Assessment and Accountability to Support Meaningful Learning

Scott Marion
National Center for the Improvement of Educational Assessment
&
Paul Leather
New Hampshire Department of Education
United States

Abstract: This paper presents an overview of New Hampshire's efforts to implement a pilot accountability system designed to support deeper learning for students and powerful organization change for schools and districts. The accountability pilot, referred to as Performance Assessment of Competency Education or PACE, is grounded in a competency-based educational approach designed to ensure that students have meaningful opportunities to achieve critical knowledge and skills. These opportunities are judged by the outcomes students achieve and not by inputs such as seat time. Therefore, students must achieve these competencies before moving on to the next major learning targets and/or graduating from high school. High quality performance assessments play a crucial role in the PACE system because of the need to have assessments that measure the depths of student understanding of these complex learning targets. Performance assessments are used as both summative and interim

measures in the PACE system as a way to document student learning of the competencies and to support remediation or extension interventions. The paper describes the system of assessments being implemented as part of the PACE pilot as well as providing a discussion of the technical quality issues the state is working to address as part of this accountability pilot. For example, being able to produce valid and comparable annual determinations for all students each year is a considerable technical challenge as well as documenting the degree to which all students are held to the same threshold expectations (equity). The paper concludes by relating the PACE initiative to the push for deeper and more meaningful learning for students.
Keywords: Assessment; accountability; meaningful learning; equity; PACE.

Evaluación y modelos de responsabilidad educativa para apoyar aprendizajes significativos

Resumen: Este artículo presenta una visión general de los esfuerzos del estado de New Hampshire para implementar un sistema piloto de responsabilidad educativa diseñado para apoyar aprendizajes más consistentes para los estudiantes y de cambio organizacional substantivos para las escuelas y los distritos. El sistema piloto de responsabilidad educativa, denominado Evaluación del Desempeño de Competencias Educativas o PACE (por su sigla en inglés), es un enfoque educativo basado en competencias diseñado para asegurar que los estudiantes tengan oportunidades significativas para alcanzar conocimientos y habilidades críticas. Estas oportunidades son juzgadas por los resultados que los estudiantes logran y no por los insumos utilizados como el tiempo que pasan sentados. Por lo tanto, los estudiantes deben alcanzar estas competencias antes de avanzar a nuevos objetivos de aprendizaje y/o de graduarse de la escuela secundaria. Evaluaciones de desempeño de alta calidad desempeñan un papel crucial en el sistema PACE debido a la necesidad de contar con evaluaciones que miden la profundidad de comprensión de los alumnos de estos objetivos de aprendizaje complejos. Las evaluaciones de desempeño se utilizan como medidas tanto sumativa y provisionales en el sistema PACE como una manera de documentar el aprendizaje de las competencias y para apoyar intervenciones de rehabilitación y ampliación. Este trabajo describe el sistema de evaluación que se implementa como parte del programa piloto PACE, y también una discute cuestiones técnicas sobre calidad que el estado de New Hampshire está trabajando para abordar en el marco de este proyecto piloto de responsabilidad educativa. Por ejemplo, ser capaz de producir determinaciones anuales válidas y comparables para todos los estudiantes cada año es un gran desafío técnico, así como documentar el grado en el que todos los estudiantes deben cumplir con las mismas expectativas de aceptación (equidad). El documento concluye relacionando la iniciativa PACE para estimular aprendizajes más profundos y significativos para todos los estudiantes.
Palabras clave: evaluación; responsabilidad educativa; aprendizaje significativo; equidad; PACE.

Avaliação e modelos de responsabilidade educacional para apoiar aprendizagens significativas

Resumo: Este artigo apresenta uma visão geral dos esforços do estado de New Hampshire para implementar um sistema piloto de responsabilização educacional projetado para apoiar aprendizagem mais consistente para os estudantes e mudança organizacional substantiva para escolas e distritos. O programa piloto de responsabilização educacional, chamado Avaliação do Desempenho de Habilidades Educativas ou PACE (por sua sigla em Inglês), é uma abordagem educativa baseada em competências que foi desenhado para assegurar que os alunos tenham oportunidades significativas para adquirir conhecimentos e competências críticas. Estas oportunidades são julgados pelos resultados que os alunos alcançam e não os insumos utilizados como o tempo que passam nas cadeiras. Portanto, os alunos devem alcançar essas habilidades antes de avançar para novos objetivos de aprendizagem e /ou completar a escola secundária. As avaliações

