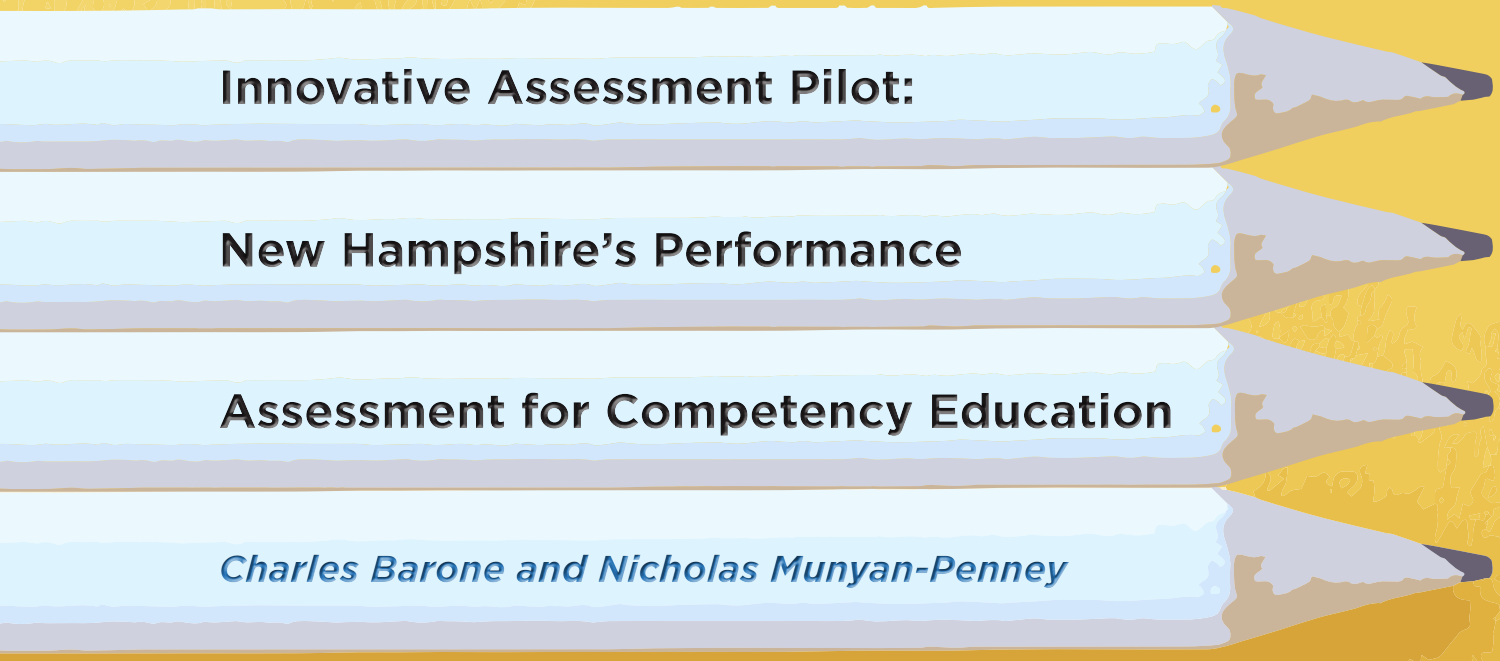


Innovative Assessment Pilot: New Hampshire's Performance Assessment for Competency Education

Charles Barone and Nicholas Munyan-Penney

JANUARY 2020

ERN

EXECUTIVE SUMMARY

In 2014, New Hampshire was the first state to be granted an innovative assessment waiver from the U.S. Department of Education, and in 2018 was the second state approved to participate in the Innovative Assessment Demonstration Authority (IADA) pilot program under the Every Student Succeeds Act (ESSA). Districts participating in New Hampshire's Performance Assessment for Competency Education (PACE) are free of federal requirements that the same summative assessments be administered in math and English Language Arts (ELA) in grades 3-8 and that all students in the state, with some exceptions,[1] participate in the same statewide assessment.

