Innovative Assessment and Accountability Systems that Support Continuous Improvement under ESSA: Practical Considerations and Early Research

Carla Evans
Center for Assessment
Andresse St. Rose
Center for Collaborative Education
Paul Leather
Center for Innovation in Education

CCSSO 2018 National Conference on Student Assessment
June 29, 2018
Setting the Context

• ESSA allows up to 7 states (or groups of states) to apply for flexibility under Section 1204: Innovative Assessment and Accountability Demonstration Authority.

• Broadly, this authority allows states to pilot an innovative assessment system in a subset of schools for up to seven years, as the state scales the system statewide.
Section 1204

• The application was due at the beginning of April 2018 and only three states applied in this first round: New Hampshire, Louisiana, and Puerto Rico.

• Other states were interested in applying, but decided not to apply for many reasons, including the regulations are not necessarily very flexible.
(b) **Innovative assessment system.** A demonstration that the innovative assessment system does or will--

1. Meet the requirements of section 1111(b)(2)(B) of the Act, except that an innovative assessment--

   (i) **Need not be the same assessment** administered to all public elementary and secondary school students in the State during the demonstration authority period described in 34 CFR 200.104(b)(2) or extension period described in 34 CFR 200.108 and prior to statewide use consistent with 34 CFR 200.107, if the innovative assessment system will be administered initially to all students in participating schools within a participating LEA, provided that the statewide academic assessments under 34 CFR 200.2(a)(1) and section 1111(b)(2) of the Act are administered to all students in any non-participating LEA or any non-participating school within a participating LEA; and

   (ii) **Need not be administered annually in each of grades 3-8 and at least once in grades 9-12** in the case of reading/language arts and mathematics assessments, and at least once in grades 3-5, 6-9, and 10-12 in the case of science assessments, so long as the statewide academic assessments under 34 CFR 200.2(a)(1) and section 1111(b)(2) of the Act are administered in any required grade and subject under 34 CFR 200.5(a)(1) in which the SEA does not choose to implement an innovative assessment;
Some Reasons Why States Chose Not to Apply in IADA Round 1

- Didn’t think they were ready yet (issues around building capacity for this work, especially in large states).
- Believed the state could continue innovative assessment design process without yet touching accountability realm.
- Concerns about scaling the innovative system statewide in seven years with no funding provided by the federal government.
- Concerns about ensuring comparability between the results of two state assessment systems.
- Other reasons…
Purpose of this Symposia

• The purpose of this symposia is to discuss practical considerations related to the design and implementation of innovative assessment and accountability systems, as well as early research about effects of such systems on student achievement outcomes.
Symposia Overview

• **Presentation #1:** Effects of NH’s PACE Pilot on Student Achievement Outcomes (2014-2017) – Carla Evans

• **Presentation #2:** MA Consortium of Innovative Education Assessment (MCIEA): Building a New Model of School Accountability – Andresse St. Rose

• **Discussant Remarks:** Paul Leather

• **Q &A/Discussion**
Presentation #1:

Carla M. Evans, Ph.D.
Center for Assessment
cevans@nciea.org
Study Purpose

• To examine the effects of a pilot program that utilizes performance-based assessments to make determinations of student proficiency in a school accountability context.

• New Hampshire’s Performance Assessment of Competency Education (PACE) pilot was officially approved by the U.S. Department of Education in March 2015 and currently operates under a first-in-the-nation waiver from federal statutory requirements related to state annual achievement testing.

  — PACE is now in its fourth year of implementation (2014-15 to 2017-18)—this study examines the first three years.
<table>
<thead>
<tr>
<th>Grade</th>
<th>English Language Arts</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Statewide achievement test</td>
<td>Local and common performance assessments</td>
</tr>
<tr>
<td>4</td>
<td>Local and common performance assessments</td>
<td>Statewide achievement test</td>
</tr>
<tr>
<td>5</td>
<td>Local and common performance assessments</td>
<td>Local and common performance assessments</td>
</tr>
<tr>
<td>6</td>
<td>Local and common performance assessments</td>
<td>Local and common performance assessments</td>
</tr>
<tr>
<td>7</td>
<td>Local and common performance assessments</td>
<td>Local and common performance assessments</td>
</tr>
<tr>
<td>8</td>
<td>Statewide achievement test</td>
<td>Statewide achievement test</td>
</tr>
<tr>
<td>9</td>
<td>Local and common performance assessments</td>
<td>Local and common performance assessments</td>
</tr>
<tr>
<td>10</td>
<td>Local and common performance assessments</td>
<td>Local and common performance assessments</td>
</tr>
<tr>
<td>11</td>
<td>Statewide achievement test</td>
<td>Statewide achievement test</td>
</tr>
</tbody>
</table>

**The PACE Pilot scales each year:**

Four school districts self-selected into the PACE Pilot in Year 1 (2014-15), another four districts joined in Year 2 (2015-16), and one more district joined in Year 3 (2016-17)
What is the NH PACE Pilot?

