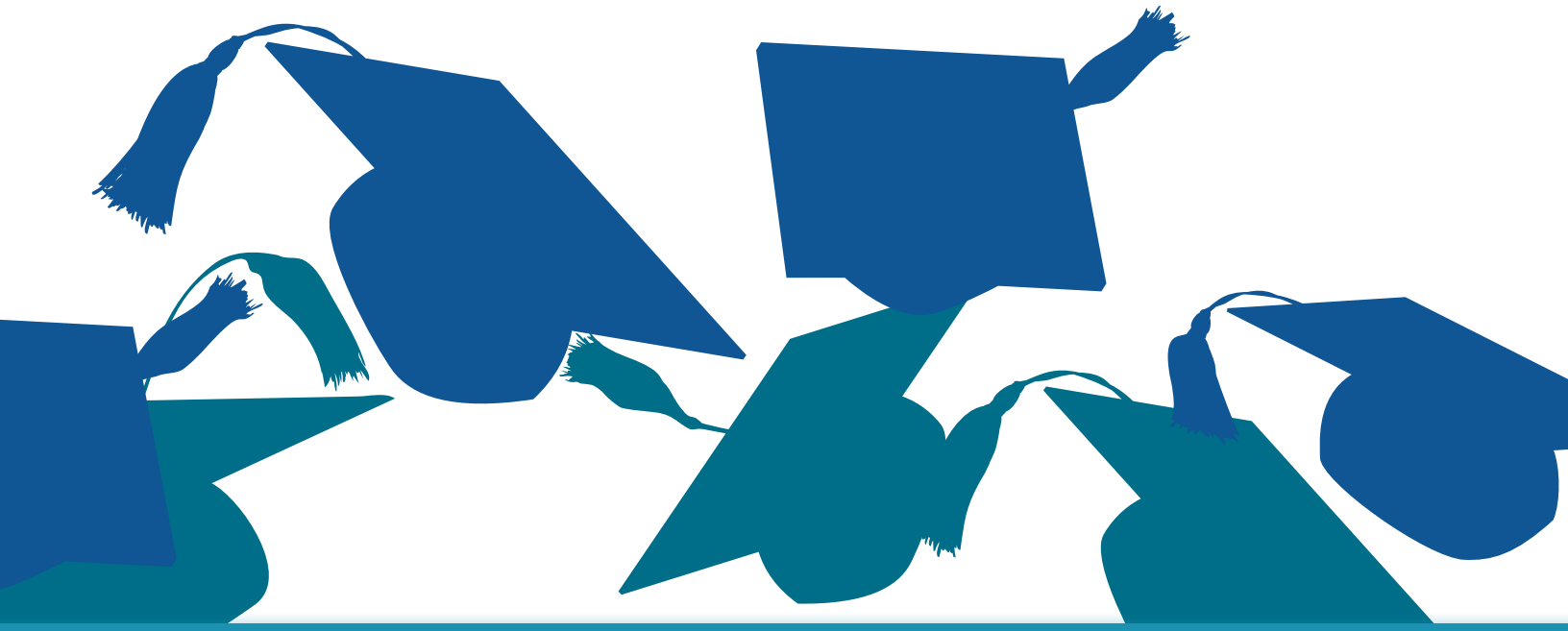


Assessment to Support Competency-Based Pathways

Achieve

Center for Assessment

Table of Contents

Assessment to Support Competency-Based Pathways

Introduction

Across the nation, state and local leaders have embraced two critical goals for public education: quality and equity. Students should be better prepared than they have been in the past; the goal now is for all students to graduate from high school ready for college and career. All students, regardless of race, economic background or geographic location, should reach or exceed college and career readiness. A number of states, districts and schools have determined that competency-based pathways (CBP) are a promising way to meet these goals. Focusing on competency requires students to demonstrate mastery on all essential knowledge and skills rather than merely attain lower levels of understanding. Providing pathways acknowledges that to ensure equity, students need personalized instructional supports and may reach and demonstrate their mastery in different ways.

A core tenet of CBP is that accurate, credible and useful student assessment information plays an essential role in supporting learning. **Assessments should be a meaningful learning experience for students, provide rich information to educators so they can provide targeted support to students, and send students and parents clear signals about students' readiness for next steps.** Assessment systems should be designed to help monitor the quality and consistency of determinations about whether a student is ready to move on and signal rigorous definitions of mastery or proficiency. Assessments should vary enough to reflect the personalization inherent in different pathways yet be comparable enough to ensure quality. Students should be able to take assessments as soon as they are ready, rather than wait for the end of the year, so that they can move ahead when they need to do so.

What Characteristics Differentiate Assessment that Supports CBP?

In a competency-based system, like in a traditional education system, assessments have a variety of uses such as to inform instruction, diagnose student needs and communicate student progress and outcomes to families and broader communities. Assessment in a competency-based system has an additional use – to help validate the determinations that teachers make about student proficiency or mastery of standards.

Additionally, assessments that support CBP often have these characteristics:

- Allow students to demonstrate their learning at their *own point of readiness*
- Contribute to student learning by encouraging students to *apply and extend their knowledge*
- Require students to *actually demonstrate* their learning
- Where possible, provide flexibility in *how* students demonstrate their learning (e.g. through a presentation, research paper or video)

Yet, little practical guidance is currently available for those designing assessments associated with CBP systems. For example, can the basic principles and practices of assessment design be applied to CBP systems without modification? What is it about CBP that makes the design of comprehensive assessment systems challenging? Are there unique considerations for validating CBP assessment systems? This document addresses the role of summative assessment in supporting CBP, clarifies key assessment challenges, and provides examples and recommendations that will be useful in guiding those who wish to design and implement assessment systems to support CBP.

Foundations of CBP

For the last 40 years, there has been ongoing interest in competency-based education from many quarters. In 2011, the International Association for K-12 Online Learning (iNACOL) and the Council of Chief State School Officers (CCSSO) sponsored a national summit to develop a common, working definition of "competency-based pathways." Soon thereafter, Achieve launched a working group of states and national organizations to explore the essential state policy decisions for graduation requirements, assessment and accountability that would lead to a CBP system that supports all students reaching and exceeding college and career readiness. The Achieve group adapted the working definition into the following points.

A Starting Point for Defining CBP

The following description, adopted by Achieve's Competency-Based Pathways Working Group, is adapted from the iNACOL/CCSSO working definition of CBP (Patrick & Sturgis, 2011):

- Students advance upon demonstrated mastery.
- Competencies include explicit, measurable, transferable learning objectives that empower students.
- Assessment is meaningful and a positive learning experience for students.
- Students receive rapid, differentiated support based on their individual learning needs.
- Learning outcomes emphasize competencies that include the application and creation of knowledge.
- The process of reaching learning outcomes encourages students to develop skills and dispositions important for success in college, careers and citizenship.