New Hampshire's expressed goal for PACE is to "structure learning and assessment opportunities that allow students to gain and demonstrate their knowledge and skills at a depth of understanding that will transfer beyond K-12 education to success in careers and college."

Unlike many other assessment innovations that tinker around the edges, PACE departs radically from systems that rely on standardized, statewide annual assessments.

The PACE system has a lot of moving parts including performance tasks in ELA, math, and science intended to assess the full depth and breadth of the state's academic standards. The assessment system for the 11 districts participating has the following components:

- **PACE.** Innovative assessment system that determines student proficiency by combining scores from:
  - **Locally Administered Performance Tasks.** Districts develop their own standards-aligned tasks to determine student mastery;
  - **Common Performance Tasks.** Students complete a PACE Common Performance Task, which calibrates scoring and is intended to provide some degree of comparability across districts.
- **New Hampshire Statewide Assessment System.** In grades 3-8, students participate in New Hampshire's traditional Statewide Assessment System (NH SAS) in a single subject (math, ELA, or science) each year;[2]
- **SAT.** All high school juniors also take the SAT in lieu of statewide summative assessments in math and ELA.

Unlike many other assessment innovations that tinker around the edges, PACE departs radically from systems that rely on standardized, statewide annual assessments. This poses a number of opportunities to improve instruction and boost outcomes, as well as some significant risks including a lack of comparability of results between participating districts.

Regardless of how successfully NH implements PACE across the state, we caution against it being a model for other states considering joining IADA. New Hampshire is relatively unique compared to states across the country: it is less demographically diverse, small in size, has few students, and has no large districts. As a result, a complex system such as PACE will likely face additional hurdles if adopted in other states.

OVERVIEW

New Hampshire launched its Performance Assessment for Competency Education (PACE) in 2014 with a waiver from the US Department of Education under the No Child Left Behind Act as an effort to complement the state's move towards competency-based instruction and grading. The pilot was extended in 2018 under the Innovative Assessment Demonstration Authority (IADA) contained in the "Every Student Succeeds" Act (ESSA),

which now requires NHDOE to scale PACE statewide and integrate it into the state's accountability structures. As of the 2018-19 school year, 11 districts and 34 schools (5% of schools statewide) were implementing the PACE system in at least one subject area. Many more are exploring the program informally through professional development opportunities offered by the state.

Opportunities. If PACE meets the state's ambitious goals, it may:

- Complement the knowledge and skills assessed under the original system by requiring students to synthesize information in multi-step problems;
- More readily reflect what students are actually learning in the classroom and what knowledge and skills they still need to master;
- Allow students to engage with material and demonstrate learning in multiple ways, particularly important for non-traditional learners, including students with disabilities (SWD) and English Learners (EL);
- Provide more actionable and timely information through flexible administration of multiple assessments throughout the year which allow educators to personalize instruction; and,
- Reduce classroom disruptions and potential student anxiety that take place when statewide assessments are concentrated over a limited, high-intensity time period.

New Hampshire's PACE system presents an opportunity to provide more actionable and timely information through flexible administration of multiple assessments throughout the year which allow educators to personalize instruction.

Risks. However, PACE poses a number of challenges and inherent risks:

- The PACE system could strain state and local capacity. Educators need extensive training in assessment development, administration, scoring, and application of results in the classroom;[3]
- The processes and results of the PACE system are more complicated and less transparent than most states' accountability systems that are centered around public reporting of summative assessments and other common indices.
- The fact that assessments differ from grade-to-grade hinders the reliable measurement of student growth and progress as compared to traditional systems that make year-to-year, apples-to-apples comparisons.[4]
- Tasks may vary in what they measure and their level of rigor. "Common tasks" vary across districts which makes inter-district comparisons difficult and may hold students across the state accountable to different standards;
- Current pilot districts don't reflect racial/ethnic demographics of the state, further limiting comparability efforts and raising questions about whether the pilot is valid with regard to students from historically disadvantaged groups,

Tasks may vary in what they measure and their level of rigor which makes inter-district comparisons difficult and may hold students across the state accountable to different standards.