Local performance assessments → Competency 1
Local performance assessments → Competency 2
Local performance assessments → Competency 3
Local performance assessments → Competency 4

District-Level Competency Scores + Teacher Judgment Surveys

PACE Common Performance Tasks

SBAC/SAT once per grade span

PACE Annual Determinations of Student Proficiency

FIGURE 1. NH PACE pilot annual determination graphic. [Color figure can be viewed at wileyonlinelibrary.com]
Explicit involvement of local educational leaders in designing and implementing the accountability system

Reciprocal support for local districts including technical, policy, and practical guidance

Use of competency-based approaches to instruction, learning and assessment.

Use of curriculum-embedded, high-quality performance-based assessments

Fosters positive organizational learning and change by supporting internally-driven motivation

Builds local capacity of teachers and administrators

Restructures the rigor and content representation of curriculum, instruction, and assessment

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

Figure 1. NH PACE Theory of Action
Research Questions

1. What is the average effect of the PACE pilot on Grade 8 and 11 student achievement in mathematics and English language arts in the first three years?

2. To what extent do effects vary for certain subgroups of students?

3. To what extent does the number of years a district has implemented the PACE pilot affect student achievement outcomes? (i.e., dosage effects)
Study Design

• **Sample Selection Process**
  
  – All NH public school students in Grades 8 and 11 during the first three years of the PACE pilot (2014-15 to 2016-17) that also have prior achievement test results and student background/demographic information available (N= ~36,000 students/grade and subject area).

  – Cross-sectional, not longitudinal (different students analyzed across years).
Making Appropriate Comparisons

• Gold standard of all research is random selection from the population and then random assignment into treatment with control; that is not possible in almost all research.

• PACE districts self-select into the pilot → selection bias

• How did I account for pre-existing differences between PACE and non-PACE districts?
  – Propensity score weighting tries to mimic random assignment so we can accurately compare PACE vs. non-PACE student performance. It is still not random assignment, but it as close as we can get.
District Characteristics of Groups are Roughly Equivalent Prior to Analyses

*Results are descriptive, not causal*

<table>
<thead>
<tr>
<th></th>
<th>Gr 8</th>
<th>Gr 11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IEP</td>
<td>FRL</td>
</tr>
<tr>
<td>Non-PACE</td>
<td>15%</td>
<td>27%</td>
</tr>
<tr>
<td>PACE</td>
<td>14%</td>
<td>29%</td>
</tr>
</tbody>
</table>
Analytic Approach

• RQ#1: Since students are nested within schools, I used multilevel modeling to estimate the average treatment effects of the PACE pilot on Grade 8 and 11 math and ELA achievement.

• RQ#2: I then examined cross-level interactions between the treatment variables and student-level characteristics (prior achievement, gender, IEP status, socioeconomic status) in order to see if effects varied for certain subgroups.

• RQ#3: Dosage effects were also examined (one, two or three years).
RQ#1: Grade 8 Average Effects

G8Math

- **2015-16**: 0.18
- **2016-17**: 0.09

Small Positive Effect (d=0.14)

G8ELA

- **2015-16**: Non-PACE 0.03, PACE 0.04
- **2016-17**: Non-PACE 0.03, PACE 0.06

Almost “No” Effect (d=0.06)
RQ#1: Grade 11 Average Effects

G11Math

Almost “No” Effect (d=0.03)

<table>
<thead>
<tr>
<th>Standard Deviations</th>
<th>Non-PACE</th>
<th>PACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-16</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>2016-17</td>
<td>0.00</td>
<td>0.02</td>
</tr>
</tbody>
</table>

G11ELA

Small Positive Effect (d=0.09)

<table>
<thead>
<tr>
<th>Standard Deviations</th>
<th>Non-PACE</th>
<th>PACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-16</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>2016-17</td>
<td>0.00</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Quick Summary of RQ#1 Findings

• Findings suggest that there were small positive effects of the PACE pilot in all examined grades and subjects – range in magnitude from about 3% to 14% of a standard deviation.