de desempenho de alta qualidade desempenham um papel crucial no sistema PACE, devido à necessidade de avaliações que medam a profundidade da compreensão dos alunos sobre esses objetivos de aprendizagem complexos. As avaliações de desempenho são usados como medidas provisórias e sumativa no sistema PACE e como uma forma de documentar as habilidades de aprendizagem e intervenções de apoio e expansão. Este artigo descreve o sistema de avaliação implementada como parte do programa PACE piloto, e também discute questões técnicas sobre a qualidade que o estado de New Hampshire está trabalhando para resolver como parte deste projeto piloto de responsabilidade educacional. Por exemplo, ser capaz de produzir determinações anuais válidas e comparáveis para todos os alunos de cada ano é um grande desafio técnico, bem como documentar a medida em que todos os alunos cumpram as mesmas expectativas de aceitação (equidade). O artigo conclui relacionando a iniciativa PACE para estimular aprendizagens mais profundos e significativos para todos os alunos.
Palavras-chave: avaliação; responsabilidade educativa; aprendizagem significativa; equidade; PACE.

Introduction

States have held schools accountable for academic performance for many years. The federal role and requirements for such accountability systems were first implemented comprehensively with the passage of ESEA in 1965, but it was later reauthorizations in 1994 (IASA) and ramped up in 2001 (NCLB) where we have seen state-led school accountability systems become a prominent feature on the educational landscape. There is no question that the United States has experienced improvements in educational outcomes since the 1960s and more recently since the passage of NCLB, however most would agree that these trends are far short of the policy promises behind these initiatives. Further, when compared to rate of improvement observed in many other countries, the performance in the United States looks stagnant. So how do we improve performance at scale and is there a role that school accountability can play to help bring about these improvements?

Current U.S. accountability system designs appear to run counter to significant bodies of research about both organizational change and human learning. Research on organizational change/reform and human learning supports the notion that real change/learning must be internally controlled and motivated (e.g., Bransford, Brown, & Cocking, 2000). One could make a case that many of the current designs, using both "carrots and sticks," follow some premises of incentive-based economic perspectives, but if the goal is to improve performance, it does not seem to make sense to essentially ignore the research about how to actually improve organizational and individual performance.

Several states involved in the Council of Chief State School Officers' (CCSSO) Innovative Laboratory Network (ILN) have been exploring alternatives to current accountability system with a goal of deepening and improving student learning. New Hampshire has been a key state member of the ILN and has advanced efforts to pilot an accountability system designed to foster more meaningful learning for students (CCSSO, 2012; Domaleski & Hall, 2013). This paper presents a discussion of this initiative. Specifically, we first present a brief overview of the learning theory literature as it informs New Hampshire's work. Next we describe how a competency-based approach for organizing instruction and assessment, particularly performance-based assessments, can support the goals of deeper learning. Given state and federal accountability demands, the instructional and assessment initiatives described in this paper must be coupled with an accountability framework that supports, rather than hinders, deeper learning. We describe New Hampshire's accountability pilot, Performance Assessment of

Competency Education (PACE), as an example of an accountability system designed to support more meaningful individual and institutional learning. We conclude with a discussion of some challenges and opportunities of supporting local expertise necessary for high quality implementation.

Deeper Learning

There have been multiple theoretical lines of inquiry attempting to better explain the way in which humans learn and develop expertise. Both the cognitive (e.g., Pellegrino et al., 2001) and sociocultural perspectives (e.g., Lave & Wenger, 1991; Wertsch, 1991) provide information useful for informing this work. Several authors have suggested that there are enough areas of overlap between the two perspectives that can advance our understanding of student learning (Anderson, Greeno, Reder, & Simon, 2000; Pellegrino et al., 2001; Shepard, 2000), including: (a) cognitive abilities are influenced in large part by cultural and social factors, (b) learners construct knowledge within a social context, (c) new learning builds on and is greatly influenced by prior knowledge with includes social and cultural factors, (d) metacognition is a crucial component of the development of advanced knowledge and skills, and (e) deep understanding is characterized by the capability of the learner to transfer that understanding to new situations (Anderson et al., 2000; Shepard, 2000). We briefly discuss each of these areas of overlap below.

Learners Construct Knowledge within a Social Context

Vygotsky provided a conceptual framework for understanding how social interactions influence the construction of knowledge. Whether one assumes a sociocultural perspective and believes that culture "constitutes" learning or a more cognitive perspective whereby culture simply influences learning, it is clear that the social and cultural forces on individuals must be considered in discussions of learning (Shepard, 2000, p. 19).