Conceptions of Competency

The term "competency" has been used in two distinct ways in the educational and psychological literature. The first use of "competency" is a state of being, of having ability. The other use of "competency" is the definition of the "things" in which a person could have demonstrated a set of knowledge and skills — "competency in writing," "competency in mathematical problem-solving" and so on.

Both of these meanings of "competency" are in play today, and both have important implications for the definition, assessment and eventual use of competency determinations. Achieve's working definition of CBP incorporates both of these perspectives: There is attention to specified knowledge, skills, abilities, and dispositions that can be made explicit and measured, specified and demonstrated with mastery.

Some Important Dimensions of Definitions of the Nature of Competency

- What are the abilities or "things" in which a person should be skilled and/or knowledgeable? For example, what competencies do all students need for college and career readiness?
- Given the educational purpose, what level of specificity is needed to define these abilities or "things" – from smaller building blocks to larger, integrated units?
- To what extent will they focus primarily on academic knowledge and skills, or go beyond to other cognitive, metacognitive, or non-cognitive skills?

Scope: Building Blocks or Larger Units?

Considering the scope or grain size of the competencies to be assessed is also critical for the design of learning and assessment systems, depending upon the purpose and use of the assessment under consideration. For example, are competencies to be assessed the same as individual content standards? Sometimes the purpose and use of the assessment may require smaller, more specific units of knowledge or skills aligned to the standards. For example, the content standard may cover understanding and being able to do addition of whole numbers up to 100. However, the instructional course may break up learning this standard into smaller chunks, so an instructor might want to know that the student is competent on a smaller chunk before going on to the next one.

CBP that focus on deeper learning may also often involve larger units — typically involving multiple standards. In this case, the purpose and use of the assessment requires measuring and understanding how students have integrated or applied several aspects of knowledge and skills to produce a performance that goes beyond a single content standard. This view often focuses on the "big ideas" or "enduring understandings" (e.g., Wiggins & McTigue, 2005) of the academic content areas as the learning goals. For example, a good mathematical assessment task might ask a student to solve a multistep problem that requires the integration of standards from more than one conceptual category.

Academic Knowledge and Skills or More?

Another aspect is to assess standards and competencies that combine knowledge and skills beyond the academic knowledge and skills typically found in a state's academic content standards. This can be a strategy to help all students reach or exceed college- and career-ready standards and expectations, assuring that diverse learners advance through their course of study with the full array of knowledge and skills they will need to be prepared. This view may reinforce the focus of assessment on application — demonstration of skills such as working with others to complete a complex task and/or performance in a specified or "real-world" context such as choosing a project topic rather having the teacher assign it or communicating the results to a nonschool audience. A focus on integration of procedure and understanding is a key part of the Common Core State Standards (CCSS) Standards for Mathematical Practice and therefore a critical part of CCSS-aligned assessments. Achieve's working definition of CBP includes attention to these skills: *The process of reaching learning outcomes encourages students to develop skills and dispositions important for success in college, careers and citizenship.*

These varying conceptions of competency highlight their importance in the design of learning and assessment systems. One can imagine very different types of assessment designs depending on the intended grain size of the standards or competencies measured and the extent to which they require demonstration of more applied, integrated, real-world learning.

Model Illustrations

The model illustrations on the following pages are intended to provide additional clarity about how the intended purposes and uses of assessments in a CBP model lead to design choices, and to make potentially abstract concepts more concrete. However, these illustrations are not intended to set apart any specific approach as best practice or exclude another potentially strong approach.

Illustration 1: Flexible-Pacing Model: A Standardized Assessment Solution

Background

Creekside School District (CSD) has developed a number of mathematics courses to support the state's adopted college- and career-ready academic standards. The state requires all students to take and pass a summative test at the end of grade 10 that covers standards typically encountered in algebra and geometry courses. Like many districts, CSD offers flexibility for the type, timing and sequence of math courses. For example, some students may take geometry in 9th grade, while others take it in 10th grade.

System leaders were troubled that many students were not successful on the summative assessment in grade 10, and they lacked good data to understand where or why academic preparation was deficient. Moreover, algebra and geometry teachers in the district reported that many students began their courses lacking important prerequisite skills, while other students were very well prepared and could access more rigorous content if the opportunity were available.

Proposed Approach

CSD has resolved to create a series of eight curriculum and assessment units covering the content spanning the two courses: algebra and geometry. Instructional resources will be developed for each of the eight units culminating in an assessment. Each student will work with his or her teacher to complete the units at his or her own pace. When the unit is complete and the teacher determines it is appropriate, the student will take an assessment to determine if he or she is ready to move on to the next unit.

The units are intended to be used flexibly to fit the scope and sequence of common courses at CSD. For example, units A, B, C and D might cover the content in a traditional algebra course, or units A, B, E and F might be suitable for an integrated algebra and geometry course. Schools will be able to incorporate the units with existing courses or design new approaches with the units. Importantly, there will be no seat time requirement. Rather, successful completion of the unit series associated with a course will be deemed sufficient for awarding course credit. Moreover, students completing all eight units should be well positioned for success on the state's grade 10 summative assessment. Ultimately, district and state leaders would like to study the feasibility of relying on student outcomes from the math units in lieu of a state summative test. That is, if a student successfully completes all eight unit assessments, he or she would be exempt from the state summative assessment.

Implementation

CSD started by assembling a committee of expert math educators, who determined the content for each unit. Educators grouped the standards into related domains deemed appropriate for an instructional unit and then identified the key learning targets for each unit. Then, educators identified the building blocks for achieving the learning target. They did this by unpacking the standards and identifying the specific knowledge and skills that should be emphasized in the unit, considering the prerequisite knowledge to support subsequent concepts.

CSD has determined that the primary purpose of each end-of-unit assessment is to produce a classification that indicates whether a student's level of performance is sufficient to move on to a subsequent unit (e.g., pass/fail). Additionally, CSD leaders want the outcomes to be comparable. That is, every time a student takes a different end-of-unit test, the level of performance required to pass should be the same. Given that students will be taking the tests at different times and allowing for retests, the system will need several equivalent forms for each unit test.

CSD determined that additional capacity would be necessary to design and implement a very robust and standardized assessment system. The district sought external support for development and operational tasks to include:

- Developing blueprints for each test based on the units defined by the committee, including both selected-response items and performance tasks;
- Developing form specifications such that each test will be well designed to support a "move-on" classification;
- Conducting item writing and review workshops with CSD teachers to ensure a sufficient quantity of high-quality items will be developed for multiple forms each year for each unit;
- Implementing a process to determine an appropriate cut score for each test;
- Producing statistically equated scores to ensure that the performance expectations will be comparable across multiple forms;
- Developing a scoring rubric and a standardized, defensible process for CSD teachers to evaluate performance tasks; and
- Facilitating computer-based delivery, scoring and reporting.