• There does not appear to be a consistent pattern of effects in one subject area as effects vary by grade.
## RQ#2: Subgroup Analysis

<table>
<thead>
<tr>
<th>Student Subgroup</th>
<th>Differential Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Prior Achievement</td>
<td>Positive</td>
</tr>
<tr>
<td>Male</td>
<td>Negative</td>
</tr>
<tr>
<td>Students with Disabilities</td>
<td>Positive/Negative</td>
</tr>
<tr>
<td>Free-and-reduced price lunch</td>
<td>Positive/Negative</td>
</tr>
</tbody>
</table>

_Caution: Share of students falling into these categories was small._
Implications

• Findings could be used to provide assurance to key stakeholders that PACE students are “not harmed” as a result of participating in the PACE pilot and provided an equitable opportunity to learn the content standards → political coverage for other states interested in applying in future IADA Rounds?

• Provides early evidence that learning gains exhibited by students resulting from this large-scale performance assessment program may be transferring or carrying over to a very different assessment of student proficiency—the state achievement. If true, signals that deeper learning has taken place.

• These are early effects and this study has limitations. It is important to continue to study effects over time and with other outcomes as well.
NH PACE Practical Considerations
Re:1204 Application

• Leadership changes/political will
• **Funding**: state education funding (no income or property taxes) and role of NHLI
• **Building LEA capacity around assessment literacy at scale**
• **Data collection demands** – LEA leadership support, capacity (small districts vs. large districts), and “fatigue” over time
• **Technology-related issues** – no product out there that meets our needs; we are now working with Motivis to design a custom-made solution
• **Scaling issues in a local control state**
NH PACE Technology Wish List

1. Collaborative synchronous and asynchronous performance assessment development;
2. Searchable warehousing of performance tasks along with accompanying administration documentation;
3. Distributed double-blind scoring for the purposes of calibration and monitoring inter-rater reliability;
4. Secure uploading, storage and sharing of student portfolios of work; and
5. Data capturing system that works seamlessly with a diverse set of district student information systems to transfer student-level task scores, competency scores, and teacher judgment scores.
NH PACE Proposed Strategy for Scaling

Section 1204 requires scaling statewide by the end of 7 years. We think there are multiple paths to “scaling” as illustrated here.

- Pedagogical expectations for all educators
- Personalized by student
- One subject area, one grade span (e.g., middle school science)
- All grades, one subject or all subjects, one grade span
- All grades & subjects
Presentation #2:
Massachusetts Consortium of Innovative Education Assessment (MCIEA): Building a New Model of School Accountability

Andresse St. Rose, Ed.D.
Center for Collaborative Education
astrose@ccbos.org
Where is MCIEA?

MCIEA is a partnership of public school districts and their local teacher unions from Attleboro, Boston, Lowell, Revere, Somerville, and Winchester. MCIEA is partnering with the Center for Collaborative Education and the University of Massachusetts, Lowell.
What is MCIEA?

The Massachusetts Consortium for Innovative Education Assessment (MCIEA) is committed to establishing fair and authentic ways of assessing student learning and school quality that champions students, teachers, and communities. MCIEA seeks to increase achievement for all students and close prevailing achievement gaps among subgroups.

High-stakes Testing:
- Narrows the curriculum
- Devalues teachers
- Misinforms the public about school quality
MCIEA defines “performance assessments” as multi-step, fair assignments with clear criteria, expectations, and processes that enable students to interact with meaningful content and that measure how well a student transfers knowledge and applies complex skills and dispositions to create or refine an original product and/or solution.
Performance Assessments with Technical Quality

**Local/School-level**
- QPA Institutes
- PLCs
- Coaching

**District-level**
- Cross-district PA Committees

---

**Capacity Building**

---

**PA Creation**
- Create Local PAs
- Quality PA Review

---

**Field Testing**
- Create Local PA Bank
- Administer Local PAs
- Scoring by Local Teachers
- Administer Cross-district PAs (1 PA per grade per discipline)
- Cross-district Scoring
- Achieve Inter-rater Reliability

---

**A Student’s Body of Work**
- Student Work Samples from Multiple Local PAs
- Student Work Sample from One Cross-district PA

---

**Proficiency Determination**
- Benchmarking Proficiency
- Cross-district Scoring
- Reporting on Proficiency and Growth
### MCIEA Logic Model