• Differences in local assessments and small numbers of students in some districts make statistical analysis more complex and less reliable; and,

• Until studies are conducted to establish reliability and validity for all measures and to link proficiency on PACE to success on other tests like NH SAS and SAT, or positive postsecondary outcomes, it cannot be known whether PACE is an accurate indicator of college and career readiness.

Despite stating that PACE can effectively be used in all aspects of the state's accountability system, including identifying schools for targeted and comprehensive support, NHDOE hasn't explained how or when annual scores from PACE will be integrated into the state's existing accountability structure for measuring growth.

Demographic data for participating districts (see *Appendix A*) show that while there is wide variation in the percentage of students from low-income families, every district currently in PACE has a lower percentage of Latinx students than the state, and 11 of 12 of the districts have a higher percentage of White students. This raises concerns about how well the pilot serves students from historically disadvantaged groups. Not having comparable data and/or not having diverse enough districts in the pilot could mean that results from the pilot are of limited value in informing expansion and statewide implementation.

The PACE system has a lot of moving parts including performance tasks in ELA, math, and science intended to assess the full depth and breadth of the state's academic standards. The assessment system for the 11 districts participating has the following components:

• **PACE.** Innovative assessment system that determines student proficiency by combining scores from:

  • **Locally Administered Performance Tasks.** Districts develop their own standards-aligned tasks to determine student mastery;

  • **Common Performance Tasks.** Students complete a PACE Common Performance Task, which calibrates scoring and is intended to provide some degree of comparability across districts.

• **New Hampshire Statewide Assessment System.** In grades 3-8, students participate in New Hampshire's traditional Statewide Assessment System (NH SAS) in a single subject (math, ELA, or science) each year;[5]

• **SAT.** All high school juniors also take the SAT in lieu of statewide summative assessments in math and ELA.

Table 1: PACE Districts Assessments Overview

Assessment	Grades 3 – 8	High School	Frequency
PACE Local Tasks	X	X	6-25x per year
PACE Common Task	X	X	1x per year
NH SAS	X	X (Science)	1x per year
SAT		X (Math & ELA)	1x per year

Student annual assessment scores (a range of 1-4) are determined through a combination of locally developed performance assessments—administered throughout the year to show mastery of individual competencies—and one or more of the PACE Common Tasks (districts are not required to administer the same one). IADA schools conduct between six and 25 assessments in each subject throughout the year, with median of 18 local summative assessments each year. NHDOE argues that this score determination avoids the

pitfalls of assessments that only measure a single moment in time and allows students to "show what they know" while also producing actionable data throughout the year.

However, despite stating that PACE can effectively be used in all aspects of the state's accountability system, including identifying schools for targeted and comprehensive support, NHDOE hasn't explained how or when annual scores from PACE will be integrated into the state's existing accountability structure for measuring growth. This will also require linking scores from NH SAS and PACE as well as to Smarter Balanced, New Hampshire's state assessment through 2017.

Table 2: Grade Level Assessments by Subject

Grade	ELA	Math	Science
3	NH SAS	PACE	Local Assessments (PACE in development)
4	PACE	NH SAS	Local Assessments (PACE in development)
5	PACE	PACE	NH SAS
6	PACE	PACE	Local Assessments (PACE in development)
7	PACE	PACE	Local Assessments (PACE in development)
8	NH SAS	NH SAS	PACE
11	SAT & PACE	SAT & PACE	NH SAS & PACE

Accessibility for students with disabilities and English learners

NH DOE has worked to make PACE accessible to all students, including SWD and EL, in both the design and administration of performance tasks, with the state making it clear in its IADA application that one of the main benefits of performance task assessments are their ability to provide "a more coherent educational experience for students with disabilities," compared to traditional assessments such as NH SAS or SAT.

