Invitational Research Symposium on
Through-Course
Summative Assessments

Putting the Pieces Together: Summary Report of the Invitational Research Symposium on Through-Course Summative Assessments

May 2011

Center for K – 12 Assessment
& Performance Management at ETS

Statement of Purpose

This report summarizes expert presentations and discussions that took place during an Invitational Research Symposium on Through-Course Summative Assessments hosted by the Center for K–12 Assessment & Performance Management at ETS on February 10–11, 2011, in Atlanta, Georgia.

A through-course summative assessment is an assessment system component or set of assessment system components that is administered periodically during the academic year. The symposium highlighted many potential benefits of a through-course design. It also raised many challenges the comprehensive assessment consortia -- the Partnership for the Assessment of Readiness for College and Careers and to a lesser degree the SMARTER Balanced Assessment Consortium -- will face in designing, developing, and implementing through-course summative assessments.

This summary report and the research papers, presentation slides and videos of discussions at the symposium are available at www.k12center.org/events.html.

Putting the Pieces Together: Summary Report of the Invitational Research Symposium on Through-Course Summative Assessments

Craig D. Jerald

Nancy A. Doorey and Pascal D. Forgione, Jr., PhD

Center for K-12 Assessment & Performance Management at ETS

Overview

This is an historic moment for education reform. Forty-four states and the District of Columbia have adopted Common Core State Standards in mathematics and English language arts developed by the National Governors Association and the Council of Chief State School Officers (Gewertz, 2011). Initial reviews have found the standards to be not only rigorous, but also clear, focused, coherent, and evidence-based, reflecting the most valuable knowledge and skills students will need to succeed in postsecondary education in the 21st century.

Because these next-generation standards require assessments that can measure complex, 21st century skills, states have begun the difficult but exciting work of designing new testing systems aligned with the Common Core State Standards. Spurred by a desire to encourage efficiency as well as innovation, many states have joined forces to develop common assessment systems, including two consortia that together are receiving a total of $377 million through the federal Race to the Top Assessment Program, the Partnership for Assessment of Readiness for College and Careers (PARCC), and the SMARTER Balanced Assessment Consortium (SBAC).[1]

The application for Race to the Top funds encouraged consortia to go beyond designing assessment systems that merely are common across states and aligned with the new standards. At this level of funding, the two consortia seek to leverage cutting-edge technologies and recent innovations in assessment design to overcome many of the limitations of current state tests.

[1] Two additional state consortia have received federal funding to design alternate assessments for students with significant cognitive disabilities that will align with the comprehensive assessment systems being developed by PARCC and SBAC.

One such innovation is *through-course summative assessment*, a strategy described in, but not required by, the Race to the Top Assessment Program application. A through-course design would enable states to shorten or break up the lengthy summative test typically administered near the end of each year and administer focused *through-course components* periodically throughout the year, combining results at the end of the year to determine a student's final or annual score.

Official Definition

Through-course summative assessment means an assessment system component or set of assessment system components that is administered periodically during the academic year. A student's results from through-course summative assessments must be combined to produce the student's total summative assessment score for that academic year. (U.S. Department of Education,. 2010, p. 18178)

PARCC embraced through-course summative assessment as a major feature of its proposed testing system (see Figure 1). PARCC plans to develop three through-course components that consist of a small number of performance tasks that assess a few priority standards, to be administered after students have completed approximately one-quarter, one-half, and three-quarters of instruction in each subject, respectively. These will be supplemented by an end-of-year component, which will sample the full set of grade level standards and consist of a mix of item types including innovative technology-enhanced questions. According to PARCC, the design will enable summative assessment to take place closer in time to when key topics are taught and provide teachers with more frequent *actionable* results to inform instruction.