New Knowledge Construction is a Product of Prior Knowledge

Prior knowledge has a tremendous influence on the formation of new knowledge. Vast stores of discipline-based concepts, algorithms, skills, and processes that can be recalled efficiently to solve problems or construct new knowledge characterize subject-matter expertise. Novices do not possess nearly the same amount of these facts and skills as experts, but more importantly, they lack well-developed schema to organize this information. Instruction needs to capitalize on the prior knowledge and cultural practices of students to help them build more efficient cognitive structures and to help them become more fully participating members of a community. Assessments, therefore, should be able to determine not just who has developed advanced knowledge compared with those who have not, but how students' prior knowledge structures influences their performance on assessment tasks.

Metacognition is a Crucial Component of the Development of Advanced Knowledge

Experts are characterized by having strong metacognitive abilities allowing them to monitor their learning and choose efficient means for solving problems. Metacognition is not reserved for experts; many types of learners can develop metacognitive skills (Palincsar & Brown, 1984). However, metacognition cannot be taught out of context of a particular subject matter domain and these strategies are bound by the structure of a given discipline (Bransford et al., 2000). Because a student's' metacognition will influence their performance on an assessment of knowledge and skills, assessments should also attempt to determine the sophistication of a student's metacognitive skills (Pellegrino et al., 2001).

Deep Subject Matter Understanding Supports Transfer

Much of what has been discussed already, particularly metacognition, the role of prior knowledge in shaping new knowledge, and the influence of social and cultural factors on knowledge are important because they support the development of deep (or expert-like) understanding. Deep understanding, or expert knowledge, is not only characterized by knowledge of a large body of facts and skills, but by the transformation of factual information into usable knowledge (Bransford et al., 2000). The literature on transfer is quite clear that when knowledge is organized into conceptual schemas and is efficiently retrievable, students are able to apply (transfer) this knowledge to new situations and to learn additional, related information more quickly (Bransford et al., 2000). This can easily be considered the most important purpose of school learning—to have students develop deep understandings that they can use in contexts beyond the classroom where it was first learned.

The development of deep understanding happens rarely in United States K-12 settings (Schmidt, McKnight, & Raizen, 1997). The development of advanced knowledge would require that students learn fewer concepts in greater depth echoing the calls of TIMMS researchers (Schmidt et al., 1997) when comparing the poor performance of U.S. students to those in other countries. Among other limitations, many large-scale assessment programs contribute to the teaching and learning of superficial content knowledge. Teachers, in their rush to ensure that all of the standards have been "covered," do not feel like they can ignore certain concepts and teach for deep conceptual understanding (Bransford et al., 2000). Further, assessing for deep understanding may not always be possible in large-scale assessments where the use of consistent administration and scoring procedures is of paramount importance. Because large-scale assessment and accountability programs drive much of what goes on in classrooms, we need to design programs to support the teaching and learning of deep understanding.

Performance Assessment to Support Meaningful Learning

New Hampshire Department of Education (NH DOE) is attempting to design a coherent accountability system to foster deep understanding of learners. Many current educational accountability systems have stated goals of promoting deeper learning for students to, among other goals, improve college and career readiness. The NH initiative is based on the premise that performance-based and related assessment approaches must be meaningfully incorporated into accountability systems if we are to do more than pay lip service to these policy goals. We rely on the following definition for performance assessment:

> Performance assessments are generally multi-step activities ranging from quite unstructured to fairly structured. The key feature of such assessments is that students are asked to produce a product or carry out a performance (e.g., a musical performance) that is scored according to pre-specified criteria, typically contained in a scoring guide or rubric.[1] In fact, the rubric is a critical component in establishing the validity of the score inferences since it is the bridge between the student work and the resulting score, the basis for the inference. Occasionally, performance assessments target key processes or skills, such as communicating to diverse audiences, engaging in critical thinking, and listening to multiple viewpoints, that students employ when wrestling with a problem or

---

[1] *Products* are sometimes thought of as a separate category of assessment form, but we argue that products are really one possible outcome or piece of evidence derived from a performance assessment.

participating in an event such as a debate or a mock presentation to a simulated (or real) city council. Like "authentic assessments," performance assessments suffer from definitional problems in that this one term can encompass many different types of assessments. For example, performance assessment can range from 15-20 minute tasks (i.e., quite short) to multi-day activities with many scoreable units (Marion & Buckley, in press).