Illustration 1: Flexible-Pacing Model: A Standardized Assessment Solution

CSD allowed for a year of development work and another year for pilot testing before operationalizing the tests in year three. A process is described for ongoing development and monitoring to ensure that the system is functioning as intended each year and refined as needed.

Discussion

This illustration is purposefully highly standardized and requires a substantial investment from the district to implement well. As such, this illustration may represent an extreme on a continuum of CBP alternatives. However, such an approach is largely aimed at increasing confidence in the comparability of scores. This model offers little to promote deeper learning nor does it describe a process likely to provide rich diagnostic feedback. Rather, the emphasis is on uniform measurement and classification precision. It is certainly possible to conceive of other designs in which the assessments are characterized by less uniformity in design and scoring, which may be appealing if strict comparability of outcomes is not essential.

Design and Implementation Considerations

In this section we address specific design and implementation considerations for assessments in a CBP model. These considerations are linked to selected questions in Achieve's CBP *State Policy Framework* (2013), which presents a number of key issues that should be addressed in developing an assessment model that incorporates CBP. In particular we address the following questions:[1]

1. How is content represented on the assessment?
2. How is the assessment administered?
3. How is assessment quality evaluated and sustained?
4. How are results reported and used?

Because design is inextricably connected to the purpose and uses for the assessment, we will use an example scenario to consider these questions for an assessment system intended to determine whether students have attained the knowledge and skills needed to move on during and at the end of a particular course/class. Please see Illustration 1: Flexible Pacing Model (page 4) to review this scenario.

How is content represented on the assessment?

A distinguishing characteristic of the CBP approach is that teachers, working with students, are given flexibility to determine the pacing and, potentially, the sequence of content. Because students are learning content at different paces and at different times, this approach requires a departure from the more typical practice of administering a single cumulative test for all students at one point in time (e.g., end of course or end of year) and requires careful thinking about how content is defined and sequenced. At the outset, ensuring that the assessments align to and reflect the full set of the state's standards within a grade/subject or course is critical. The process often starts with grouping content into competency-defined modules based on learning objectives associated with coherent domains from the superset of course competencies and determining the scope and structure of the individual assessments.

[1] Achieve's full framework is organized by graduation requirements, summative assessment and accountability. Please refer to this document for a more complete treatment of the range of issues that should be considered. The focal questions addressed here combine some elements and draw from multiple sections to provide a concise design illustration of two selected approaches.

As discussed in the previous section, scope refers to the grain size of the competencies. One approach is to create assessments for groups of standards. For example, one could group like standards by domain (e.g., functions and measurement) or select an individual domain and divide it into smaller blocks (e.g., functions subdivided into linear functions, quadratic functions and so forth). Depending on the intended purpose and design of the assessment (particularly for performance assessments), grouping standards in such a way as to connect to "big ideas" and enduring understandings such as mathematical processes and applications may be desirable.

Structure refers to the order and desired overlap of the modules. The assessment modules can represent discrete content groupings intended to be taught and tested in a prescribed order or in an order of the teacher and/or student's choosing. Alternatively, the assessments can be organized by spiraled content such that each module requires increasingly complex demonstrations of similar tasks. For example, a task requiring a student to construct a written response to literature may be grouped into modules based on increasing text complexity.

To determine the nature and sequence of the assessment modules, the system should be designed to represent a learning model that specifies a pathway for how students build content expertise. For example, Hess has developed research-based learning progressions for the CCSS in mathematics (2010) and English language arts (2011).[2] This approach assumes that students will encounter the content and be assessed in the prescribed sequence, although the pace may vary. Another approach is to have content experts group the learning targets by similar characteristics, allowing for different learning progressions. In this manner, the teacher and student may choose both the sequence and the pace in which the content and assessments are encountered. With either approach, a pilot program and ongoing evaluation can be helpful to evaluate the efficacy of the selected model.

Standard test development practices may be used to design items and tests in the CBP model. While each assessment may focus on a limited set of learning targets or competencies, representing the depth of the construct is important and is likely to require items and tasks that go beyond multiple choice. An approach such as evidence-centered design is useful to help develop items and/or tasks that are likely to elicit target demonstrations (Mislevy, Steinberg & Almond, 1999). Such an approach is particularly important in an assessment that is used to determine whether students have attained mastery. Such an assessment must include items and tasks that can elicit evidence to support inferences about the desired performance. Ultimately, the development decisions are typically codified in test specifications and test blueprints that clearly define the features of the assessment.

How is the assessment administered?

A central characteristic of CBP models is that students and educators dictate the pace of learning. Therefore, the timing and frequency of assessments must vary so they can be administered after the student has completed each learning unit. In general, we recommend assessments should be relatively robust with a variety of item types. For example, it would be reasonable to conceive of a system with four to six modularized assessments per course, with each module containing 30 or more score points derived from a variety of types of items or tasks. This example is illustrative and reflects the minimal threshold to represent the construct and have sufficient information to support a "move-on" classification with adequate reliability. In practice, the particular specifications for any module should be established by specific content and psychometric targets. For instance, if the content cannot be represented and/or if conditional standard error at the cut score is insufficient with 30 items, the length of the assessment may be adjusted.

Given that such assessments are likely to have a summative purpose, security and standardization of the assessments must be addressed. When a single or limited number of administrations are offered, a small number of standard forms

[2] Many scholars contend that mastery of content standards is not about accumulating a collection of skills and knowledge but about making meaningful connections among skills and knowledge and building domain-specific schemas over time. Further, while it is recognized that competence can develop along multiple pathways to reach the same understandings, research indicates that some pathways will be followed more often than others. These typical paths provide the basis for developing learning progressions (Hess, 2012). Unlike standards, learning models reflect the systematic consideration of interactions among the learner, the content and the context for learning (e.g., situational, sociocultural, nature of support/scaffolding), as well as the dynamic, cumulative outcomes of these interactions (Cameto, Bechard & Almond, 2012). This is at the heart of CBP models based on a theory of how students learn.

can be developed and administered simultaneously to mitigate security concerns. However, it may be necessary to offer on-demand administrations and multiple retests, which will necessitate numerous administrations at different times.