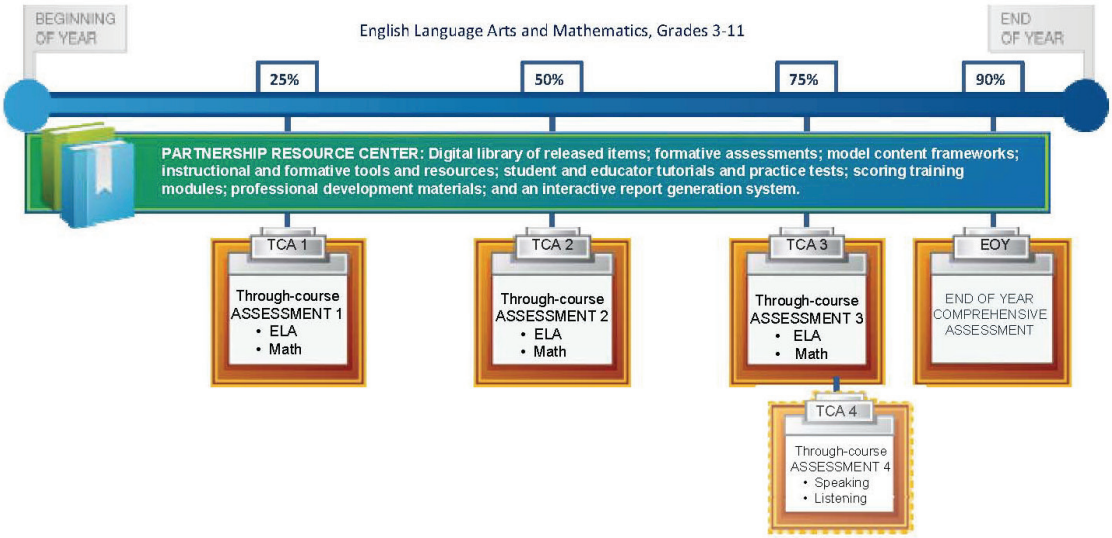

Figure 1. PARCC planned assessment program.

SBAC's planned assessment system does not incorporate a through-course design, relying instead on a set of performance tasks and computer-adaptive assessments administered during a 12-week window near the end of the year (see Figure 2).[2] SBAC also plans to develop optional interim assessments that states may administer periodically throughout the year, but results from those assessments will not count toward summative scores. However, the SBAC will investigate whether to provide member states with the flexibility to administer an alternative summative system that would rely on distributed through-course components. Such components would be computer-adaptive assessments "based on content clusters to allow students to demonstrate mastery of content and skills throughout a course and at the most appropriate time for each student" (SMARTER Balanced Assessment Consortium, 2010, p. 43).

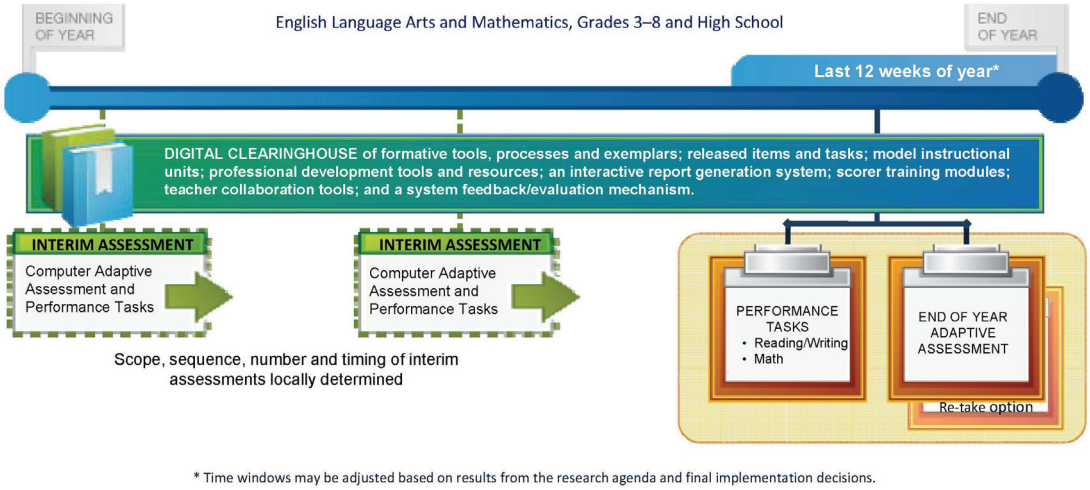

Figure 2. SBAC planned assessment program.

This report summarizes expert presentations and discussions that took place during an Invitational Research Symposium on Through-Course Summative Assessments hosted by the Center for K–12 Assessment & Performance Management at ETS (the K–12 Center) on February 10–11, 2011, in Atlanta, Georgia.

