructed to select nine students representing 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. Districts are also instructed to collect samples of student work from those nine students for each of the state competencies in a given grade/subject area. The student work samples included in the Body of Work (BOW) portfolios are from major classroom summative assessments throughout the year in order to demonstrate student achievement for each of the grade/subject competencies.

Teachers from across PACE districts come together at the PACE Summer Institute each year to participate in a modified Body of Work standards validation process. The purpose of the validation process is to review portfolios of student work and make judgments about student achievement relative to the PACE Achievement Level Descriptors. Teachers are randomly assigned to cross-district teams of two to four people and asked to independently rate bodies of work from other districts using the PACE Achievement Level Descriptors. The independent ratings take place in two rounds. The teams discuss their independent ratings with their assigned partners between each round using evidence from the body of student work to support their ratings.

Rather than using the median value of the Round 2 ratings—as is traditionally done with the body of work standard setting method—we only use scores of those raters who agreed on a given achievement level for the portfolios of work. We decided on this approach because we typically have only 2-3 raters at a table and there is still considerable variability in the quality of the student work portfolios submitted (though we continue to see improvements over time in the quality of evidence submitted). This consensus rating inspires more confidence that the quality of the body of work is sufficient for making a consistent judgment about student performance. We then compare this score (rating) to the teacher judgment survey (TJS) rating used to set standards as both judgments are based on the PACE Achievement Level Descriptors. Because the PACE annual determinations are grounded in the work that students produce throughout the year, this "body of work" analysis provides particularly useful validity evidence to support the PACE innovative assessment system.

In general, BOW analyses across years shows that there is strong agreement for students at Level 3 and 4, but the BOW ratings are more stringent than the TJS ratings for Levels 1 and 2. This finding is consistent with the measurement literature where it is well-documented that the body of work method is more rigorous than other standard setting approaches[12]. Findings across years also show a high degree of exact and adjacent agreement between the BOW ratings and TJS ratings (typically > 90%); however, the strength of this validity evidence would improve with higher exact agreement rates.

Though findings to date for BOW analyses do not provide confirmatory evidence, at this point, to support the validity of PACE annual determinations, however that is likely due to challenges with implementing the BOW method in this context. 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 of mixed quality, which made it difficult to find evidence to support inferences.

The NH DOE and Center for Assessment continue to try out different approaches to support educators and provide training on the purpose and nature of the bodies of evidence they should be collecting throughout the year to support the collection of higher quality BOW samples. Based on the improvement in these samples we have seen over the past several years, we expect to see continued improvement going forward.

[12] See, for example: Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice,* 22*(1),* 22–32.

Section 4: Extensive Comparability Analyses and Evaluation

The PACE system is designed to ensure annual determinations of student proficiency are comparable *within* districts, *among* districts, and *across* PACE and non-PACE districts. The validity of the PACE system includes 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.