This definition does not distinguish among traditional academic and more cross-cutting (e.g., critical thinking, problem solving) knowledge and skills, because the principles for assessment design and validation apply to the multiple assessment targets. Shepard (2000) and others have argued that high quality tasks and assessments provide teachers and students the opportunity to learn more about the content being assessed than they could from selected-response items. Additionally, good assessments, especially performance tasks in which students have to generate solutions and reveal and/or explain their thinking can provide opportunities for teachers to develop sophisticated understandings about the nature of student learning (see also NRC, 2014). Although such insights are not impossible to obtain with selected response items, they are more likely to emerge from examining student work associated with complex performance tasks.

Performance Assessment of Competency Education (PACE)

New Hampshire is committed to raising the bar for all students by defining college and career-readiness to encompass the knowledge, skills, and work-study practices that students need for post-secondary success including deeper learning skills such as critical thinking, problem-solving, persistence, communication, collaboration, academic mindset, and learning to learn. However, NH's educational leaders recognize that the level of improvement required cannot occur with the same type of externally-oriented accountability model that has been employed for the past 12 years. In fact, the state argues that the current system is likely an impediment for moving from good to great. The state is piloting an accountability system with significantly greater levels of local design and agency to facilitate transformational change in performance. As part of this shift in orientation, the state is supporting a competency-based approach to instruction, learning, and assessment contextualized within an internally-oriented approach to accountability to best support the goal of significant improvements in college and career readiness. The information learned through competency-based assessments would then be used to support accountability determinations and, hopefully, better inform school improvement (e.g., Hargreaves & Braun, 2013).

A competency-based system relies on a well-articulated set of learning targets that helps connect content standards and critical skills leading to domain proficiency. Such a system requires careful tracking of student progress and ensures that students have mastered key content and skills before moving to the next logical set of knowledge and skills along locally-defined learning trajectories. Current systems th at rely on compensatory systems (e.g. averaging) for grading and related record-keeping may allow students to slip through the cracks in terms of possessing necessary knowledge for building deep understandings in the focal disciplines.

The PACE system is designed to foster deeper learning on the parts of students than is capable under current systems. This requires timely assessments linked closely with curriculum and instruction. The PACE system is based on a rich system of local and common (across multiple districts) performance-based assessments that are necessary for supporting deeper learning as well as allowing students to demonstrate their competency through multiple performance assessment measures in a variety of contexts. Thus, the accountability option was established to enable schools and districts to demonstrate student achievement and learning

growth through means other than or in addition to standardized tests, with an emphasis on performance assessment.

In the PACE option, the New Hampshire Department of Education (NH DOE) has created a route for districts and schools to demonstrate quality not solely or primarily dependent upon state standardized tests. The creation of the PACE accountability option reflects NH DOE's belief that school accountability works best if the responsibility for design and implementation is shared by districts and the state, rather than top-down mandates. Known as "reciprocal accountability," districts and schools are responsible for determining and reporting on local accountability measures, while the state is responsible for support and oversights in helping districts establish strong accountability systems.

Finally, New Hampshire is committed to implementing a philosophically coherent system. If the State is encouraging districts to embrace student agency in determining learning goals, then it only makes sense for the State to embrace "district agency" in establishing its own accountability goals. In order to provide participating districts with "breathing room," NH DOE is negotiating an agreement with the United States Department of Education (USED) to limit state (or consortium) standardized testing to select grade levels (e.g., 4, 8, and 11). NH DOE is a strong supporter and governing member of the Smarter Balanced Assessment Consortium, but it argues that once per year assessments, as good as Smarter Balanced may turn out to be, are not enough to drive and support deeper learning. Further, NH DOE is concerned that having external, large-scale assessments at almost every grade will control the conversation and not allow the space for the competency-based reform to take hold. The current PACE model, described here, is not necessarily a fully realized competency-based accountability system. Rather, we are presenting a "transitional system" that incorporates expected requirements of federal/state accountability, but points the way to what a fully realized system would look like with a possible change in ESEA or other policy changes on the federal level

Implementation Plan

It is one thing to put forth a proposal for a richer approach to education, but it is another thing to create the conditions necessary for successful implementation. NH DOE is engaged in a multi-faceted implementation plan to ensure the success of the PACE option that includes requirements for participating districts, technical and professional learning support, including task development and scorer calibration activities, and wrestling with complex technical issues. Clearly, NH DOE has not solved all technical, policy, and implementation challenges. Rather, this is an ongoing journey that NH has just begun. We describe below key aspects of PACE implementation in hopes that it might help others considering similar efforts.