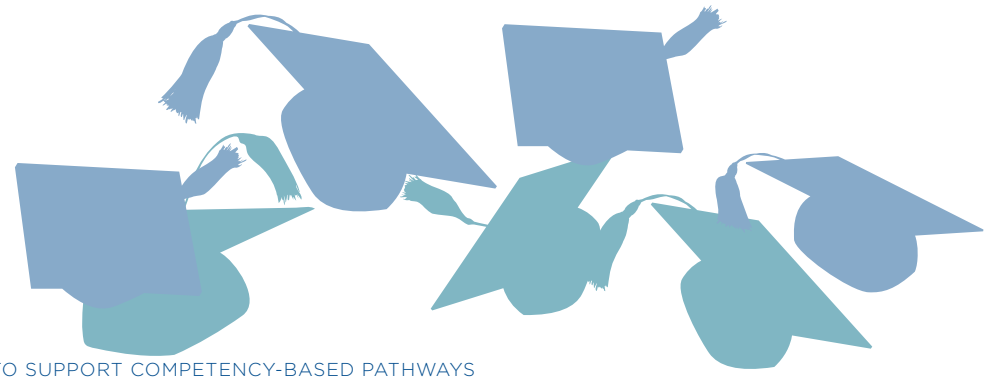
CBP assessment alternatives may vary along a continuum of standardization, with each option having its own advantages and limitations. A highly standardized approach would involve centralized production of a large number of equivalent forms or a bank of items that could be delivered using an adaptive or linear-on-the-fly algorithm. These options promote standardization and security and facilitate score comparability. On the other hand, this approach is typically time consuming and expensive to develop, given that it likely requires contracting with an external assessment developer. Moreover, assessments developed to conform to a single blueprint may not be appropriately personalized to each learner.

A less standardized approach involves creating assessments at the district or school level. For example, teams of educators may work together to develop a bank of assessment tasks and/or assessment modules for the desired courses. This approach is likely less costly and allows a high degree of customization. Standardized external scoring is often desirable to maximize consistency and security when the purpose of the test calls for it. This challenge can be overcome, however, by setting clear criteria for development and implementation and then systemically reviewing and monitoring the assessment against those criteria.

Another important consideration is how to handle scoring procedures for constructed-response items or performance tasks. Again, options exist along a continuum from standardized to flexible, each with tradeoffs. Standardized external scoring is often desirable to maximize security and consistency when tests are used for high-stakes purposes. However, this option is more time consuming and expensive. One potential alternative is use of a distributed scoring approach, in which trained, external raters evaluate student work transmitted electronically. The field of machine scoring (e.g., essay scoring software) continues to grow and may offer a promising alternative as well.

On the other hand, local scoring by the instructor or other qualified school personnel is a viable alternative that affords more flexibility and may be less operationally burdensome. Most important, local scoring, done well, greatly enhances the usefulness of assessment results for improving instruction and learning for students. However, if consistent score interpretations are desirable, measures such as developing scoring rubrics and procedures, implementing training and calibration processes, and conducting audits must be in place to bolster comparability. Additional benefits from these processes include professional learning and building capacity and depth of understanding of the standards and assessment.

When the purpose of measurement is to elicit demonstrations of more complex and/or cross-cutting competencies, assessment administration may look quite different from a traditional academic test. For example, One can imagine a biology course in which students are asked to demonstrate their understanding of biodiversity and ecosystem maintenance. By defining a relevant local ecosystem problem and engaging in the design process to design, evaluate, refine, and present an original solution, students have the opportunity to illustrate their understanding of science content through the application of knowledge and skills across both science and engineering domains. Another option is to establish rigorous criteria for students, educators and schools for the learning experiences and objectives that the student may flexibly satisfy to fulfill one or more course requirements. For example, students may receive guidelines that specify the criteria for the experience, the expected performance and how successful completion may be certified. The guidelines may allow students to select from several options (e.g., internship, conference presentation, science competition, service project), and each option may include the criteria for a successful capstone accomplishment and manner of certification.



How is assessment quality evaluated and sustained?

With respect to the quality of the assessments, a large body of literature regarding test development and validation is available, including the recently revised *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014). While we address validity in more detail in a subsequent section, there are four priority areas for ensuring quality:

1. **Content Representation:** The test items should represent what teachers should be teaching and students should be learning with respect to breadth and depth. Item/task development should be consistent with industry standards for best practice (e.g., committee review for bias), and independent validation of content and blueprints is essential.
2. **Precision:** When the primary purpose will be to classify the learner with respect to an established threshold, the assessment or, more likely, the set of assessments should provide the most precise information near the performance target — in contrast to providing deep information along the full ability range as might be useful for diagnostic information or growth.
3. **Administration/Scoring:** Tests should be administered and scored using well-documented procedures. Typically, quality assurance measures such as training and audits are recommended.
4. **Fairness/Accessibility:** Efforts are made to mitigate barriers so students can show what they know. For example, appropriate accommodations and supports for students with special needs are available. A particular advantage of CBP systems, in terms of ensuring fairness, is that essentially all systems provide students with multiple opportunities and formats to demonstrate mastery. Such an approach significantly diminishes the risks associated with any single assessment.

Illustration 2: K–12 Competency Education Model

Background

Granite State School District is located in a community of approximately 20,000 people in New England. The district has four elementary schools, two middle schools and one high school. It total, it serves approximately 1,200 students. The district leadership and the community were concerned that too many students were being moved along to the next grade and to graduation without really having the skills necessary to succeed at the next level. On the other end of the spectrum, the community was expressing concerns that high-achieving students were not being challenged appropriately.

Proposed approach

The high school graduation requirement in the Granite State School District decided that moving to a competency-based education model could address its challenges at both ends of the achievement continuum. The district adopted the state model graduation competencies and over the next several years worked in disciplinary and interdisciplinary teams both within and across grades to develop grade and course competencies. The district leadership wanted to focus instruction and assessment on a set of competencies that represent the big ideas of each discipline and tap deep levels of complexity and rigor associated with college- and career-ready skills. The competencies encompass the full range of the state's standards including important cross-cutting skills (e.g., written and oral communication, collaboration and problem-solving).

After considering a variety of approaches for providing learning opportunities for students, the district determined that the main organizing structure would be embedding competencies into instructional units. These units range between three and six weeks and are designed using an "Understanding by Design" model (Wiggins & McTigue, 2005) to support opportunities for deeper learning and allow for considerable opportunities for differentiating learning.

Implementation

The district recognized early on that performance-based assessments would be necessary to assess student learning of the competencies. Such performance-based assessments often serve as the end-of-unit assessment, but multiple assessment methods are used to evaluate and improve learning throughout the unit. Assessments are tied to competencies or components of competencies, but students are assessed summatively when they are ready to demonstrate competency. Remediation and/or reteaching and then reassessment are available for students who do not demonstrate competency at first.