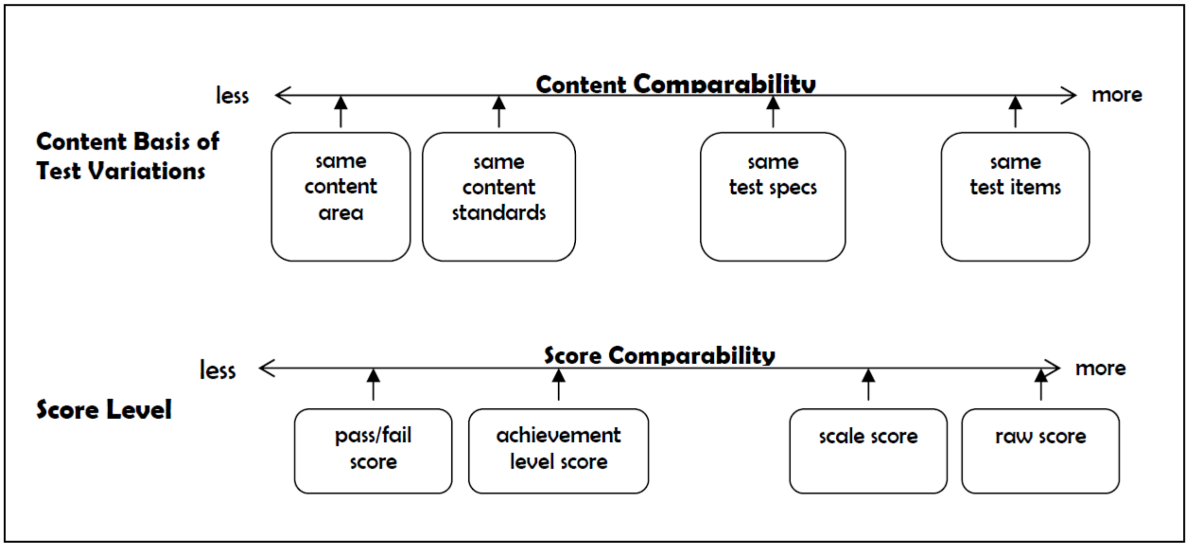
This means that evidence used to support claims of comparability will differ depending on the nature (or grain-size) of the reported scores. For example, supporting claims of raw score interchangeability—the strongest form of comparability—would likely require the administration of a single assessment 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.

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 4, 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 exists at different degrees of granularity.

Figure 4

Comparability Continuum



Note: Figure taken from Winter (2010)[13].

In the PACE system, comparability claims and associated evidence are 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 PACE or non-PACE) he or she could expect to also be deemed proficient.

Comparability at the level of annual determinations is of primary concern for two reasons. First, because NH must incorporate assessment results from the PACE districts into the state accountability system alongside the results generated from the non-PACE districts, the assessment systems must produce results that are comparable enough to support their simultaneous use in the single statewide accountability system. Second, requiring that the assessment systems produce comparable results ensures that the state and districts will not view the innovative assessment and accountability demonstration authority as a way to relax the rigorous expectations established under the current state assessment system. 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.

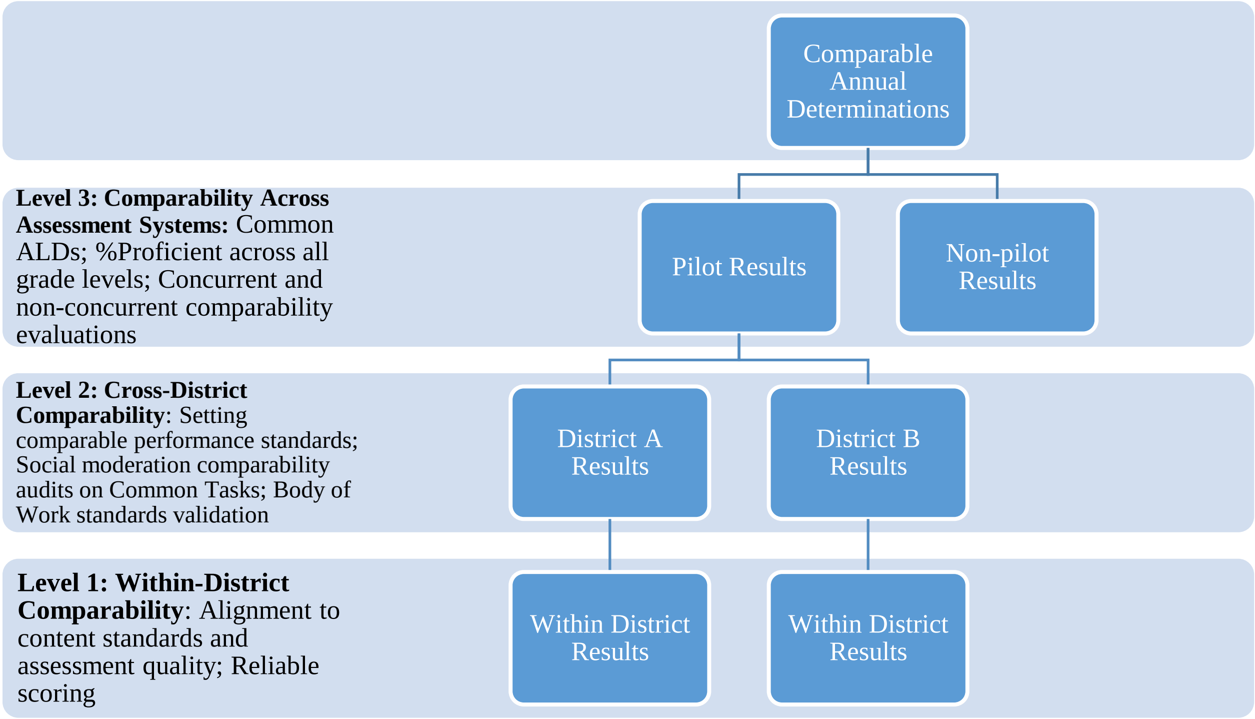
[13] Winter, P. C. (2010). Evaluating the comparability of scores from achievement test variations. Washington, DC: Council of Chief State School Officers, p. 5.