The symposium highlighted many potential benefits of a through-course design. For example, distributing summative assessment components throughout the year could be used to provide timely feedback throughout the school year to allow for midcourse interventions. In

[2] The time windows may be adjusted based on results from the research agenda and final implementation decisions.

addition, as in the PARCC design, this format could make it easier for states to administer extended performance tasks, which might be necessary to measure aspects of the Common Core State Standards critical for college and career readiness. The English language arts standards require that students be able to "conduct short as well as more sustained research projects to answer a question…or solve a problem" (Council of Chief State School Officers & the National Governors Association, p. 46). Traditional end-of-year tests are ill-equipped to assess students' ability to conduct sustained research projects.

Thus, a through-course strategy could help states better assess the full range of knowledge and skills described in the Common Core State Standards, which in turn could help ensure that summative scores provide an accurate measure of students' readiness for college and workforce training. Further, administering such assessments periodically throughout the year would help keep teachers and students focused on high-level skills critical for college and career readiness. Massachusetts Commissioner of Education Mitchell Chester told symposium participants that he believes PARCC's through-course strategy will discourage narrow test preparation by demonstrating that rich and challenging performance tasks truly are valued.

The symposium also raised many challenges the consortia will face in designing, developing, and implementing through-course summative assessments. Section II of this report describes the major measurement challenges raised in the research papers commissioned by the K–12 Center and presented during the symposium, while Section III discusses overarching themes and lessons that emerged from the symposium.

Attendees were urged by Chris Cross of Cross & Joftus and Rick Hess of the American Enterprise Institute to place this work into its larger political context. The debates that will be occurring as part of the reauthorization of the Elementary and Secondary Education Act will raise questions about the common standards, the purposes of the assessments, the role of the federal government in education, and the ultimate uses of the data. In addition, tight state and local budgets will strain implementation efforts. Being mindful of the many competing interests and policy goals will help the assessment community communicate more effectively, influence these policy discussions, and adapt as required to the changing landscape.

With that said, this remains an unprecedented opportunity to make advances in the field of measurement and assessment. "Each consortium will make unique and valuable contributions to our understanding of assessment and will push the frontiers of measurement forward—over the next four years and into the future," commented Pat Forgione of the K–12 Center. "With the active and focused support of the education measurement and assessment community, we can make a difference for the children of this nation."

Meeting Major Measurement Challenges

The K–12 Center commissioned nationally prominent assessment experts to explore key measurement issues raised by the proposals the consortia submitted in June of 2010 for funding under the Race to the Top Assessment Program. The new assessment systems are to be

operational by the 2014-2015 school year. Readers should keep in mind that assessment design and development is a dynamic process, and the consortia's plans are likely to evolve over the next three years.

While each commissioned paper focused on a particular measurement topic, authors at the symposium presented their findings in a logical sequence reflecting how consortia must consider and address various issues, and the summaries below follow the same sequence. consortia must begin by developing a coherent theory of action about how the assessment system will use multiple components to improve student learning; must identify the most essential or *keystone* topics and skills to be assessed in the through-course components; must check the reliability of results produced by each component and the components together; must determine how to best *roll up* results from various components into *annual* scores at the end of the year; must select a model for linking and scaling components to analyze student responses, report results, and monitor progress over time; and, finally, must decide the best method for drawing conclusions about student growth based on results across components and school years.

Validity and Theory of Action

In their paper for the symposium, Randy Bennett, Michael Kane, and Brent Bridgeman of ETS recommended that the consortia employ a novel two-part approach to validation to ensure not only that the interpretations and uses of the test scores are appropriate, but also that the impacts claimed in the consortia's theories of action are occurring.

Part 1. Typically, validation would focus only on whether claims about student achievement based on test scores are, in fact, justified. For example, both consortia plan to generate scores that indicate whether high school students are *college and career ready* as defined by the Common Core State Standards. Colleges could use the information to make placement decisions for incoming freshman, and high school educators could use the scores to make decisions about which students need targeted assistance before they graduate.

However, many factors could undermine that plan. For example, if the assessments fail to cover hard-to-measure skills in the Common Core State Standards that are critical for success in college and the workplace, claims that students are *college and career ready* based on the assessments would not be justified. In the short term, the consortia will need to identify such threats as they begin to develop the assessments. Over the longer term, they will need to collect evidence about how people are interpreting and using various kinds of assessment results to see if such uses are, in fact, appropriate.