Requirements for Participating Districts ("Guardrails")

Districts participating in the 2014-2015 pilot must have already adopted the State graduation competencies and developed a coherent and high quality set of K-12 course and grade competencies mapped to the State graduation competencies. These competencies were developed by teams of NH educators and approved by the NH State Board of Education. These districts must have demonstrated the leadership and educator capacity to participate effectively in the pilot. In addition to having a well-articulated set of competencies, these districts must have developed or be close to completing the development of a comprehensive assessment system tied to these competencies. Districts considered for the 2015-2016 pilot must have adopted graduation competencies and have a commitment during 2014-2015 to fully build out

their course and grade competency systems in K-12 as well as their comprehensive assessment systems.

Participating districts must be willing to participate in a peer and expert review process where they submit their systems of performance-based assessments for evaluation based on clear and rigorous criteria including alignment with state standards and competencies, consistency and accuracy of scoring, and fairness to all test takers. Further, PACE districts will be required to administer the state summative assessments (Smarter Balanced) in at least three grades, one at each level (e.g., 4, 8, and 11), which will serve as both an internal and external audit regarding school and district performance (see Table 1 below). Local districts will be expected to incorporate the results of the Smarter Balanced assessments in their local accountability systems.

All pilot districts are expected to fully participate in the development and implementation of the pilot accountability requirements such that all pilot districts will have the same general assessment requirements in the same courses and grades. As noted above, the Smarter Balanced summative assessment will be administered in select grades. The current plan involves staggering the Smarter Balanced subject areas according to when the results will be most useful for informing programs and auditing the local and common performance assessments. The current state science assessment (NECAP) will be phased out as these districts play a lead role in beginning to pilot "next generation" science assessment tasks. In fact, the National Research Council advocated in a recent report that moving to assessments of the Next Generation Science Standards must be led by classroom-based assessments rather than trying this complex endeavor with large-scale assessments first (NRC, 2014). The PACE districts will be particularly suited to pilot this new approach, given their intensive efforts in implementing complex performance assessments.

Table 1
Common summative performance-based assessments (PACE) and Smarter Balanced assessments administered by grade and content areas in all PACE districts

Grade	Competency Grading	English Language Arts	Mathematics	Science
K-2	✔			
3	✔	Smarter Balanced	PACE	
4	✔		Smarter Balanced	PACE
5	✔	PACE		
6	✔			
7	✔	PACE	PACE	
8	✔	Smarter Balanced	Smarter Balanced	PACE
9	✔	PACE	PACE	PACE
10	✔	PACE	PACE	PACE
11	✔	Smarter Balanced	Smarter Balanced	PACE
12	✔	⚜CAPSTONE PERFORMANCE ASSESSMENT		

Importantly, local performance assessment, used for competency determinations, will be administered in all subjects and grades. In certain grades and subjects, they will be "anchored" by Smarter Balanced assessment results, but in many others, they will be tied to performance assessments common to all participating districts. The competency determinations for all grades and subjects depicted above will include local (to each district perhaps) performance and other assessments designed to represent the full range and depth of the target competencies at each grade level. They were not depicted in Table 1 simply to avoid cluttering the chart. These common performance assessments (PACE) are intentionally limited to just one or two major tasks in most grade levels and content areas because NH DOE does not intend to simply replace one state assessment with another. Rather, these common performance assessments will be used to help calibrate performance expectations across participating districts and will be incorporated into local competency determinations.

The Task Bank

An ultimate goal of the PACE pilot is to enhance the capacity of educators to develop and use their own classroom assessments. However, creating a set of tasks for common administration and scoring purposes as well as helping to jumpstart local capacity is critical to the success of this project. The NH Task Bank is a repository of quality performance tasks that have been designed specifically to assess student attainment of the New Hampshire State Model Competencies. Additionally, the tasks in the NH Task Bank serve as models that teachers can use in their own assessment design work.