Overview of PACE Comparability Evaluation 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 5 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.

Figure 5

Establishing an Evidence-Base for PACE Comparable Annual Determinations



Level 3: Comparability Across Assessment Systems: Common ALDs; %Proficient across all grade levels; Concurrent and non-concurrent comparability evaluations

Level 2: Cross-District Comparability: Setting comparable performance standards; Social moderation comparability audits on Common Tasks; Body of Work standards validation

Level 1: Within-District Comparability: Alignment to content standards and assessment quality; Reliable scoring

Level 1: 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 of alignment and assessment quality. Evidence regarding alignment and assessment quality comes from the comprehensive local assessment system peer reviews discussed under Section 2.

Evidence of reliable scoring. 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.

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.

Inter-rater reliability estimates. The Center for Assessment externally audits the consistency in scoring each year by asking each district to submit a sample of papers from each PACE Common Task that have been double-blind scored by teachers within district. The collection of double scores is then analyzed using inter-rater reliability methods to estimate within-district scoring consistency. 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. We report rater consistency using percent agreement in two ways. First, we report percent agreement (exact and adjacent) between raters by task and rubric dimension. The target set for rater consistency is a 60% exact agreement rate for each dimension on the PACE Common Tasks. Exact agreement rates that do not meet this target are noted and examined further at the district level.

Second, we report rater consistency by district and subject area using percent exact and adjacent agreement. The target set for rater consistency is 60% exact agreement rate for each district and subject area. Districts with exact agreement rates below this threshold in any given subject area are noted for further review and follow-up conversations discussed in more detail below.

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.

We calculate Kappa estimates by task and rubric dimension for each subject across all districts, as well as by subject area and rubric dimension for each district. Any Kappa estimate lower than moderate agreement (< 0.40) is noted and discussed with districts. Conversations with districts center on ways they can adjust processes or procedures within-district to strengthen the quality of scoring practices and inter-rater reliability.

Generalizability analysis. In the NH PACE assessment and accountability system there could be upwards of thirty local assessments contributing to students' overall achievement estimates for a given grade/subject. 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.

For the first few years of the NH PACE system, the Center for Assessment used electronic grade book competency data provided by districts to examine the generalizability of the individual scores that go into achievement estimates (e.g., summative tests, quizzes, projects, performance tasks) in the PACE grades and subjects.

The variance of assessment (task) scores were partitioned into independent sources of variation due to differences between persons, tasks, and the residual. In these analyses, 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 was 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 is used. 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.

In general, and similar to other analyses from prior research[14], analyses found that one assessment (task) does not account for a large percent of the variance in individual student achievement. The largest variance component in all grade/subject combinations is the residual. 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

[14] Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education*, *7*(4), 323–342. Retrieved from https://doi.org/10.1207/s15324818ame0704_4

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.

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 these analyses, the G coefficient represents the proportion of variability in observed assessment scores attributable to systematic differences in students' competency. Results suggest that the collection of classroom assessments provide for stable estimates of student achievement in a given content domain.