Part 2. However, the authors urged the consortia to go beyond that traditional notion of validity to consider whether *the assessment system as a whole* has the multiple impacts that the consortium intends it to have, including both intermediate effects, such as better classroom instruction, and ultimate effects, such as higher rates of college and career readiness.

In his presentation at the symposium, Bridgeman argued that this more comprehensive approach to validation is necessary because the Race to the Top Assessment Program is meant not only to *measure* student performance, but also to *improve* it. "The essential claim for this whole enterprise is that students are going to be better prepared for college and for the workforce," he explained. "That's not just a measurement claim; that's a program impact claim. You could have tests measuring exactly what they're supposed to be measuring and still not have the desired effects on student preparedness."

That two-part validation approach will require each consortium to construct a very clear and explicit *theory of action* describing how its common assessment system will achieve its intended effects. Fortunately, the federal Notice Inviting Applications (NIA) for the Race to the Top Assessment Program required the consortia to include a theory of action section in their applications. However, the NIA did not ask for a sufficiently detailed and explicit theory of action to support the kind of evaluation Bennett, Kane, and Bridgeman envision. Therefore, as a first step, the authors urged the consortia to flesh out their respective theories of action to include all of the following elements:

- The intended impacts of the assessment system, including both intermediate and ultimate effects;
- Features of the system, such as distributed through-course assessments, that will produce the intended impacts;
- Claims that educators, policymakers, or others will make based on assessment results;
- The specific action mechanisms that describe how the various features of the assessment will cause the intended effects; and
- Any potential negative effects (i.e., unintended consequences) and what the consortia will do to mitigate them.

Pursuing that broader validation process also would require a consortium to conduct significantly more research and collect much more data, but the authors argued that the benefits would outweigh the costs. By collecting evidence about both inferential and impact claims, the consortia will be able to analyze whether the entire system is working as intended. If not, consortia can make midcourse corrections to better realize their intermediate and ultimate goals while avoiding unintended consequences.

Bridgeman provided an example: both consortia plan to provide teachers with information they can use to adjust instruction. However, those adjustments will improve learning only if teachers draw appropriate conclusions based on the assessment results they are given, which is far from a simple task. What if a teacher automatically concludes that subskills where students scored lowest deserve extra time and attention even though *other* subskills might be much more critical for career and college readiness? To mitigate that threat, the consortia will need to consider how to provide adequate professional development to teachers

and, over the longer term, collect data on whether teachers are making instructional adjustments based on appropriate inferences.

William Herrera of the Wyoming Department of Education, a governing state in the SBAC, raised another threat to that goal, which he called "drowning teachers with too much information." In the past, some states and districts have worked so hard to provide teachers with a broad range of information from student tests that they have provided too much information for teachers to interpret. "Give them too many kinds of colorful reports splitting up the achievement pie and they end up chucking it," warned Herrera. "They give up because they are overwhelmed."

Only some of the impacts described in the theories of action are intended to result from stakeholders taking action based on scores. Others are intended to result from innovative features of the assessments themselves. For example, the consortia claim that their challenging performance tasks will signal to teachers and students the kinds of high-level knowledge and skills the Common Core State Standards demand and will model the kind of instructional activities that can help students meet those standards, thereby improving teaching and learning *even before* any scores are reported. These claims must be validated by examining changes in instructional practices.

Bennett, Kane, and Bridgeman concluded by urging the consortia to ensure that assessment components cover topics in the Common Core State Standards that are critical for college and career readiness but are not easily measurable using current state tests. And they urged the consortia to seek an appropriate balance between psychometric goals, such as reliability, and other important goals in their theories of action, such as improving instruction.

Finally, because it will be impossible to document the impact of the assessment system on teaching and learning without systematic before-and-after evidence, Bennett, Kane, and Bridgeman counseled the consortia to begin collecting information from key stakeholders *now*, before new assessments are administered. Given the unprecedented size of this federal investment and the nation's urgent need to improve college and career readiness, the consortia simply cannot afford to miss any opportunity to validate that their new assessment systems actually work as intended. And the cost of a thorough research effort will be more than offset by the value of what we learn about how assessments can improve instruction.