Illustration 2: K–12 Competency Education Model

The units and tasks are developed and vetted locally but supported by a bank of vetted tasks and exemplar units that are developed by the state. Generally, the classroom teacher scores the performances to better provide feedback quickly to students, but the district decided it was important to create a local auditing (e.g., double-scoring) system as well. The units and tasks describe expectations for administration and scoring. However, standardization is not the focus. Rather, the district is using the performance tasks to build a credible body of evidence for each student to document the degree to which he or she has met the target competencies.

Competency determinations are intended to be comparable, although procedures for ensuring comparability are not always well specified. In this area the district is continuing to develop and fine-tune its approach. However, competency determinations for each grade and course are tied to performance-level descriptors associated with each grade. These grade-level determinations have been articulated to lead to readiness to succeed in college and careers.

The high school graduation requirement in the Granite State School District is to assemble of "body of evidence" over the course of high school that demonstrates that students have met the required competencies for graduation. The high school graduation competencies are similar to the previous K-12 course and grade competencies but include more cross-cutting and trans-academic skills to fill out the college- and career-ready profile. The body of evidence will necessarily be weighted toward the later years in high school because students usually need a little time before they are able to demonstrate competence in various areas. The district has adopted a sophisticated electronic portfolio system that enables it to keep track of students' work along with scores for each of the required competencies. Students are eligible to graduate once they have met all of the required competencies, generally through coursework but also through extended learning opportunities, and their full body of evidence is reviewed by a district graduation panel.

Discussion

This model was meant to illustrate a K–12 competency-based educational system focused on deeper learning of content and skills throughout and across the curriculum for all students. It culminates in a body of evidence graduation system to ensure that all students who graduate from the Granite State School District are well prepared for college and careers. This system focuses on ensuring rigorous demonstrations of learning tied closely to what students were expected to learn rather than standardization. However, the district has created systems of internal and external review to ensure that all students are being held to comparable expectations.

How are results reported and used?

In addition to design decisions associated with individual assessments, it is important in a CBP system to consider how multiple measures across assessments will be combined to produce an overall determination for a student (e.g., "mastery") and how that determination will be reported and used.

One approach is to combine tests conjunctively to provide an overall determination per tested domain. That is, each assessment in the series is used to provide a determination that the student has achieved a target level of performance (e.g., "mastery") for the competency or competencies in the tested domain. In this manner, the overall summative judgment is based on the student achieving the requisite performance level on every test in the series. Naturally, if passing each test is required to move on, the structure for this approach is in place by default. As discussed previously, this approach requires that each test is sufficiently robust to support a suitably precise classification. It should also be noted that this type of conjunctive use of multiple tests is atypical for most state end-of-course tests. A compensatory scoring model is normally used for a single summative test comprised of multiple domains, meaning that higher performance in one area can offset lower performance in another. If the conjunctive approach is used, it is likely to establish a higher bar (assuming each test's individual performance is comparable to the overall summative target) for earning credit in the course, which may or may not be desirable to policymakers. However, the conjunctive approach is conceptually coherent with a CBP philosophy in which students are not allowed to "get by" without mastering critical content and skills.

A second approach is to aggregate performance from each individual test into a composite score used for summative classification. There are various methods to accomplish this, such as joint IRT calibration or methods based on empirically or policy-weighted composites. Discussion of this topic is beyond the scope of this paper, but resources detailing approaches for combining multiple indicators are available (see, e.g., Wise, 2011).

A third approach is to use another end-of-course or end-of-domain test, such as the statewide summative assessment, to validate the summative decision. By so doing, the purpose of the along-the-way assessments is to signal that the student is ready to take the summative test. Either of the preceding two approaches could also be used to provide that information. If policymakers are concerned about comparability of assessments (e.g., are course credit decisions from schools using aggregated competency assessments comparable to those using a single summative test?), this approach addresses that concern. However, the increase in test burden may be prohibitive.

Validity Considerations

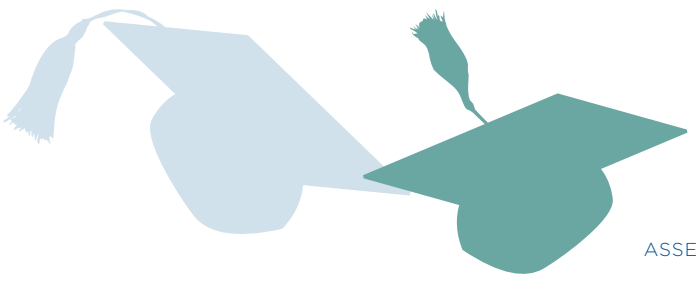
Most measurement specialists would say that validity is the most important criterion by which to judge the quality of an assessment or assessment system. In fact, the recently released *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014) emphasize the importance of validity in the evaluation of test quality.

Validity gets at the heart of test quality — does the evidence support the inferences about what a student knows based on the test score? The validity of an assessment or a system of assessments is always contingent upon the purposes of the assessment and the ways in which the assessment and/or system is intended to be used. Validating the inferences associated with CBP systems should follow closely from what we know about validating more traditional assessments, while emphasizing aspects that may be unique to CBP assessments. This section of the paper provides a brief overview of validity theory and practice while focusing on considerations for competency-based education systems.

Kane's (2006) argument-based approach is a helpful organizing framework for validity evaluation because it facilitates users prioritizing investigations and synthesizing the results of various studies to make judgments about validity. At its simplest, Kane's approach asks the evaluator to search for and evaluate all the threats to the validity of the score inferences. If these threats are not substantiated, the inferences drawn from the assessment results may be supported. Kane's framework assumes that the proposed interpretations and uses will be explicitly stated as an argument, or network of inferences and supporting assumptions, leading from observations to the conclusions and decisions. "Validation involves an appraisal of the coherence of this argument and of the plausibility of its inferences and assumptions."

Adopting Kane's framework does not mean ignoring more traditional sources of validity evidence as reemphasized in the most recent edition of the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), which includes evidence based on the following dimensions of validity.

- **Test content** is the degree to which the content and skills required by the test represent the knowledge and skills described by the learning targets (e.g., standards).
- **Response processes** are a description of the ways that students respond to items and tasks to provide evidence of the intended construct and not in ways that indicate that the item/task is tapping a different set of knowledge and skills than those described by the learning target.



- **Internal structure** is the extent to which the student responses to the items and tasks support the hypothesized claims and subclaims articulated for the assessment.
- **Relationships with other variables** is the expectation that items, tasks and assessments thought to measure a particular construct are more related to other known measures of that construct than they are to measures of different constructs.
- **The consequences of testing** are those intended and unintended uses of test scores that support or run counter to the aims of the program or other social goals.

These sources of evidence can be subsumed within an argument-based approach to the extent that the particular source of evidence is relevant to the inference being evaluated. Kane's progression from evaluating the veracity of the inferences based on scoring, generalization, extrapolation and decision fit particularly well for a CBP system.