In the D, or "decision" study, we show how increasing the number of assessments included in achievement estimates results in diminishing returns beyond approximately 20 assessments. 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 over the course of the year in a content domain.

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 innovative assessment system. Approximately 15-20 assessments per year in a content area provide for an efficient trade-off while still ensuring a high degree of relative and absolute decision reliability. Most local assessment systems include at least 15 summative assessments per year.

Level 2: Cross-District Comparability in Expectations of Student Performance

There are two main sources of cross-district comparability evidence: (a) setting comparable performance standards, and (b) social moderation comparability audits using the PACE Common Tasks.

Setting comparable performance standards. Comparable performance standards across districts assumes that there are common definitions and understandings of student proficiency shared across districts. The use of common ALDs across districts promotes this comparability, as does the submission of student bodies of work that are then re-rated by teachers from other districts and ratings compared to TJS ratings. See Section 3 for a more detailed discussion of the Body of Work performance standards validation.

Additionally, comparable performance standards across districts assumes that there are the same accommodations are offered to students on PACE common tasks as well as other local summative assessments. These accommodations are specified in the PACE Accommodations Manual and applied to students based on students' Individualized Education Plans (IEPs).

Social moderation comparability audits. In order to account for differences in the relative stringency and leniency in teacher scoring across the PACE districts, the PACE system uses common performance tasks across districts. These PACE common tasks allow us to evaluate the degree of comparability in local scoring[15]. These analyses rest on two foundational assumptions: 1) that patterns in scoring for the common tasks is representative of district relative

[15] Evans, C. M., & Lyons, S. (2017). Comparability in balanced assessment systems for state accountability. *Educational Measurement: Issues and Practice*, *36*(3), 24–34. https://doi.org/http://dx.doi.org/10.1111/emip.12152

stringency or leniency of local scoring represented in end of year competency scores, and 2) the degree of relative stringency or leniency of scoring is consistent within district for a particular grade and subject area.

The social moderation comparability 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 needed due to systematic cross-district differences in scoring, which violates one of the foundational assumptions noted above. 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.

Each summer either asynchronously or during the PACE Summer Institute, teachers and leaders from the PACE districts participate in the social moderation calibration audit. The calibration audit uses a consensus scoring method that involves pairing teachers together, each representing different districts, to score student work samples. The student work samples are gathered for each of the PACE common tasks from the implementing districts. Both judges within each pair are asked to individually score their assigned samples of student work. Working through the work samples one at a time, the judges discuss their individual scores and then agreed on a “consensus score”. If consensus cannot be reached, an expert scorer (who does not have affiliation with any particular district) decides on the appropriate consensus score. There are typically very few cases where an expert scorer is needed. If moderation is needed, it is typically for one rubric dimension.

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

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

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

Findings over the years suggest that it is much more likely that districts have positive mean deviations. This means a systematic overestimation of common task scores by the classroom teachers, or that teachers tend to score their own students more leniently than teachers from other districts. This finding is not unexpected given that classroom teachers may also rely on other information about students to inform scoring and that type of contextual information is not available to outside teachers. If mean deviations are all positive or high it is not necessarily problematic from a comparability perspective, the purpose is to look for differences among the districts in mean deviation.

Results from three-factor analysis of variance conducted over the years also reveals significant 3-way interactions for district by grade by subject combinations. This means unilateral adjustments to any one district's cut scores across the board are not justified. Instead, more nuanced decisions (if any) must be made based on follow-up analyses.

Follow-up analyses are two-fold. First, mean deviations are analyzed by district, subject, and grade. Mean deviations that are ± 0.50-points (on the scale of the rubric) different than the subject and grade level average deviation are noted, if present, as are district, subject and grade combinations with less than 10 students. Small sample sizes confound interpretations due to lack of precision and the associated uncertainty. Second, mean deviations ± 0.50-points with more than 10 students are then examined based on historical trends over time captured in the impact analyses. The impact analyses are discussed in more detail in Section 3 under "Quality Assurance Processes and Procedures." The purpose of the follow-up analyses is to examine the extent to which the district, grade, and subject combinations with larger/smaller relative mean deviations than expected also show unlikely or improbable shifts in proficiency rates given historical trends. Over the years, evidence that supports any cut score adjustment based on the calibration audit has been extremely rare and is dealt with on a case-by-case basis in consultation with the NH DOE.