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Activities</th>
<th>Outputs</th>
<th>Short-Term Outcomes</th>
<th>Mid-Term Outcomes</th>
<th>Long-Term Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funding</td>
<td>Quality Performance Assessment (QPA) Professional Development Institute</td>
<td>Five days of training on QPA design cycle and performance assessment (PA) systems for lead teams in each cohort</td>
<td>Increased PA literacy for all faculty in cohort schools</td>
<td>Increased quality of curriculum embedded, standards aligned PAs</td>
<td>Increased student learning</td>
</tr>
<tr>
<td>CCE Staff</td>
<td>Coaching</td>
<td>10 days of coaching per school per cohort with full faculty to support school-wide implementation of PAs</td>
<td>Increased confidence in PA design for all faculty in cohort schools</td>
<td>Improved school-wide professional collaboration practices for all faculty in cohort schools</td>
<td>Decreased achievement gaps by race, income, language, and/or disability</td>
</tr>
<tr>
<td>District administrators and resources</td>
<td>Cross-district Performance Assessment Committees</td>
<td>18 teacher leaders facilitate 9 cross-district PA Committees that meet 4 times annually to create common PAs</td>
<td>Increased belief that PAs are a key lever to improve student learning and achievement for all faculty in cohort schools</td>
<td>Increased perception that performance assessments are a school-wide priority for all faculty in cohort schools</td>
<td></td>
</tr>
<tr>
<td>School leadership, faculty, and resources</td>
<td>Network Calibration and Performance Assessment Validation Sessions</td>
<td>Lead teams in every cohort meet 3 times to validate common PAs and calibrate student work</td>
<td>Increased knowledge and skills in leading professional collaborative practices for teacher leaders</td>
<td>Increased presence of conditions and structures that support effective PA systems at MCIEA schools</td>
<td></td>
</tr>
<tr>
<td>Center for Assessment</td>
<td>Leadership Support</td>
<td>25-30 school leaders per cohort converse 3-4 times per year to support the implementation of PAs school-wide</td>
<td>Increased inter-rater reliability and comparability in scoring student work from PAs for lead teams</td>
<td>Improved quality of teacher instructional and assessment practices across schools towards deeper learning and personalized learning for all faculty in cohort schools</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MCIEA Webpage with Performance Assessment Bank</td>
<td>25-30 PAs are vetted and uploaded to the PA Bank per year for use by all teachers in MCIEA schools</td>
<td>Increased use of Local and common PAs for all faculty in cohort schools</td>
<td>Increased school leader knowledge of the conditions and structures that support effective PA systems</td>
<td>Increased student engagement</td>
</tr>
</tbody>
</table>
Evaluation Questions

- How and to what extent does teacher leader performance assessment literacy change after participating in the QPA professional development institute?

- How and to what extent does teacher performance assessment literacy at participating MCIEA school change after participating in professional development provided by teachers leaders?
Growth in Performance Assessment Literacy Scale Components - Teacher Leaders (n=93)

*: Difference is statistically significant at .05 level

<table>
<thead>
<tr>
<th>Scale Components</th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>3.9</td>
<td>4.6*</td>
<td>0.7</td>
</tr>
<tr>
<td>Reliability</td>
<td>4.0</td>
<td>4.2*</td>
<td>0.3</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>4.0</td>
<td>4.6*</td>
<td>0.6</td>
</tr>
<tr>
<td>Fairness</td>
<td>3.4</td>
<td>4.1*</td>
<td>0.7</td>
</tr>
<tr>
<td>Student Voice and Choice</td>
<td>3.2</td>
<td>4.1*</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Mean Scores

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
Insights from the Data - Validity (Teacher Leaders)

Create assessments that are aligned to specific habits, skills, and dispositions

Create performance assessments designed to give students the opportunity to demonstrate high levels of cognitive rigor

Design performance assessments that accurately measure student proficiency on MA State standards

Identify student work products that can be used as exemplars for other students

Use backwards design/planning to organize my units and lessons

Create assessments that are clearly aligned to MA State standards

<table>
<thead>
<tr>
<th>Pre</th>
<th>3.5</th>
<th>3.6</th>
<th>3.8</th>
<th>4.1</th>
<th>4.1</th>
<th>4.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post</td>
<td>4.4</td>
<td>4.4</td>
<td>4.5</td>
<td>4.7</td>
<td>4.8</td>
<td>4.8</td>
</tr>
<tr>
<td>Growth</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident

*: Difference is statistically significant at .05 level
Create a rubric for use with multiple assessments so students can easily track their progress and growth from one assessment to the next.

Create rubrics that have clear criteria and descriptions of student performance at each level.