Kane's (2006) Level of Inference	Applied to CBP
Scoring: From observed performance to the observed score	Does the design of a specific test or task, including the scoring rules applied (e.g., rubric), lead to accurate inferences regarding student performance on the learning target (e.g., competency or component of the competency)?
Generalization: From the observed score to the universe score	Each task or set of test items is theoretically drawn from an infinite set of possible tasks and test items. Even if not drawn from a truly infinite set of possible items, any task/item is only one of a set an enormous number of possible items and tasks from the "universe" of tasks/items. The question then relates to the degree to which the items/tasks experienced by the student generalize to all possible items/tasks measuring the specific learning target(s).
Extrapolation: From the universe score to the target score	This is an especially important inference for CBP systems. Competencies are purposefully subsets of the full domain of interest (e.g., biology), so it is critical to be able to substantiate that the set of competencies either represent the domain (i.e., the target) adequately — in a sampling sense — or represent the most essential elements of the domain so that the student will have critical knowledge for use at some future time.
Decision: From the target score to the evaluative determination	CBP systems are being designed and implemented to make determinations about whether or not a student is "competent" enough to either move on in a learning progression (very low stakes) or graduate from high school or achieve certification (high stakes). The lower stakes decision might not require collecting the evidence necessary to support the decision inference because one would not be making claims about the target score at this point. On the other hand, such evidence is necessary for higher stakes decisions such as graduation or certification.

Illustration 3: Graduation Distinction by Exhibition

Background

Mountain View School District (MVSD) has adopted rigorous academic standards associated with college and career readiness. Encouragingly, the academic performance of students as measured on state and national tests is typically very favorable. However, MVSD leaders have noted that many students moving on to college and careers report initial difficulties with applying their knowledge in a new context. For example, some students are struggling with skills such as organizing and leading work projects or designing and implementing a research study. As these skills are essential for college and career readiness, MVSD has resolved to provide opportunities for students to develop and receive feedback on their performance in this area.

Proposed Approach

MVSD decided to develop a capstone senior project for high school students, the successful completion of which will qualify the student to achieve a highly regarded diploma distinction. The project is intended to provide the student an opportunity to hone skills such applied problem-solving or innovative thinking in an academic context. The student selects and works with a faculty mentor, and together they identify a project that matches an area of interest for a career and/or postsecondary study. The student completes the project and prepares a report and exhibition for an evaluation committee, which makes the final determination regarding sufficiency of the demonstration.

Implementation

Districts leaders began by convening a committee to identify the competencies most important to elicit through this initiative. The committee was comprised of educators and leaders from MVSD in addition to representatives from the career/business sector and higher education. Committee members discussed the full range of competencies associated with postsecondary success and worked to identify the highest priority trans-academic skills that should be assessed. Prominent skills identified by the committee included written and oral communication, critical thinking, problem-solving, and contextual organization skills.

Next, a team of educators and system leaders at MVSD, including postsecondary leaders and employers, produced guidelines for demonstrations based on the committee's recommendations. The guidelines list the types of experiences or projects that are likely to elicit these trans-academic skills. The list includes: complete selected internships, design and implement a service project, develop and execute a research study, or design and teach a course or present at an academic conference, among other options.

The process calls for the student to submit a prospectus of the project to the faculty mentor prior to receiving approval. The process culminates with an exhibition based on the student's experience. Criteria are developed for the exhibition, which include preparing a written product that details the experience and includes metacognitive reflections about what the student learned. Additionally, the student is required to facilitate a presentation based on the written submission to a committee chaired by his or her faculty mentor. The committee may approve the exhibition or require the student to fulfill additional tasks to meet requirements.

Finally, MVSD set up a system to periodically follow up with the students who successfully complete the exhibition. MVSD surveys the students at regular intervals to understand the challenges and successes they are experiencing in their postsecondary pursuits to refine and improve the requirements of the exhibition.

Discussion

This illustration reflects a CBP assessment of high-level trans-academic competencies that are not easily measured on traditional academic tests. Purposefully, the student is involved in selecting and completing the project within the guidelines to promote motivation and relevance. Also, there is minimal standardization associated with the experience and exhibition. The goal is not to make statements about comparability (e.g., student A's exhibition was equal in rigor or quality to student B's). Rather, the goal is that every successful demonstration meets a threshold standard for rigor and quality judged sufficient to promote and exhibit key competencies associated with postsecondary success.

Scoring

The scoring inference is fundamental in that educators, students, parents and others are making inferences as a result of a particular score on a task or test and want to know if these inferences can be supported. For example, if a student receives a score of 3 on a four-point rubric for a mathematics problem, most would interpret that to mean that the student had reasonable control of the knowledge and skills represented by the learning targets that the problem was intended to measure. But if the rubric was poorly designed such that almost all students received a 3, whether or not they really knew the material, those inferences would not be warranted.

The scoring inference should be evaluated by collecting data in at least the following categories from the *Standards for Educational and Psychological Testing:*

- Test content;
- Response processes;
- Internal structure; and
- Relationships with other variables.

Generalization

The sources of evidence outlined above are critical for evaluating the quality of inferences from performance to scores. Generalizable inferences are focused on the degree to which a performance on a single assessment represents the student's knowledge and skills if we could have administered all possible assessments of the same learning targets under all possible conditions to that student. This is true for all types of assessments, but the use of performance tasks in measuring competencies — which is very appropriate — requires other sources of evidence, especially when one is concerned with generalization inferences. Psychometric techniques such as reliability and generalizability analyses allow us to estimate, based on administering a limited number of assessments to a student, the amount of uncertainty associated with each score. Said another way, such estimates of uncertainty help us describe the likeliness that a person's observed score is a good approximation of their "true score" (or "universe" score in generalization terminology).

Much has been written about generalizability, which is essentially the measurement analogue to learning transfer, especially in terms of performance-based assessment (e.g., Brennan, 1992; Cronbach, Linn, Brennan & Haertel, 1996; Shavelson, Baxter & Pine, 1992).

One of the challenges in conceptualizing the generalizability of inferences is to be clear about the target of generalization and the factors that would influence or threaten such generalizations. It is easy to get seduced by the apparent simplicity of terms such as "universe score" or similar characterizations, but thinking about it from the learning and instruction side leads one to ask questions about whether the results of one assessment or a small set of assessments provide credible evidence that the student really knows what is being claimed based on the assessment scores. This is a concern for all instruction and assessment programs, but it might be more of a concern for certain types of CBP systems that are designed to declare that students have demonstrated mastery of the competency and they can move on without having to demonstrate such mastery of the same competency again (e.g., on an end-of-year test).