All performance tasks are designed using Universal Design for Learning (UDL), which is intended to make each assessment assessable to as many students as possible and is explicitly connected to SWD and EL in the Task Development Framework. Additionally, the PACE Accommodation Standards include a comprehensive list of research-based design considerations specifically for EL, which are referred to in the development framework. The assessment review rubric used for local assessments and the Common Task also contains a series of items related to accessibility.

Accommodations guidelines are adapted from Smarter Balanced assessments and are used for both PACE and NH SAS. However, a Human Resources Research Organization (HumRRO) evaluation of the PACE system notes that NHDOE needs to provide more training in both UDL and accommodations to ensure accessibility across all participating districts; currently this training is limited only to content leads.

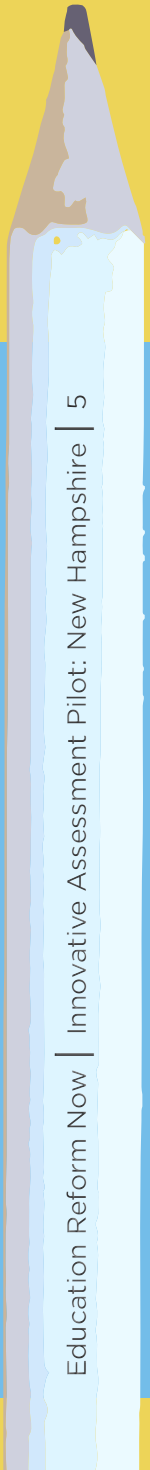
RELIABILITY AND VALIDITY

In order for PACE to function within New Hampshire's accountability system, the state needs to establish evidence that the assessments are reliable and valid measures of

student achievement. Reliability and validity are important to ensure that assessments actually measure what students know and are able to do. The tables in *Appendix B* summarize what we know about the reliability and validity of the PACE assessment system.

NHDOE has released a number of internal and external studies that begin to address these concerns (see *Appendix B*), but there is still much more work to be done as PACE develops and scales statewide. For example, without clear proof of inter-rater reliability (i.e., results should be the same no matter who scores the test and when they score it), students' scores from different districts—both inside and outside of the PACE system—cannot be compared to one another. More importantly, without clear evidence that students deemed proficient under PACE have strong post-secondary outcomes, such as college enrollment and completion, the PACE system cannot be assumed to be valid for its intended purposes. Similar concerns could be directed towards NH SAS assessment, for which no studies of validity and reliability are publicly available. However, unlike PACE assessments, NH SAS was developed by the American Institutes for Research, which has a strong reputation and whose assessments are used by states around the country.

While innovation is to be valued and encouraged, we need to be mindful of the reasons that statewide standards and assessment systems were implemented in the first place lest, in years ahead, we see an ever-accelerating race to the bottom.



WHY COMPARABILITY MATTERS

One key theme throughout ESSA is that standards and assessments must be the same, statewide, for all students. The words "all," "same," and "statewide," as applied to standards, assessments, schools, and students appear, consistently, multiple times across what is really the heart of the entire 400-page law. As words go, "all," "same," and "statewide" are about as precise as it gets and, generally speaking, these provisions have a 25 year history under various iterations of the Elementary and Secondary Education Act.

Local assessment systems have broad policy and political appeal but two key reasons ESSA, outside the innovation pilot, requires that state assessments be the same for all students are:

1) Such measures cannot be compared against one another; and

2) Students in different local education agencies could be held to very different standards, even though they would ultimately be applying to the same colleges and competing for the same jobs.

Despite the best intentions, there are immense political and economic pressures at the local level to cast schools in the best light possible. If we abandon statewide assessment systems, poor and minority students, students with disabilities, and English Learners—who historically, prior to advent of the standards and assessment movement, were held to lower standards—might return to a time when they repeatedly were told they were doing fine, only to graduate from high school and discover they didn't have the skills needed to succeed in college and the workplace. Moreover, resources that are now allocated on the basis of accountability systems geared to a single and apples-to-apples comparable set of state tests—those, for example, for after-school and summer programs, tutoring, teacher training, and new curricula—might be misdirected away from areas that actually need them most, because each district or school would then be measured by different standards and different yardsticks.