Keystone Elements and Learning Progressions

Both consortia plan to develop *learning progressions* that will enable their assessment systems to track progress that students are making toward college and career readiness goals. In its proposal, the SBAC defined learning progressions as "empirically validated descriptions of how learning typically unfolds within a curriculum domain or area of knowledge or skill."[3] To

[3] The SBAC's definition is based on a paper published by the K–12 Center last year. See Darling-Hammond and Pecheone (2010).

develop learning progressions, the consortia will need to identify the *keystone topics*, or *cognitive targets*, that students must master as they develop increasingly sophisticated knowledge and skills on the way toward college and career readiness.

Though the concept is mentioned in documents related to the Common Core State Standards, it is a relatively new one in American education, and little is known about which keystone topics and skills actually underlie the normal progression of student learning in mathematics and, especially, in English language arts. As Nancy Doorey framed the challenge in a recent K–12 Center publication, "What are those skills and concepts, at what level of mastery, and in what sequence, if one such sequence exists?" (Doorey, 2011, p. 15).

The K–12 Center commissioned three university-based experts to explore those questions: Sheila W. Valencia of the University of Washington, Seattle; P. David Pearson of the University of California, Berkeley; and Karen K. Wixson of the University of North Carolina, Greensboro. Valencia, Pearson, and Wixson chose to focus their paper and presentation on reading comprehension, since it is a challenging and relatively unexplored area for identifying learning progressions, and because it is central to learning across the disciplines and the workplace. "Reading is not topical the way other subject areas are," Wixson reminded symposium participants, "so identifying keystone elements is a very different enterprise in English language arts than it is in mathematics or science."

A Multidimensional View of Reading Comprehension

Valencia, Pearson, and Wixson offered the following example to illustrate how reading comprehension depends on the interaction of multiple factors:

…imagine Henry as an eighth grader who scores at the fourth-grade level on a standardized test [and who] might look like a struggling reader in his history class when reading the state adopted textbook. But give him a book on a topic he knows about and has a passion for, basketball for example, and he'll look like a 10th grade reader. Or change the purpose and/or context of the reading from studying for a test to working on a community project that might benefit his neighborhood, and his comprehension of relevant books and articles might rise to his grade level. Or change the setting and the accountability from one in which everyone reads it on his or her own and writes an individual summary to a small group competing with other small groups for the best book-jacket snippet for the novel the class has just read, and his comprehension might look different still. Or frontload the chapter on Egyptian cultural contributions with an engaging movie of the archeological discoveries of ancient Egypt, and his comprehension might improve dramatically. (Valencia, Pearson, & Wixson, 2011, p. 16)

Unfortunately, after reviewing the available evidence, the authors concluded that today, "as a field, we cannot convincingly identify learning progressions that reflect the complex nature of higher levels of comprehension needed for college and career success." Current reading comprehension tests have not been designed around empirically derived models of complex reading comprehension, nor are results reported in ways that represent understanding and learning from text. "There are studies going back to before the sixties showing that skills like main idea, inferencing, drawing conclusions, and so on are not discrete comprehension

skills, yet we still continue to design tests and report scores that way," Valencia said at the symposium.

Therefore, the consortia must begin by redefining the concept of reading comprehension itself as a necessary first step in the development of learning progressions and assessment components. The authors recommended a six-step process based on the approach employed by Irwin Kirsch and others to develop the National Assessment of Adult Literacy (NAAL) and the International Adult Literacy Survey (IALS; Kirsch, 2001). That process begins by defining reading comprehension and continues through organizing the domain, specifying assessment task characteristics, identifying and operationalizing the variables for reporting scores, validating those variables, and building a scheme for interpreting results.

To kick-start the process, Valencia, Pearson, and Wixson analyzed four key sources they believe to be essential for developing a defensible definition of reading comprehension—the report of the RAND Reading Study Group, the National Academy of Science report *How People Learn*, research by Patricia A. Alexander of the University of Maryland, and the Common Core State Standards themselves. Based on those perspectives, the authors concluded that "reading comprehension is an interactive and multidimensional process in which students understand, learn from, and use text to accomplish specific purposes in educational, workplace, and everyday settings." Although this conception of reading comprehension is not new, it has rarely been used to develop reading assessments.