Develop common rubrics with other educators.

Calibrate scoring of student work with colleagues using a common rubric.

Identify student work samples that can be used as anchors for scoring.

Use a rubric to score student work.

### Reliability - Component Mean Scores - Teacher Leaders Only (n=94)

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create a rubric for use with multiple assessments so students can easily track their progress and growth from one assessment to the next</td>
<td>3.4</td>
<td>3.9*</td>
<td>0.5</td>
</tr>
<tr>
<td>Create rubrics that have clear criteria and descriptions of student performance at each level</td>
<td>3.8</td>
<td>4.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Develop common rubrics with other educators</td>
<td>4.0</td>
<td>4.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Calibrate scoring of student work with colleagues using a common rubric</td>
<td>4.2</td>
<td>4.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Identify student work samples that can be used as anchors for scoring</td>
<td>4.2</td>
<td>4.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Use a rubric to score student work</td>
<td>4.4</td>
<td>4.6</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident.

*: Difference is statistically significant at .05 level.
Insights from the Data - Data Analysis (Teacher Leaders)

Data Analysis - Mean Component Scores - Teacher Leaders Only (n=93)
*: Difference is statistically significant at .05 level

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create performance</td>
<td>3.2</td>
<td>4.3</td>
<td>1.1</td>
<td>4.3</td>
</tr>
<tr>
<td>assessments that</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>provide actionable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>feedback about your</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>students’ learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personalize instruction</td>
<td>3.9</td>
<td>4.5</td>
<td>0.5</td>
<td>4.5</td>
</tr>
<tr>
<td>for individual students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>based on student</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>assessment data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analyze and reflect on</td>
<td>4.0</td>
<td>4.6</td>
<td>0.6</td>
<td>4.6</td>
</tr>
<tr>
<td>student assessment data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>on my own</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjust instruction for</td>
<td>4.1</td>
<td>4.6</td>
<td>0.6</td>
<td>4.6</td>
</tr>
<tr>
<td>particular groups of</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>students based on</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>student assessment data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modify instruction for</td>
<td>4.2</td>
<td>4.6</td>
<td>0.4</td>
<td>4.6</td>
</tr>
<tr>
<td>students based on</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>student assessment data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discuss and interpret</td>
<td>4.2</td>
<td>4.8</td>
<td>0.6</td>
<td>4.8</td>
</tr>
<tr>
<td>student assessment data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with colleagues</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
Insights from the Data - Fairness (Teacher Leaders)

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop performance assessments that incorporate content on diverse cultures and traditions</td>
<td>3.0</td>
<td>3.9*</td>
<td>0.9</td>
</tr>
<tr>
<td>Design performance assessments that provide students with multiple pathways to demonstrate their knowledge</td>
<td>3.4</td>
<td>4.2*</td>
<td>0.8</td>
</tr>
<tr>
<td>Incorporate accommodations into assessments for English Language Learners</td>
<td>3.5</td>
<td>4.2*</td>
<td>0.7</td>
</tr>
<tr>
<td>Design assessments that are free of stereotypes about cultural and linguistic groups</td>
<td>3.5</td>
<td>4.1*</td>
<td>0.6</td>
</tr>
<tr>
<td>Incorporate accommodations into assessments for students with disabilities</td>
<td>3.8</td>
<td>4.3*</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
## Insights from the Data - Student Voice and Choice (Teacher Leaders)

**Student Voice and Choice - Mean Component Scores - Teacher Leaders Only (n=91)**

*: Difference is statistically significant at .05 level

<table>
<thead>
<tr>
<th>Component</th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create performance assessments that allow students to set their own learning goals</td>
<td>2.7</td>
<td>3.8*</td>
<td>1.1</td>
</tr>
<tr>
<td>Design performance assessments that provide students with feedback to make decisions about their learning</td>
<td>3.1</td>
<td>4.1*</td>
<td>1.1</td>
</tr>
<tr>
<td>Design performance assessments that allow students to exercise ownership and decision making</td>
<td>3.3</td>
<td>4.1*</td>
<td>0.9</td>
</tr>
<tr>
<td>Develop performance assessments that provide students with opportunities to reflect on their learning</td>
<td>3.3</td>
<td>4.3*</td>
<td>1.0</td>
</tr>
<tr>
<td>Develop assessments that promote an academic growth mindset</td>
<td>3.4</td>
<td>4.3*</td>
<td>0.9</td>
</tr>
<tr>
<td>Create performance assessments that focus on addressing authentic problems</td>
<td>3.4</td>
<td>4.4*</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
## Insights from the Data - Performance Assessment Literacy Scale (Non-Teacher Leaders)