One of two key sources for performance tasks are those designed and submitted by New Hampshire teachers, most of who have participated in New Hampshire's Quality Performance Assessment Initiative over the past three years. These teachers received training in task design, quality assurance, analysis of student work and calibration. Tasks that are submitted to the NH Task Bank undergo a rigorous vetting and revision process. The NH task bank is organized according to content-specific competencies arranged along a developmental trajectory. The second key source of performance tasks is through the CCSSO's ILN Performance Assessment Project. The ILN project is collecting and curating a set of quality performance tasks that will populate an open-source, vetted task bank accessible to teachers. The emphasis of the work is on the type of performance-based measures that support assessment of deeper learning.

Professional Learning Support

The professional learning opportunities associated with PACE are embedded in the actual work of PACE, including task development, scorer calibration activities, system design, and peer review. The implementing schools established work groups, creating common developmental competencies in the key content areas aligned to the state graduation competencies as well continuing to build the state task bank. Sharing and analyzing student work is the core of any meaningful professional learning activity, therefore a key aspect of such learning opportunities for PACE teachers involves learning how to carefully analyze student work using established protocols to engage in common scoring sessions designed to foster consistent and accurate scoring of complex tasks.

Technical Issues and Considerations

In order for this reform initiative to be credible to New Hampshire stakeholders and to satisfy USED requirements, NH DOE is focused on ensuring the technical quality of the PACE system. Some of the key technical challenges include: creating comparable annual determinations, documenting longitudinal student progress (growth), measuring and reporting

the performance of key student groups (equity), and establishing systems for the effective use of assessment and accountability results (utility).

Comparability of Annual Determinations

One of the major challenges with the PACE pilot accountability system is ensuring that students from all NH schools receive meaningful opportunities to learn the required knowledge and skills. One of the ways to evaluate these opportunities is to require all students to participate in the same assessment of the same knowledge and skills. But it is not the only way. There are many examples, both with educational programs and outside of education, where we recognize that the "same" is not the only way to define comparability. For example, consider students applying for a competitive music program. Students will play different songs, perhaps using different instruments, but judges will have to determine who should be admitted to the program. We accept that judges are able to weigh the different types of evidence to make "comparable judgments." Why do we accept this? Because we have great trust in expert judges and their shared criteria. When the criteria are not explicit and applied systematically, then people have concerns (remember some of the Olympic figure skating fiascos in past years).

True psychometric comparability (i.e., "interchangeability") across districts administering different systems of assessment cannot be assured. In fact, it is not expected. However, NH DOE is taking important steps to ensure that students in pilot districts receive a high-quality education that meets or exceeds the expectations for non-pilot districts held to the same high expectations. For example, students deemed proficient in a particular grade or content area likely should be considered proficient regardless of the type of assessment.

Comparability efforts should not be focused on individual assessments administered throughout the year, rather the focus of comparability must be on the annual determinations of "proficient," "on-track," "competent," or any other label. NH DOE has proposed an approach to do just that. The Smarter Balanced achievement level descriptors (ALDs) are the basis for establishing cutscores on the Smarter Balanced assessments (this process was recently completed). The ALDs serve as the narrative descriptions of performance and the role of the standard setting panelists is to match the narrative descriptions with actual performance on the test. Therefore, NH DOE has decided to require all PACE districts to anchor their annual determinations of proficiency (competency) to the Smarter Balanced ALDs for the respective grade level and subject area.

Of course, it is one thing to use common descriptors, but having assessment evidence to evaluate against these descriptors is another critical component of comparability. Therefore, all PACE districts have agreed to participate in a common standard setting process based on thoughtfully-identified set of summative competency assessments administered throughout the year along with the common summative PACE performance assessment. Participating in a common standard setting process, where student work is compared with the ALDs will allow for comparably rigorous achievement standards to be established in all PACE districts.

To audit the extent to which the intended comparability has been achieved, NH DOE will rely on the results of the Smarter Balanced assessments in math and ELA in at least three grades and NH DOE is closely examining the Smarter Balanced interim assessments to replace or augment current local benchmark assessments to support comparability while raising the level of performance expectations. These common state assessments provide both an internal and external audit for locally-designed systems of assessment, evaluating the degree to which student performance on the local performance assessment system relates to performance on the statewide assessments. Discrepancies between local and state/consortium assessment results do

not mean that the local results are wrong. Rather, it should lead to conversations and inquiries to try to understand the reason for any large differences between the two sets of results.

All districts participating in the PACE pilot will be expected to participate in a peer review process during the first two years of implementation in order to examine their system design, assessment results, and annual determinations. Peer review will be structured to provide support and technical assistance to districts to ensure that local systems maintain high quality.