In other words, reading comprehension is a much more sophisticated process than one in which readers simply apply a basic set of decoding and comprehension skills to a text of a given difficulty level. Instead, reading competence varies in response to the interaction of a wide variety of factors related not only to the reader and the text but also to the nature of the activity and the setting (or sociocultural context) in which the activity occurs. Readers themselves do not just bring different kinds of cognitive skills to bear, but also different levels of motivation and various types of knowledge that can vary greatly from one activity to the next within the same hour or day. (See Figure 1.) And the relative importance of different kinds of knowledge, motivation, and skill changes as a reader develops from a novice into a competent or expert reader over time. It is the ability to adapt to different reading situations, texts, and task demands that is the hallmark of reading competence.

"Reading comprehension assessments must reflect the complex, dynamic nature of comprehension," contended Valencia, Pearson, and Wixson. "Otherwise assessments cannot serve as models for good instruction or as valid predictors of college and career success." They went on to propose a unique approach to developing such assessments, "Text-Task Scenarios (TTS)," which would assess readers' abilities in relation to various combinations of purposes for reading, range of texts, and types of comprehension tasks, as well as their interests, prior knowledge, and reading strategies .

The authors concluded by outlining an ambitious, two-stage research agenda to "get it right." In the short term, the consortia could identify existing assessments that might be revised

to better match the complex definition of reading assessment; analyze how well results from current assessments predict college and career readiness while simultaneously developing better criteria for success (beyond broad measures like dropout rates or college GPA); and identify high-quality texts already "permissioned" for use in assessments as a basis for developing and piloting TTSs. During the second stage, the consortia would begin more systematic development of TTSs to pilot in classrooms where teachers already engage in high-quality instruction, as well as conduct a number of studies to answer a range of psychometric questions related to scaling, consequential validity, generalizability, and other issues.

Experts at the symposium agreed with their conclusions and recommendations. David Coleman of Student Achievement Partners commended the authors for "a thoughtful discussion about not reporting reading subscores in terms of the classic skills" because they can be so misleading to teachers:

It is not that these skills are meaningless, but rather that they correlate so tightly together that it is a fantasy or falsehood of data-based instruction that your students are weak at "main idea" or "inference." We've almost misled teachers by giving them a false precision that implied that if today they addressed "main idea" and tomorrow they did "point of view," they were teaching well, based on the data.

At the same time, Coleman cautioned consortia and their partners to maintain a healthy skepticism moving forward "rather than allowing a new dogma to enter, because that false precision could mislead us once again."

Reliability

Michael Kolen of the University of Iowa cautioned the consortia that estimating reliability of scores will be particularly challenging, given certain features of their planned assessment systems, including the following:

- The assessments will generate many different types of information, and the consortia will need to estimate reliability for each score reported, as well as for different student populations.
- Some of the planned assessment components will include only a handful of longer performance tasks, which will make it very difficult to estimate reliability of individual component scores, due to the relatively small number of items administered.
- The consortia will need to estimate the overall reliability of annual composite scores created by combining results from different kinds of assessments given at different times of the year.
- The consortia plan to rely on automated scoring of performance tasks using advances in artificial intelligence, even though methods for estimating reliability of such automated scores have not yet been fully developed.

Given these challenges, the consortia will need to conduct a wide variety of pilot studies to ensure that scores are sufficiently reliable for their intended purposes by the time the new assessments are formally administered in 2014-2015. Kolen also cautioned the consortia not to assume that a student's proficiency is constant over various types of assessment tasks and items *or* over different points in the school year; instead, the consortia should estimate reliability of each through-course component individually and use psychometric procedures that are designed to assess reliability for composite scores.

Moreover, beyond simply estimating reliability of results, the consortia will need to make decisions about how to *maximize* the reliability of many kinds of scores while dealing with the same set of challenges. For example, certain components will have lower reliability than others because they include far fewer items. PARCC's first three through-course components and SBAC's performance-task component will each be composed of one to three tasks, according to their proposals, compared with 40 to 65 items per subject that PARCC plans to include on its end-of-year component and SBAC plans for its computer-adaptive component.

From a purely psychometric standpoint, that would argue for giving greater weight to components that rely less heavily on small numbers of performance tasks (i.e., give more weight to PARCC's end-of-year component and SBAC's computer-adaptive component) in order to boost the reliability of combined annual scores. However, such a weighting scheme might compromise other important policy goals, such as improving instruction. If results of performance tasks contribute less toward students' combined scores, will teachers and students take those tasks less seriously? If that happens, will the performance tasks lose their beneficial signaling and modeling impact on teaching and learning?