While innovation is to be valued and encouraged, we need to be mindful of the reasons that statewide standards and assessment systems were implemented in the first place lest, in years ahead, we see an ever-accelerating race to the bottom.

NH DOE has created various safeguards and conducted multiple evaluations to determine validity through the comparability between PACE and the state's main assessment system. Both use the same Achievement Level Descriptors and use the same accommodations for students with disabilities and English language learners. It also conducts score comparisons between the two systems, both internal and external, which NHDOE states support the validity of the PACE system. Existing comparability studies show around 70% agreement in student proficiency determinations between PACE and Smarter Balanced (SBAC). However, since NH SAS just replaced SBAC in 2018 new comparison studies have not yet been released. Importantly, NHDOE cautions against expecting or requiring PACE and NH SAS to "tell the same story about student achievement," arguing that such an expectation would effectively stifle innovation and undermine the overarching goals of PACE.

Comparability: PACE Common Tasks

The most rigorous processes for reliability exist around comparability across PACE districts through the development and administration of PACE Common Tasks. Common Tasks are developed by cross-district teams of teachers and leaders in a nine-step process that includes two external reviews and task pilots. Teachers in PACE districts also participate in score calibration audits with the goal of creating consistent, reliable grading practices. During calibration audits, teachers work in cross-district pairs to determine "consensus scores" for a sample of student work. These consensus scores are then compared to the actual scores these students received from their home district. Any district-grade-subject combinations with large, statistically significant differences are then compared to score distributions in the district for other grades in the same subject. If this comparison shows statistically significant differences in the same direction, adjustments to district cut scores are made.

Without clear evidence that students deemed proficient under PACE have strong postsecondary outcomes, such as college enrollment and completion, the PACE system cannot be assumed to be valid for its intended purposes.

However, despite the name of these assessments, PACE districts are not required to administer the same Common Task in a given grade-subject. Instead they can choose from any of a number of assessments in a "bank" of tasks. NHDOE argues that it is neither "feasible [n]or necessary" for all districts to use a single task, given the rigorous development process. And the Common Task only accounts for an undefined small portion of student annual scores, with local performance tasks contributing a vast majority of achievement data. Additionally, NHDOE does not make sample Common Tasks available on its website—unlike NH SAS, which has publicly available practice items—making it impossible for external stakeholders to examine and assess these tasks.[6]

According to NHDOE, the development of Common Tasks and cross-district scoring audits serve the dual purpose of modeling practices that local districts should use when developing, administering, and scoring local performance tasks. Yet, only a small subset of local performance tasks is subjected to cross-district review to ensure quality. Generalizability studies based on local assessment tasks have produced strong results. However, these only assess generalizability within a district, and all other reliability studies are focused solely on Common Tasks. A HumRRO evaluation of the PACE system flagged the lack of guardrails in place to ensure the rigor of local assessments. As a result, NHDOE has contracted with Stanford University to conduct external reviews of local performance tasks, though no details or timeline of these reviews have been released.

The state also plans to greatly expand the Common Task bank, so districts can use these as local assessments. Given that local assessments determine a large proportion of a student's annual score, these external reviews are critical to ensure the rigor and consistency of local performance tasks. But despite hiring an external reviewer, NHDOE downplays the need to ensure the quality of each individual performance task, stating that the PACE system score is "greater than the sum of the parts."

SCALING UP

As a part of the innovation assessment pilot, NH is committed to scaling up the PACE system to all districts across the state. However, the state is taking a unique approach to this process because of NH's strong tradition of local control. The state is not instituting a top-down mandate requiring that all districts adopt PACE. Instead, they are using a "social movement" method to encourage local adoption, spreading the word about the assessments through blogs, talks, and conferences, communicated through staff in participating districts as well as outside partners. Additionally, the state has created a tiered adoption system and opened PACE professional development opportunities to non-participating districts to support the building of local capacity and buy-in. Districts can enter the PACE system in a single grade or subject, or even a single grade-subject combination before fully implementing PACE. Additionally, NHDOE has made no indication that it plans to phase out its traditional assessment system (NH SAS) once PACE has scaled statewide.