When competencies are evaluated with performance-based assessments or other open-response tasks, legitimate concerns may be raised about the generalizability of student scores because of a the influence of the specific tasks used, the raters used to score the tasks, the occasions when the tasks were given and potentially other factors. Such "error" or uncertainty introduced by different facets can lead to considerable challenges to generalizability. This variability associated with tasks, raters and occasions can be evaluated using generalizability methods (e.g., Brennan, 1992), and the threats to generalizability can be ameliorated by ensuring that enough tasks are employed and that rater accuracy and consistency are monitored.

Extrapolation

One of the biggest challenges in any accountability system — and competency-based education is often used for student-level accountability — is determining approaches for aggregating and weighing evidence and then using the aggregated evidence to establish mastery or other identified performance levels tied to an inference of competence. We base the following discussion on the assumption that we can improve the measure of competencies, especially college- and career-ready competencies, with multiple assessments. Therefore, the results from multiple assessments will need to be aggregated in a defensible way to determine if students have achieved the required performance on the competencies.

Generalization is focused on evaluating the degree to which inferences based on scores from an assessment can support interpretations about how well the student knows and/or is able to do the specific learning targets (competencies) represented by the assessment. Extrapolation has to do with inferences from the specific learning target(s) to the larger domain of which that learning target is a part. For example, if assessments yield evidence of successful performance by students on learning targets having to do with classification, cell theory, genetics and scientific inquiry, one would like to conclude that the student is competent in high school biology. To substantiate such an inference, one would first have to determine the degree to which each assessment can support generalizability inferences and then evaluate the degree to which each assessment and the set of assessments adequately represent the target domain.

As discussed in the previous section, competency-based education systems have been designed, in part, to move away from compensatory-based assessment systems, in which as long as the student, on average, knows enough of a large domain (e.g., biology), he or she is not generally required to know enough of each important subarea (e.g., within biology) to be considered successful (competent). Most CBP systems organize the required competency by course, subject area, and/or interdisciplinary knowledge and skills, and while documenting students' mastery of specific competencies within each content area is valuable, most stakeholders care about whether or not students are competent at these larger levels. Therefore, validating extrapolation inferences should be an important consideration in evaluating the validity of CBP assessment systems.

Judgmental approaches that examine and map the content of each of the assessments to the target domain would provide useful information regarding the degree to which the set of assessments can support inferences about that domain. Additionally, the ways in which the individual assessment results are combined can to lead to more or less defensible inferences regarding both generalization and extrapolation.

Decision

Ultimately, CBP lead to evaluative decisions, and these decisions may have policy implications, illuminate barriers to the system that is envisioned, or require states and districts to anticipate potential consequences. These decisions can range from low-stakes decisions that are used to guide instruction to high-stakes decisions such as certification and graduation. The validity evidence needed to support low-stakes decisions such as moving on to the next learning target will likely have been provided already by the evidence described above for scoring and generalization. This is because there are few consequences for being wrong about such determinations. If a student is mistakenly advanced to the next learning target, teachers should quickly be able to discover and correct this error. Similarly, if a student is erroneously (or cautiously) held back from advancing to the next learning target, this error should become evident and corrected. There may be some loss of efficiency, but this loss is likely not very critical in the larger scheme.

On the other hand, some states or school districts with competency-based systems may have policies that include high-stakes decisions such as graduation or certification determinations. Being wrong about such determinations may have significant consequences for students. Given that students in a competency-based learning environment should have flexibility to demonstrate mastery through various methods of assessment, it is more likely that these decisions are being made using multiple assessment measures. States and districts, therefore, must consider how they will establish performance standards for making determinations of student competency across various methods of assessment. These decisions should occur at the overall determination level, but if such overall determinations (e.g., graduation) are based on conjunctive methods for

aggregating individual competency decisions, then deliberative approaches for establishing "good enough" (i.e., performance standards) must be employed as well. In more formal settings, standard setting is the process of converting written descriptions of performance into cut scores on an exam or other collection of quantitative data. If done correctly, it is a deliberate and systematic process designed to develop shared meaning among those selected to participate in the activity (the panelists) so they can establish cut scores based on a common understanding of performance. Much has been written about validity requirements for performance standards (e.g., Hambleton, 2001; Shepard, 1993) as a result of such standard-setting activities, and evidence substantiating overall competency determinations must be provided and evaluated.

Values and Consequences

Kane (2006) and others (e.g., Shepard, 1993) suggested that the evaluator must attend to values and consequences when evaluating a decision procedure such as when a testing program is used as a policy instrument, as is the case with essentially all state tests. When conducting such a validity evaluation, the values or uses inherent in the testing program must be made explicit, and the consequences of the decisions as a result of test scores must be evaluated (e.g., Cronbach, 1971; Lane & Stone, 2002; Linn, Baker & Dunbar, 1991; Messick, 1989, 1995; Shepard, 1997).

CBP assessments may be incorporated into larger accountability systems and used to serve policy aims, and the consequences of these decisions must be considered in validity evaluations. More directly, CBP instructional and assessment systems are used in a growing number of places to support graduation decisions, which certainly have consequences. Therefore, when evaluating the validity of a CBP system, stakeholders and evaluators must be particularly attentive to unintended negative consequences that may arise from denied or reduced opportunities for advancement.

Another aspect of consequential evidence relates to the larger CBP system and not exclusively to the assessment system. Part of a validity argument for a competency-based system should relate to the claim that meeting the competencies will lead to important outcomes (e.g., enduring understandings) for students because the competencies represent constellations of critical content, skills and dispositions. As noted throughout this section, the assessments should reflect this complexity, but the question of whether mastering the overall or particular set of competencies leads to the intended outcomes is one that needs to be addressed. Such questions are likely beyond a validity evaluation of a CBP assessment but should still be evaluated as part of an overall program evaluation of the CBP system.

Conclusion

CBP models represent a promising approach for improving equity and excellence toward college and career readiness for all students. To realize this promise, CBP require accurate and useful assessment information. Indeed, many education leaders look to CBP to help promote high-priority outcomes to prepare all students for success in college and career. We have outlined some of the key decision points as well as strategies for design and validation. Establishing quality assessment systems in a CBP framework is not without challenges, but good options exist. Using a competency-based approach can help provide a close linkage between the learning targets and assessments, but this coherence cannot be assumed. Designing assessment systems for CBP requires thoughtful considerations of such features as grain size, learning sequences, aggregation and decisions, and consequences. Again, none of these features requires a new conceptualization of assessment, rather existing measurement knowledge and procedures can be adapted to serve CBP systems.

References

Achieve. (2013). *Advancing Competency-Based Pathways to College and Career Readiness: A State Policy Framework for Graduation Requirements, Assessment and Accountability.* Competency-Based Pathways Working Group.