Growth in Performance Assessment Literacy Scale Components - Non-Teacher Leaders (n=333)

\*: Difference is statistically significant at .05 level

<table>
<thead>
<tr>
<th>Component</th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>4.2</td>
<td>4.4*</td>
<td>0.2</td>
</tr>
<tr>
<td>Reliability</td>
<td>4.1</td>
<td>4.4*</td>
<td>0.3</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>4.2</td>
<td>4.4*</td>
<td>0.2</td>
</tr>
<tr>
<td>Fairness</td>
<td>3.9</td>
<td>4.1*</td>
<td>0.3</td>
</tr>
<tr>
<td>Student Voice and Choice</td>
<td>3.7</td>
<td>4.0*</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Mean Scores

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
Create assessments that are aligned to specific habits, skills, and dispositions

Create performance assessments designed to give students the opportunity to demonstrate high levels of cognitive rigor

Design performance assessments that accurately measure student proficiency on MA State standards

Use backwards design/planning to organize my units and lessons

Create assessments that are clearly aligned to MA State standards

Identify student work products that can be used as exemplars for other students

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create assessments</td>
<td>3.9</td>
<td>4.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Design performance</td>
<td>4.0</td>
<td>4.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Use backwards</td>
<td>4.0</td>
<td>4.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Create assessments</td>
<td>4.2</td>
<td>4.4</td>
<td>0.1</td>
</tr>
<tr>
<td>Identify student</td>
<td>4.4</td>
<td>4.6</td>
<td>0.2</td>
</tr>
<tr>
<td>work products</td>
<td>4.7</td>
<td>4.8</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
Insights from the Data - Reliability (Non-Teacher Leaders)

Create a rubric for use with multiple assessments so students can easily track their progress and growth from one assessment to the next.

Create rubrics that have clear criteria and descriptions of student performance at each level.

Develop common rubrics with other educators.

Calibrate scoring of student work with colleagues using a common rubric.

Identify student work samples that can be used as anchors for scoring.

Use a rubric to score student work.

Reliability - Component Mean Scores Non-Teacher Leaders Only (n=321)
*: Difference is statistically significant at .05 level

<table>
<thead>
<tr>
<th></th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create a rubric for use with multiple assessments</td>
<td></td>
</tr>
<tr>
<td>Create rubrics that have clear criteria and descriptions of student performance at each level</td>
<td></td>
</tr>
<tr>
<td>Develop common rubrics with other educators</td>
<td></td>
</tr>
<tr>
<td>Calibrate scoring of student work with colleagues using a common rubric</td>
<td></td>
</tr>
<tr>
<td>Identify student work samples that can be used as anchors for scoring</td>
<td></td>
</tr>
<tr>
<td>Use a rubric to score student work</td>
<td></td>
</tr>
</tbody>
</table>

Pre 3.7 4.0 4.2 4.2 4.3 4.5
Post 4.0 4.3 4.4 4.4 4.6 4.7
Growth 0.3 0.3 0.3 0.2 0.2 0.2

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
Insights from the Data - Data Analysis (Non-Teacher Leaders)

<table>
<thead>
<tr>
<th>Action Description</th>
<th>Mean Score</th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create performance assessments that provide actionable feedback about your students’ learning</td>
<td>3.7</td>
<td>4.0*</td>
<td>4.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Personalize instruction for individual students based on student assessment data</td>
<td>4.2</td>
<td>4.4*</td>
<td>4.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Analyze and reflect on student assessment data on my own</td>
<td>4.2</td>
<td>4.5*</td>
<td>4.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Discuss and interpret student assessment data with colleagues</td>
<td>4.3</td>
<td>4.6*</td>
<td>4.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Adjust instruction for particular groups of students based on student assessment data</td>
<td>4.3</td>
<td>4.5</td>
<td>4.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Modify instruction for students based on student assessment data</td>
<td>4.4</td>
<td>4.5*</td>
<td>4.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Data Analysis - Mean Component Scores - Non-Teacher Leaders Only (n=317)

*: Difference is statistically significant at .05 level

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
Insights from the Data - Fairness (Non-Teacher Leaders)

Fairness - Mean Component Scores - Non-Teacher Leaders Only (n=316)
*: Difference is statistically significant at .05 level