The HumRRO evaluation of PACE states that this approach to scaling up the PACE program is resulting in very authentic engagement in the process and strong implementation. However, the evaluation also notes that early adoption has occurred in high-motivation, high-capacity schools and districts, and that adoption will likely become more challenging as districts with lower capacity join PACE. While NHDOE expresses confidence that the entire state will be PACE participants within the seven-year timeline of the IADA pilot (including a two-year extension), the lack of a state mandate and the likelihood of lower capacity among late-adopters make this timeline seem rather ambitious. It's important that NHDOE maintain its commitment to quality and strong development of local capacity as PACE continues to grow.

ENDNOTES

1 ESSA allows an alternate assessment for students with the most significant cognitive disabilities. The law and accompanying regulations cap the use of alternate assessments at 1% of all students statewide although a number of states have applied for and received waivers of the 1% cap.

2 In non-participating districts, students take NH SAS in math and ELA in grades 3-8 and for science in grades 5, 8 and 11.

3 Thus far, the state is doing a commendable job with this, working to develop grassroots support of PACE, hosting cross-district trainings, and providing additional support for districts that need it. But pressure to reach the goal of statewide adoption under the Innovative Assessment Demonstration Authority (IADA) has the potential to undermine these efforts, so the state must remain diligent even if this means extending their initial timeline.

4 While New Hampshire Department of Education (NHDOE) is making progress on these fronts and the state clearly believes the potential benefits of their innovative assessment outweigh the risks, using PACE for accountability now may be putting the cart (far) before the horse.

5 In non-participating districts, students take NH SAS in math and ELA in grades 3-8 and for science in grades 5, 8 and 11.

6 Multiple requests to NHDOE officials to obtain a sample Common Task went unanswered.

7 FRPL, Race/Ethnicity, ELL enrollment: https://www.education.nh.gov/data/attendance.htm; SWD enrollment: https://www.education.nh.gov/instruction/special_ed/data_profiles/index.htm

APPENDIX A

New Hampshire IADA School Districts—Demographic Information[7]

District Name	% FRPL	% White	% Black	% Latinx	% ELL	% SWD
1. SAU 8-Concord	37%	79%	10%	3%	9%	16%
2. SAU 9-Conway	36%	93%	1%	2%	1%	12%
3. SAU 14-Epping	24%	93%	1%	3%	**	19%
4. SAU 17-Sanborn	14%	92%	1%	5%	1%	17%
5. SAU 23-Haverhill Cooperative	32%	96%	0%	2%	**	16%
6. SAU 30-Laconia	56%	89%	2%	5%	2%	17%
7. SAU 35-Bethlehem	36%	88%	1%	5%	**	9%
8. SAU 39-Amherst	5%	92%	1%	3%	1%	11%
9. SAU 43-Newport	53%	97%	0%	1%	2%	19%
10. SAU 54-Rochester	42%	89%	1%	4%	1%	18%
11. SAU 77-Monroe	25%	97%	1%	1%	0%	24%
12. Seacoast Charter School	9%	94%	0%	2%	**	15%
Statewide	**27%**	**85%**	**2%**	**7%**	**3%**	**16%**