American Educational Research Association, American Psychological Association and the National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Brennan, R. L. (1992). *Elements of Generalizability Theory.* Iowa City, IA: American College Testing Program.

Cameto, R., Bechard, S., & Almond, P. (Eds.). (2012). *Third Invitational Research Symposium: Understanding Learning Progressions and Learning Maps to Inform the Development of Assessment for Students in Special Populations.* Manuscript in preparation. Menlo Park, CA, and Lawrence, KS: SRI International and Center for Educational Testing and Evaluation.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373–399.

Hambleton, R. H. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In Cizek, G. J. (Ed.). *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Hess, K. (Ed. & Principal author). (2010). *Learning Progressions Frameworks Designed for Use with the Common Core State Standards in Mathematics K-12.* National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment. Retrieved from: www.nciea.org/ publications/Math_LPF_KH11.pdf

Hess, K. (Ed. & Principal author). (2011). *Learning Progressions Frameworks Designed for Use with the Common Core State Standards in English Language Arts & Literacy K-12.* National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment. Retrieved from: www.naacpartners.org/publications/ela_lpf_12.2011_final.pdf

Hess, K. (2012). Learning progressions: Connecting ongoing learning to assessment. *Communication Points Newsletter*, 5. Lexington: National Center and State Collaborative, University of Kentucky.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence Centered Assessment Design.* Educational Testing Service. Retrieved from: www.education.umd.edu/EDMS/mislevy/papers/ECD_overview.html

Patrick, S. & Sturgis, C. (July 2011). *Cracking the Code: Synchronizing Policy and Practice to Support Personalized Learning.* iNACOL. Retrieved from: www.inacol.org/cms/wp-content/uploads/2012/09/iNACOL_CrackingCode_full_report.pdf

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21 (4), 22–27.

Wiggins, G. & McTighe, J. (2005). *Understanding by Design: Expanded Second Edition.* Alexandria, VA: ASCD.

Wise, L. (2011). *Combining Multiple Indicators.* Paper prepared for the Partnership for the Assessment of Readiness for College and Careers (PARCC) Technical Advisory Committee. Retrieved from: www.parcconline.org/sites/parcc/files/PARCCTACPaper-CombiningMultipleIndicatorsRevised09-06-2011.pdf

Acknowledgments

Achieve and the National Center for the Improvement of Educational Assessment would like to thank the individuals and organizations who contributed to this paper.

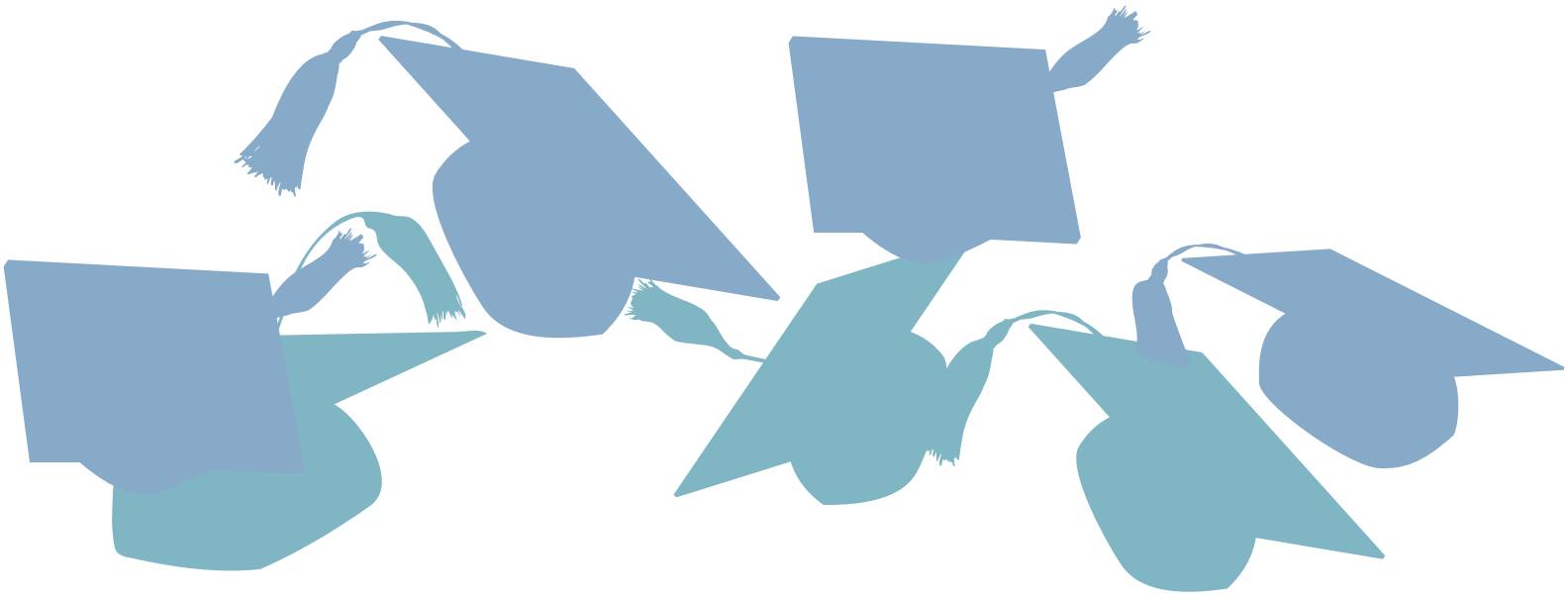
We would like to thank Chris Domaleski, Brian Gong, Karin Hess and Scott Marion, National Center for the Improvement of Educational Assessment, for their leadership in writing the paper. Cory Curl, Senior Fellow, Assessment and Accountability, and Alissa Peltzman, Vice President of State Policy and Implementation Support, Achieve, contributed essential guidance and feedback on the paper.

We would also like to thank the team at KSA-Plus Communications, Inc. for their editorial contributions and Rings Leighton for their design work.

Finally, we would like to express gratitude to the Bill & Melinda Gates Foundation for providing generous funding for this paper.

Michael Cohen
President
Achieve

Brian Gong
Executive Director
National Center for the Improvement
of Educational Assessment



Published in April 2015.

CC-BY 4.0 Achieve 2015. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA.

Suggested citation: "Domaleski, C., Gong, B., Hess, K., Marion, S., Curl, C., Peltzman, A. (2015). Assessment to Support Competency-Based Pathways. Washington, DC: Achieve."

1400 16th Street NW, Suite 510 • Washington, DC 20036
P 202.419.1540 • www.achieve.org

31 Mount Vernon Street, Dover, NH 03820
P 603.516.7900 • www.nciea.org