Level 3: Across Assessment System Comparability of Annual Determinations

The comparability processes and audits that occur at both the within-district level and the cross-district level are all in an effort to support the claim of comparability in the annual determinations. Until the PACE system scales statewide, a major ESSA comparability requirement is that the innovative assessment system results are comparable with the statewide standardized results.

The following five procedures are used to formally promote and evaluate the comparability of the annual determinations of student proficiency across PACE and non-PACE districts: (a) common ALDs shared across assessment systems, (b) common accommodations shared across assessment systems, (c) percent proficient evaluations across assessment systems, (d) concurrent comparability evaluations, and (e) 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.

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 at least these 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 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.

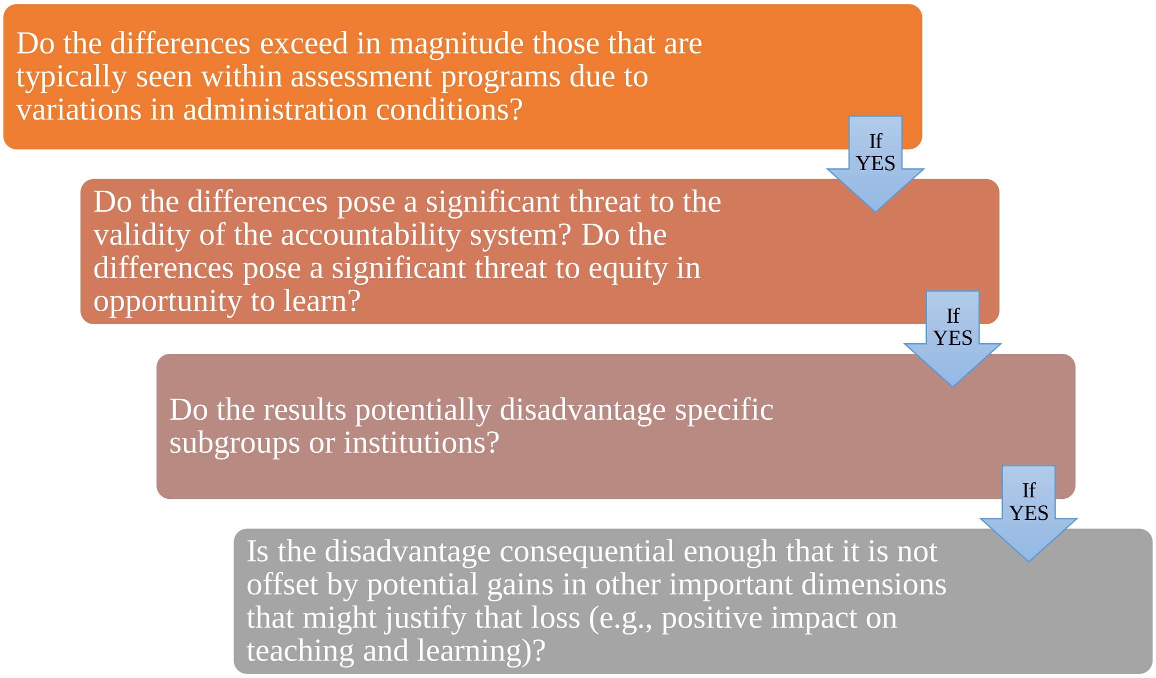
However, it is worthwhile to consider what might be reasonable to expect for the amount of variability in proficiency classifications across the two assessment programs. We argue that a reasonable upper bound for comparability across PACE and non-PACE systems is the degree to which comparability is achieved 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.[16] 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 PACE and non-PACE systems 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 PACE and non-PACE 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 are detected, the state should evaluate the practical implications of those differences for decision making within the accountability system. Figure 6 presents a series of questions that could determine whether or not the levels of comparability seen are appropriate for the intended purposes.