<table>
<thead>
<tr>
<th>Mean Score</th>
<th>Develop performance assessments that incorporate content on diverse cultures and traditions</th>
<th>Design performance assessments that provide students with multiple pathways to demonstrate their knowledge</th>
<th>Incorporate accommodations into assessments for English Language Learners</th>
<th>Incorporate accommodations into assessments for students with disabilities</th>
<th>Design assessments that are free of stereotypes about cultural and linguistic groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>3.5</td>
<td>3.7</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Post</td>
<td>3.9</td>
<td>4.0</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>Growth</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident
Insights from the Data - Student Voice and Choice (Non-Teacher Leaders)

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create performance assessments that allow students to set their own learning goals</td>
<td>3.4</td>
<td>3.6</td>
<td>0.2</td>
</tr>
<tr>
<td>Design performance assessments that provide students with feedback to make decisions about their learning</td>
<td>3.6</td>
<td>3.9</td>
<td>0.3</td>
</tr>
<tr>
<td>Design performance assessments that allow students to exercise ownership and decision making</td>
<td>3.7</td>
<td>4.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Create performance assessments that focus on addressing authentic problems</td>
<td>3.7</td>
<td>4.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Develop performance assessments that provide students with opportunities to reflect on their learning</td>
<td>3.8</td>
<td>4.1</td>
<td>0.3</td>
</tr>
<tr>
<td>Develop assessments that promote an academic growth mindset</td>
<td>3.8</td>
<td>4.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Student Voice and Choice - Mean Component Scores - Non-Teacher Leaders Only (n=309)

*: Difference is statistically significant at .05 level

Scale: 1 = Not at all confident, 2 = A little confident, 3 = Moderately confident, 4 = Confident, 5 = Very confident, 6 = Completely confident

Mean Score

- 6.0
- 5.0
- 4.0
- 3.0
- 2.0
- 1.0
Implications

- Results provide early evidence on a key mediating factor - increased performance assessment literacy of teacher leaders.
- The results also provide suggestive evidence on a short-term outcome - increased performance assessment literacy of faculty schoolwide. But we know that implementation at the school-level, i.e., scaling was inconsistent across schools and not uniform over time.
- Major limitation is that all evidence is based on self-reports (we also have some focus group data that supports and provides insight to the quantitative results).
A Re-Look at the HumRRO Formative Assessment Results and the Problem of Scale

Paul Leather
Director, Local and State Partnerships
3 critical cornerstones essential for successful performance assessment scale-up initiatives –

• **robust, sustained professional development** to build teacher capacity to create high-quality, curriculum-embedded performance assessments;

• **technical quality** to ensure that performance tasks are valid and student work is scored reliably; and

• **political leadership and policy support** that enables performance assessment initiatives to be successful and sustaining.

Tung & Stazesky. CCE 2010
Rethinking Scale – Cynthia Coburn, 2003

Four Dimensions:

– Depth of Pedagogical Change
– Sustainability
– Spread
– Shift in Reform Ownership

HumRRO PACE Formative Evaluation:
https://docs.wixstatic.com/ugd/10b949_696ca7f8484c4418825bee921fbc6c5f.pdf
We understand that the PACE stakeholders are not test design experts and, therefore, that the AERA, APA, & NCME Standards are not firsthand knowledge for this audience. Consequently, our discussion with these stakeholders referred more generally to “high-quality assessment.”
A Review of Progress Against the Claims

- **Claim 1a. Local leadership is clearly committed** -- The overwhelming majority of PACE participants reported high levels of commitment.

- **Claim 1b. Participating districts collaborate with one another** – The evaluators found multiple ways in which PACE districts collaborate.

- **Claim 2a. Teachers developing performance tasks are trained and knowledgeable of the Joint Standards for test development** -- PACE teachers demonstrated high levels of assessment literacy during training sessions, scoring, and standards setting meetings.

- **Claim 2b. Performance assessments must adhere to the Joint Standards, including ensuring equity** -- PACE results are compared with an external reference assessment (Smarter Balanced)... largely parallel the processes of large-scale testing companies that adhere to the Joint Standards and they contribute to a high quality assessment system.
Teacher Training Effectiveness

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

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

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

• **Teachers across districts expressed that implementing performance tasks has had a positive impact on their instruction.**
Student Engagement in PACE

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

- Teachers report higher engagement for their students and deeper learning of the content, during PACE assessments and as a result of improvements in their instructional practice that they attribute to participating in PACE. The majority of students report that they would rather take a PACE assessment than an end-of-year comprehensive test like Smarter Balanced or the New England Comprehensive Assessment Program (NECAP) test.
Effectiveness of Training Scoring of Tasks