Lastly, NH DOE is taking steps to ensure scoring comparability by promoting accurate and consistent scoring of performance assessment tasks across classrooms, schools, and districts. NH DOE will sponsor Professional Development Institutes, including summer and school-year Quality Performance Assessment institutes on assessment literacy, competencies and designs for teaching them (knowledge, skills, and dispositions), assessment task design and validation, scoring calibration, and data analysis to track student progress and inform instruction. Regional task validation sessions will be conducted to assist districts in fine-tuning assessment tasks to ensure they measure target knowledge, skills, and dispositions. Regional calibration scoring sessions will be conducted to build inter-rater reliability and consistency in scoring across districts. These sessions are designed to build expertise among a core group of participants who can then lead task validation and calibration scoring sessions at the local level.

Equity

The competency-based educational system at the foundation of this pilot is, by design, more equitable because educators focus on the learning needs of every student and do not allow any students to fall through the cracks. That said, the state will continue to aggressively monitor and report the performance of student groups as outlined in New Hampshire's approved ESEA waiver. In addition, districts participating in the PACE pilot will be subject to additional examination of student group performance through their required participation in a peer review process to evaluate aggregate and student group performance results.

Student Progress

Student Learning Objectives (SLO) continue to be the main component of NH's educator evaluation system for all NH districts. This was the clear intention of the NH Task Force on Effective Teaching (NH DOE, 2013). The state believes that it can successfully document changes in student learning while supporting positive changes in local assessment and instruction. Pilot districts, because of the improvements in their assessment capacity, will be able to produce higher quality SLOs than most NH schools and districts. Therefore, the question should focus more on can pilot districts produce valid educator evaluation results and less on specific (and distal) approaches for calculating current achievement conditioned on prior achievement (e.g., SGPs, VAM).

NH has been using Student Growth Percentiles (SGP, see Betebenner, 2009) for school accountability purposes for many years and plans to support districts in incorporating aggregate SGP results into educator evaluations starting in the 2015-2016 school year. The NH Task Force on Effective Teaching recommended not attributing SGP results to individual teachers, unless the district's specific evaluation plan requires such use. The Task Force recommended, and NH DOE agreed, that aggregate SGPs must be used at least as part of a "shared attribution" approach according to a district's (or school's) theory of improvement (e.g., grade-level or content area teams). This is an important distinction because a similar—but not exactly the same—model can be applied in the PACE schools. In other words, NH proposes to use Smarter Balanced assessments at select grades to calculate SGPs and use the results aggregated at the

school level. These school-level results can be used to audit the individual SLO results and compare the "growth" of students in the pilot schools with other schools in the state.

Utility

Henry Braun stated that utility is the most important technical criterion by which we should judge the quality of accountability systems (Braun, 2012). Utility refers to the degree to which the policy/accountability system is able to support its intended aims. In the case of PACE, this would mean that the accountability system provides structure and information to help transform educator practices and deepen student learning. Focusing on utility changes the accountability conversation from one of labeling and sorting to one focused on using the results to bring about desired improvements in schools and student learning (Hargreaves & Braun, 2013).

Discussion

The purpose of encouraging schools and districts in this type of reform effort is to connect deeper learning at both the individual student and institutional levels. If students are expected to more fully engage in deeper learning, requiring them to follow a lockstep approach to learning runs counter to the research base. Similarly, if schools are going to support deeper and more flexible learning for students, then it appears incoherent for states to dictate to schools and districts performance expectations for students.

NH DOE originally conceived of an accountability system where districts were identifying their own goals and designing their own programs, indicators, and evaluation system. However, one of the most important things we are learning is that the cross district collaboration is a better professional learning structure than almost anything the state (or individual districts) could have supported on its own. Therefore, instead of a long-term goal where districts design locally-tailored systems, having districts join networks of districts focused on similar goals seems to be more effective and sustainable strategy. We also note that PACE is an incremental improvement over past practice, due to both the current USED regulatory requirements and state and local capacity.

The State is not blind to well-known challenges with implementing performance assessments as part of accountability systems as well as with the challenges of building the local capacity necessary for raising the level of student learning, improving local performance assessments, and supporting local accountability determinations. The State is not attempting to meet the levels of standardization and psychometric specifications associated with a state-controlled assessment and accountability system (e.g., AERA, APA, NCME, 2014). NH argues that the theories of action for such systems are impoverished with little evidence that such state-led systems bring about the levels of student and organizational learning the NH DOE would like to see. Rather, NH DOE is willing to engage in the challenge of supporting local capacity and agency in order to bring about transformational changes in student learning.