[16] See, for example: 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

Figure 6

Decision Tree for Determining Degree of Comparability Achieved



Do the differences exceed in magnitude those that are typically seen within assessment programs due to variations in administration conditions?

Do the differences pose a significant threat to the validity of the accountability system? Do the differences pose a significant threat to equity in opportunity to learn?

Do the results potentially disadvantage specific subgroups or institutions?

Is the disadvantage consequential enough that it is not offset by potential gains in other important dimensions that might justify that loss (e.g., positive impact on teaching and learning)?

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 PACE system are arguably the most important for demonstrating that results across PACE and non-PACE 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. Additionally, if the evidence for comparability across the two systems of assessment is strong, comparability need not be re-evaluated every year. Once it has been established, the state should provide evidence that the processes and procedures in place are sufficient for replicating the program across years.

Common Achievement Level Descriptors (ALDs) shared across assessment systems. 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., cut scores) for an assessment program. One of the goals of the PACE system 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 NH SAS ALDs as the basis for the NH PACE ALDs and to ensure careful alignment between the ALDs throughout the design and review process described below.

The NH PACE ALDs were revised during the 2018-19 school year through a comprehensive process to ensure alignment with the new Statewide assessment system—NH SAS—and aide PACE teachers in making accurate and reliable judgments about student proficiency on the Teacher Judgment Survey at the end of the year. When PACE was originally implemented in the 2014-15 school year, New Hampshire was administering Smarter Balanced. Once the NH SAS published ALDs, the NH DOE and its technical partners set about reviewing and revising the PACE ALDs.

The PACE ALD revision process began in January 2019 when PACE content leads were invited to participate in the first round of PACE ALD revisions, which occurred on February 5, 2019. PACE content leads are teachers from participating PACE districts with demonstrated assessment literacy expertise such that they lead teams of teacher task developers from across PACE districts to design and pilot PACE Common Tasks each year (see Section 1: Content Leads for more information).

After the first round of revisions to the PACE ALDs, the Center for Assessment provided drafts to the entire group of PACE content leads at the next content lead meeting. The purpose was to solicit and gather feedback about the structure of the revised PACE ALDs and the extent to which summary ALDs better served the purpose of supporting accurate and reliable teacher judgments of student proficiency at the end of the year. The PACE content leads were overwhelmingly positive about the benefits of the new structure, format, and content of the PACE ALDs.

A group of PACE content leads met during the PACE Summer Institute 2019 and asked to finish the revisions. Center for Assessment staff facilitated this two-day revision event.
The Center for Assessment then completed a thorough and detailed review of the revised PACE ALDs for each subject area and then across subject areas to check for alignment to the NH SAS ALDs, consistency of format and language, and quality of the summary descriptions of student achievement at each of the four performance levels. Finalized versions of the PACE ALDs were then curated.

The NH PACE ALDs include Grades 3-8 ELA and Math and Grades 5 and 8 Science. Some of these grade/subject area combinations are non-PACE accountability grades, but are used in order to produce non-reported PACE annual determinations to compare student-level PACE results with the student-level results on the NH SAS. More detail about those analyses is below under the concurrent and non-concurrent validity evaluation.

Common accommodations shared across assessment systems. The second way comparability of annual determinations across assessment systems in the State is promoted is through common accommodations. Again, given the switch to the NH SAS, the NH DOE and its technical consultants revised the PACE Accommodations Guide so that it was consistent with the NH SAS Accommodations Guide. The PACE Accommodations Guide is identical to the accommodations on the statewide academic assessment and both are based on principles of Universal Design for Learning. Participating PACE districts and schools agree to implement the allowable accommodations on their local and common assessments. This coherence increases the comparability of results across assessment systems for students with disabilities and English learners.

Percent proficient evaluations assessment systems. Detailed analyses that compares the percent proficient or above by subject across the PACE and NH SAS system for the PACE districts is included in the impact analyses described in more detail under Section 3: Quality Assurance Processed and Procedures.