***Fewer than 10 students, redacted by NHDOE.*

APPENDIX B

Reliability

Type	Explanation	Evidence from PACE	Implications
Reliability of student performance	One student should be able to take a test on Monday and then again on Tuesday and get very similar results each day.	There are currently no reports that directly address this issue, though generalizability reports (below) indicate student performance is consistent across different tasks throughout the year.	In theory, PACE should be strong in this area given that performance tasks are designed to show what students really know and are conducted throughout the year to coincide with instruction.
Inter-rater reliability	The results should be the same no matter who scores the test and when they score it.	A 2017 analysis of inter-rater reliability in nine LEAs found that in most LEAs teachers' scores of student work is exactly the same 65% or more, but in two districts agreement was only about 50% for all subjects. Adding adjacent scores (one teacher scores a 3 and the other 4), brings agreement to above 90% in all districts and subjects. The overall results meet existing statistical standards, though some individual districts fall below this threshold.	Many districts are proving to have strong consistency between assessment scorers and NHDOE plans to work with struggling districts improve their scoring calibration practices. However, inconsistent implementation of scoring procedures will likely be a recurring problem as PACE continues to scale up.
		Calibration analyses in 2016 and 2017 reveal no systematic differences in scores across districts.	These results are promising for PACE reliability. Plus, if systematic differences are found in a given district, PACE plans to adjust cut scores to compensate, improving comparability across districts. However, these studies ignore the potential for vastly different assessment quality between districts.

		After complaints from districts about proficiency rates under PACE, an analysis of the cut score determination process revealed systematic inflating and deflating of proficiency rates for a number of districts. As a result, PACE reverted to a system of producing cut scores for each district-subject-grade combination, which have small sample sizes.	Creating cut scores for every district-subject-grade combination will only become more arduous as PACE scales statewide, and small districts and schools across the state will limit the reliability of these estimates, while statewide cut scores have been proven to artificially shrink the range of proficiency rates across participating districts.
Reliability between different forms of the same test	Different forms of a test have slightly different questions in a slightly different order. However, the content and difficulty level are the same and a student should perform similarly on both tests.	Generalizability studies in 2016 and 2017 found that individual performance tasks are reliable estimates of student achievement of all tasks. Reliability estimates reach the ideal of 90% or greater when students complete at least 15 performance tasks.	Students participating in PACE score similarly on different tasks in the same subject help prove reliability. However, the state likely should require districts to administer 15-20 performance tasks in order to ensure the strength of this type of reliability.

Validity

Type	Explanation	Evidence from PACE	Implications
Construct Validity	The adherence of a measure to existing theory and knowledge of the concept being measured.	All PACE LEAs submit assessment maps that demonstrate the coverage of and alignment to standards, and local assessment reviews validate these maps.	While local assessment reviews help improve validity, these reviews are only for a small sample of local tasks despite a less rigorous development process than PACE Common Tasks.
		PACE Common Tasks are developed through a nine-step process including cross-district collaborations and multiple external reviews to ensure task quality.	PACE Common Task development is a strong component of the PACE system, however the system, perhaps naively, assumes that this process will ensure high-quality local performance tasks.

Content Validity	The extent to which the measurement covers all aspects of the concept being measured.	NHDOE argues that PACE has much stronger content validity than NH SAS or other standardized tests because it is able to better assess the depth of the standards through extended, complex performance tasks and the breadth of the standards by testing throughout the year.	The ability of PACE to assess students more deeply and authentically highlights the promise of the state's innovations. Yet, as with construct validity, it hinges on the quality of local assessments.
Criterion Validity	The extent to which the result of a measure corresponds to other valid measures of the same concept.	Both concurrent and non-concurrent comparisons to SBAC show agreement in proficiency of approximately 70% in all subject-grade combinations. However, PACE tends to over-identify students in achievement levels 2&3 and under-identify students in achievement levels 1&4, compared to SBAC. Additionally, no comparison studies have been released since NH SAS replaced SBAC at the state's main assessment.	Earlier studies show strong comparability to SBAC when it comes to overall proficiency rates, but NHDOE should work with districts to improve score calibration procedures to produce a distribution of achievement more comparable to SBAC. More importantly, NHDOE needs to provide substantially more evidence that students deemed proficient under PACE are likely to score well on NH SAS and the SAT in addition to connections to postsecondary outcomes, before PACE can reliably be used for accountability purposes.