The State's major concern is scaling such efforts to all NH schools. The current PACE accountability system, even if wildly successful, is based on a voluntary proof of concept pilot with high-capacity schools. Improving chronically low-performing schools will be an enormous challenge. The State is committed to supporting the development of local leadership and capacity to help low performing schools implement the PACE system with fidelity. However, there are no illusions that this will happen overnight. In fact, the networked approach supported through PACE and other NH reform initiatives is likely the only viable strategy for bringing PACE to scale. This would involve growing this reform at a rate that can be managed and

supported, while continuing to focus on building local expertise as part of regional and statewide networks. Again, NH DOE does not assume that implementing a reciprocal accountability will be easy or smooth, but is committed to employing an approach couched in research on individual and organization learning to realize the deeper learning for students envisioned by many NH stakeholders.

References

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.

Anderson, J. R., Greeno, J. G., Reder, L. M, & Simon, H. A. (2000). Perspectives on learning, thinking, and activity. *Educational Researcher, 29, 4*, 11-13. http://dx.doi.org/10.3102/0013189X029004011

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending discourse. *American Educational Research Journal, 14,* 367-382. http://dx.doi.org/10.3102/00028312014004367

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, Vol. 28(4), 42–51. http://dx.doi.org/10.1111/j.1745-3992.2009.00161.x

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.

Braun, H. (2012). Using student work for educator accountability: The road from authenticity to validity. Presentation at the Reidy Interactive Lecture Series. Retrieved from www.nciea.org on September 30, 2012.

Council for Chief State School Officers (2011). *Roadmap for next generation state accountability principles*. Retrieved from: www.ccsso.org/Resources/Publications/ Roadmap_for_Next-Generation_State_Accountability_Principles.html

Domaleski, C. & Hall, E. (2013). *A vision for improved education accountability systems*. Council for Chief State School Officers.

Hargreaves, A. and Braun, H. (2013). *Data-driven improvement and accountability*. National Education Policy Center. Retrieved on October 23, 2013 from: http://nepc.colorado.edu/publication/data-driven-improvement-accountability.

Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511815355

Marion, S. F. & Buckley, K. (in press). Design and implementation considerations of performance-based and authentic assessments for use in accountability systems. In Braun, H. (ed.). *Meeting the challenges to measurement in an era of accountability*. NCME Book Series.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Committee on Developing Assessments of Science Proficiency in K-12. Washington, DC: The National Academies Press.

New Hampshire Department of Education (2013). The New Hampshire Task Force on Effective Teaching: Phase II. Retrieved from: http://www.education.nh.gov/teaching/documents/phase2report.pdf

Palincsar, A. S. & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and monitoring activities. *Cognition and Instruction, 1,* 117-175. http://dx.doi.org/10.1207/s1532690xci0102_1

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

Schmidt, W. H., McKnight,C. C., & Raizen, S. A. (1997). *Splintered vision: An investigation of U.S. science and mathematics education.* Norwell, MA: Kluwer Academic Publishers.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29, 7*, 4-14. http://dx.doi.org/10.3102/0013189X029007004

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental process.* Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language.* Cambridge, MA: MIT Press.

Wertsch, J. V. (1991). *Voices of the mind: A sociocultural approach to mediated action.* Cambridge, MA: Harvard University Press.

About the Authors

Scott Marion
National Center for the Improvement of Educational Assessment
Smarion@nciea.org
Scott Marion is the Vice-President of the National Center for the Improvement of Educational Assessment where his current projects include developing and implementing reform-based educator evaluation systems, serving as a technical expert for numerous state assessment and accountability systems, and designing and implementing high quality, locally-designed performance-based assessments.

Paul Leather
New Hampshire Department of Education
Paul.Leather@doe.nh.gov
Paul Leather is the Deputy Commissioner of Education for the New Hampshire Department of Education where he has been leading the competency education efforts for many years. He is also a key member of CCSSO's Innovative Laboratory Network helping to shape a national vision for competency-based and deeper learning.

About the Guest Series Editor

Linda Darling-Hammond
Guest Series Editor
Stanford University
ldh@stanford.edu
Linda Darling-Hammond is Charles E. Ducommun Professor of Education at Stanford University where she is Faculty Director of the Stanford Center for Opportunity Policy in Education. Her latest book is *Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning* (Wiley, 2014).