Concurrent comparability evaluations. The concurrent analysis calculates PACE annual determinations for the grades that are currently taking NH SAS (Gr 3 ELA, Gr 4 Math, and Grade 8 ELA/math) and compares the results. PACE annual determinations are not reported for these subjects and grades and no common performance task was administered, however, the same procedure for producing PACE annual determinations was used in these grade levels as for the PACE reported annual determinations.

Annual concurrent comparability evaluations compare (a) the overall percent of students scoring proficient or above by subject and grade level between the two assessment systems, (b) achievement level frequency counts and percentages for the two sets of annual determinations, (c) cross tabulation of achievement levels for the two sets of annual determinations by subject and grade level, (d) aggregates of the crosstabs showing the percentage of exact agreement, adjacent agreement and percentage of exact or adjacent agreement by grade and subject area, and (e) classification accuracy across the assessment systems for all student groups and by waiver-reported subgroups . "Classification accuracy" refers to the percentage of students who received the same proficiency classification (i.e., 'proficient'=Yes or 'not proficient'=No) across the two years.

Findings over the years suggest a high degree of comparability of the students scoring at the reported achievement levels. Importantly, there is typically about 90% exact or adjacent agreement on achievement levels for all grades and subjects between the two assessment systems. The classification accuracy tends to be lower—around 75% to 80%. While this agreement is high, there are a variety of reasons why there may be legitimate differences in the results produced by the different assessment systems. First, the degree of agreement is limited by the reliability of each assessment system. In other words, an assessment cannot correlate more with another assessment than it can with itself (i.e., reliability). Therefore, because both PACE and NH SAS are not perfectly reliable, we may be approaching the upper bound of the relationship between the two assessment systems. Additionally, New Hampshire's PACE assessment system is in place to measure the state-defined learning targets differently than they are measured in the statewide assessment system. The purpose is to measure the standards more deeply and authentically through performance-based assessments. Additionally, the PACE assessment system is intended to measure the set of standards more completely (e.g., including the listening and speaking standards). The demonstrated 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.

Non-concurrent comparability evaluations. The non-concurrent analysis compares performance for the same students on the two assessment systems across years. The Center for Assessment conducts two non-concurrent comparability evaluations.

Non-concurrent analysis #1. The first analysis compares last year's performance on NH SAS in grade 3 ELA and grade 4 math with this year's performance on PACE for students in grade 4 ELA and grade 5 math. Only students with a NH SAS achievement level in 2018 and a PACE achievement level in 2019 are used for these analyses.

Annual analyses investigate (a) the percent proficient or above for the matched cohort of students across years, (b) achievement levels with frequency counts and percentages across the two assessment systems by grade level and subject area, (c) cross tabulation of achievement levels from SAS to PACE by grade level and subject area, (d) aggregates of the cross tabs showing the percentage of exact, adjacent, or exact plus adjacent agreement by grade and subject area across the assessment systems, and (e) classification consistency tables (2 x 2 tables) showing the percentage of students who receive the same proficiency classification (proficient or not proficient) across the two years.

Results across years show that, in general, PACE has fewer students at Levels 1 and 4 than NH SAS, which is designed to more evenly spread students across the distribution of performance levels. Importantly, while there is variation across the two assessment programs over two years, the degree of agreement is high across years (> 90% exact plus adjacent agreement). The strength of the correlations between the two assessment programs across years is typically moderate (r = ~ 0.60). The strength of the correlations is quite high given the intentional differences in design and purpose. Also, these analyses assume that students did not change their performance levels across years when, in fact, we know that not to be true.

For the final part of this analysis, classification accuracy may be a misnomer since students can and do legitimately change in their classifications across years. In fact, schools are purposefully trying to improve the performance of students across years. We expect and do see evidence that from annual analysis that students either stay proficient/non proficient or move from non-proficient to proficient from one year to the next.