- **Claim 4a. Scorers are effectively trained**
- the overall scoring consistency is quite high and few adjustments are necessary to the initially set cut scores due to inconsistent scoring (either too lenient or too strict) within the districts, indicating **effective training for the scoring of PACE tasks**. This process ensures consistency of scoring across districts. It is also the way that scores are made comparable across years.
Reliability

- **Claim 4b. Scorers attain successful rates of interrater agreement and reliability**
- The Center for Assessment computes within-district rater agreement statistics (e.g. % exact agreement, % adjacent agreement) and Cohen’s Kappa statistics for a sample of the double-scored common tasks (Evans & Lyons, 2016). **Pairs of raters had exact agreement rates of between approximately 60 and 85%**. There were substantial differences by grade, subject, dimension, and by district, but **nearly all districts achieved greater than 60% exact agreement rates across all grade subjects**. Kappa statistics indicate moderate to substantial agreement of ratings across all grades and subjects as well
Pedagogical Change and Teacher Ownership

- **Educators are in charge of nearly all aspects of the program.** Teachers decide what is assessed, how it is assessed, and they even design the scoring rubrics. By placing the responsibility for creating the tasks on the primary users of the assessment data, PACE gives teachers more say in how their students will be assessed than in more traditional testing systems. **Educators at all levels described ownership of the system as a major contributor to buy-in.**

- Unlike end-of-year comprehensive statewide assessments, which sample from the past year’s curriculum, **PACE is targeted to the learning that is occurring at the time of administration.** Since there is no specific testing window for PACE, and since the tasks are targeted to one broad curricular topic, teachers can administer the tasks when it makes the most sense.

- Teachers routinely design assessments to check progress on the content they teach, and they did so prior to the PACE program. **PACE adds the competency aspect,** though many schools had implemented some form of competency education previously, placing the focus of the assessment on competency rather than progress or performance relative to peers.
The Issue of Sustainability

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

• PACE must prove that it is scalable. New districts are joining PACE, but NH DOE recognizes the considerable challenges involved in scaling PACE statewide as it is currently conceived, as indicated by NH DOE leadership and reiterated by district superintendents during interviews.

• there is a great deal of preparation a district must do to become a Tier 1 PACE district. It would be difficult to suddenly implement PACE on a much broader scale because of the integrated nature of task development, teacher professional development, and collaboration. Getting a full state’s population of teachers to suddenly begin to effectively collaborate seems unlikely. In New Hampshire, PACE began with a few highly motivated districts and is expanding carefully.
Shift in Ownership – Content Teacher Leaders

• A collaboration mechanism -- the naming of multiple Content Leads (about 30 total) for each grade level and content area combination. These teachers were identified as leaders in PACE and were recommended by peers and ultimately selected by the PACE District Leads to help coordinate subject/grade-specific activities. Most have been PACE participants and task developers since the beginning of the PACE pilot program. The Content Leads program allows PACE to build deep expertise among local educators without requiring all educators to attend every meeting and activity. The Content Leads helped PACE address the expansion of the program. They act as liaisons to the educators in their districts and also in a “buddy district,” which might not have a Content Lead.
Developing a Theory of Action to Scale PACE

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

- Getting new staff members oriented to complex new ways of educating students takes considerable time and effort. If the experienced teachers train the new ones, they will need time to do so. They will need time in addition to the time they spend implementing PACE in their own schools and classrooms.

- There may also be performance gaps between the experienced and newly joined districts. These issues, as well as potential changes in the political and economic climate in which PACE is being implemented will likely challenge PACE. The sustainability of PACE will rely on demonstrating that the benefits of PACE continue to outweigh the challenges. For this to happen, PACE will require continuous feedback and improvement as the system expands.
Theory of Action
(How top down meets bottom up)

START with SEAs where existing district/CMO work lives and interest in Competency Education is high

POLICY FORMATION
Problem (Define/Re-define)
Solution (Try/Iterate)

TRANSLATE into policy formation

TEST and COLLECT evidence, lessons learned, and models of practice

INVEST in networks of diverse local school systems that are putting pieces of CBE in practice

Re/frame the “mental model” for SEA/LEA leaders

Develop elements of “policy” model + practices

Develop elements of a “system” model
Questions?
Thank you!

Carla Evans  
*Center for Assessment*  
Andresse St. Rose  
*Center for Collaborative Education*  
Paul Leather  
*Center for Innovation in Education*