Non-concurrent analysis #2. The second non-concurrent validity analysis compares last year's performance on PACE in grade 3 math, grade 7 ELA/math with this year's performance on NH SAS for students in grade 4 math and grade 8 ELA/math. Only students with a PACE achievement level in the previous year and a NH SAS achievement level in the current year are used for these analyses.

Annual analyses investigate (a) the percent proficient or above for the matched cohort of students across years, (b) achievement levels with frequency counts and percentages across the two assessment systems by grade level and subject area, (c) cross tabulation of achievement levels from PACE to SAS by grade level and subject area, (d) aggregates of the cross tabs showing the percentage of exact, adjacent, or exact plus adjacent agreement by grade and subject area across the assessment systems, and (e) classification consistency tables (2 x 2 tables) showing the percentage of students who receive the same proficiency classification (proficient or not proficient) across the two years.

Findings from over the years indicate that PACE is at least as rigorous as NH SAS but, in general, PACE has fewer students at Levels 1 and 4 than NH SAS, which is designed to more evenly spread students across the distribution of performance levels. The degree of exact plus adjacent agreement is high across years ranging (> 90%). The correlations between the two assessment programs across years are similar to the non-concurrent analysis #1 (*r* = ~ 0.60) for ELA and math. As mentioned previously, given the fact that no assessment is likely to correlate more highly with a different assessment than with itself, the strength of the correlations between PACE and SAS 2019 are remarkably high. The classification accuracies across years are about the same as the classification accuracies observed for the concurrent and other non-concurrent year comparisons (~ 70%).

Section 5: Valid Annual Determinations of Student Proficiency

The purpose of this technical quality manual was to present comprehensive and detailed evidentiary argument in support of the validity of the PACE system and associated annual determinations of student proficiency in PACE grades/subjects. 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.

The intended uses and interpretations of PACE assessment system results are supported based on all the evidence described in this manual and presented in the companion annual results with respect to the technical quality of the PACE system. Specifically, evidence of **alignment** to the breadth and depth of the state’s academic content standards, **fairness** of the assessments for all students, **reliability** of scoring and generalizability of the inferences about students’ knowledge and skills, and **comparability** of the assessment results for students within districts, across districts, and across assessment systems in the state. The analyses and documentation presented in this report supports the validity of inferences from PACE assessments for the intended uses in the PACE system.

Reporting

The PACE system provides student-level annual determinations of student proficiency based on the state’s academic content standards for the grade in which the student is enrolled. PACE results are consistent with the NH SAS assessment system as students receive an achievement level 1-4 based upon their achievement over the course of the year. Levels 1-2 identify which students are not making sufficient progress toward, and attaining, grade-level proficiency on such standards. Level 3 is considered proficient and Level 4 is above proficient.

The NH DOE has set up parent access to PACE results in the same format and manner as the NH SAS results. Specifically, the NH DOE has created an electronic login parents can use to access their student's PACE results and student-level reports have been produced and are going to be sent out to districts to provide to parents. The PACE student reports and NH SAS reports have a uniform format except that a scale score is provided on the NH SAS student report.

PACE system results are produced in such a way that they can be disaggregated within the State, as well as each LEA and school by all subgroups identified in federal regulations, except in such cases in which the number of students in a subgroup is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student.

Use of Results in State Accountability System

New Hampshire's Accountability Task Force—the stakeholder group responsible for the design of the approved December 2017 ESSA plan—was intently interested on ensuring that PACE continues to play a prominent role in the State's strategic plan. This focus is represented throughout each part of New Hampshire's state plan and is especially true for accountability, where the state plan ensures that PACE schools can be effectively and comparably included in all aspects of the system including the state's long-term goals for academic achievement, the academic achievement indicator, school identification for targeted or comprehensive support and improvement, and reporting on State and LEA report cards.

The PACE innovative assessment system has been designed to be comparable to the statewide system of assessments for the express purpose of use within the state accountability system. Because the annual determinations are designed to be comparable, the determinations can be used to serve the same purposes within the accountability system. This means that a school's participation in PACE does not systematically influence a school's score on the achievement indicator, and likewise the overall summative determination within the accountability system.