In his presentation, Kolen suggested that one solution for mitigating that tradeoff might be to develop performance tasks that can be broken up into multiple, separately scored subtasks, thereby increasing the reliability of component scores. However, that solution might itself entail tradeoffs. In responding to Kolen's presentation, David Coleman of Student Achievement Partners questioned whether breaking up the extended performance tasks into smaller pieces would undermine the authentic nature of such "sustained research projects" as envisioned in the Common Core State Standards.

Thus, this leaves us with a set of important policy and research questions. Should psychometric goals such as reliability always trump instructional goals and other goals for assessment? How can the consortia strike the right balance among the many policy goals set out by the Race to the Top Assessment Program and described by the consortia in their plans submitted last summer, from ensuring psychometric quality to improving instruction to supporting accountability?

Combining Through-Course Results

One of the most difficult and important decisions a consortium faces in using through-course assessments has to do with deciding how to combine results from components

administered at different times in the year to calculate *annual* measures of proficiency and growth.

Should all through-course components be weighted equally? If so, that might be unfair to students who begin with lower scores but make a lot of progress from the beginning to the end of the year. In that case, should the end-of-year component receive greater weight? If so, that would diminish the importance of performance tasks administered earlier in the year and, consequently, reduce incentives for teachers and students to focus attention on the critical college and career readiness skills measured by those tasks.

In his presentation at the symposium, Lauress Wise, principal scientist at the Human Resources Research Organization (HumRRO), argued that a consortium's choice about how to aggregate through-course results should be guided by *how and when students actually learn the material covered by the assessments*. Wise described four possible scenarios, or "learning models," in his paper:

- *One-Time Learning* assumes that students learn little or nothing about a topic until it is taught, and that their mastery of the topic neither improves nor erodes over the rest of the year.
- *One-Time Learning With Forgetting* assumes that students learn little or nothing about a topic until it is taught, and that their mastery of the topic then erodes over the rest of the year.
- *One-Time Learning With Reinforcement* assumes that students learn little or nothing about a topic until it is taught, and that their mastery then improves over the rest of the year as the topic or skill is reinforced by subsequent instruction.
- *Continuous Learning* assumes that students progressively master a topic or skill at a relatively even pace throughout the school year.

Rather than pursuing a "one learning model fits all" approach, the consortia will need to recognize that different models might fit best, depending on the content areas and types of knowledge or skill being assessed. For example, skills like reading comprehension are expected to improve continuously over time, while other material is intended to be taught and learned only at a particular point in time. "You don't know very much about exponents, then you get to the unit that teaches you about exponents, and you learn what you need to learn about exponents," Wise offered as an example of the latter. After the unit, students' knowledge of exponents might erode, might be retained, or might even be strengthened as new math topics reinforce that initial learning.

Wise used simulated results for 400,000 students to investigate how well several alternative methods for aggregating scores performed under each of the four different learning scenarios above; he found that the choice can have serious consequences. "Midyear results can significantly underestimate or, in some cases, overestimate end-of-year status and growth if

the method for aggregation is not consistent with how students actually learn," he concluded. "The biggest finding of this study is that the method you choose makes a huge difference."

For example, giving students multiple opportunities to take the same test and picking the highest score for summative purposes tends to seriously overestimate end-of-year achievement and growth. On the other hand, simply averaging scores across through-course components without regard to when material is taught would seriously *underestimate* annual growth, especially if students continue to learn material after it is assessed.

Fortunately, Wise found that some methods can work well if they are matched to an appropriate model of how students learn. Specifically:

- For knowledge and skills that are taught and learned at a given point in the school year, consortia can use a method resembling "end of unit testing" that simply adds up scores on different through-course components administered over the year.
- For knowledge or skills that improve continuously throughout the year, consortia can consider a method that employs regression-weighted projections to predict end-of-year proficiency or annual growth.

However, Wise emphasized that much more research will be necessary before the consortia can make decisions about how to aggregate results, and those decisions must be validated by examining data from "operational" results after the assessments are finally administered. To begin with, the consortia will need to conduct research on learning progressions so that the selection of aggregation model can be based on proven models of how students learn. They also will need to analyze the Common Core State Standards and other materials to "distinguish between topics or skills that are taught at particular points in the curriculum and topics or skills that are learned and practiced more or less continuously throughout the year."

After the consortia have developed assessments, they can collect *real* data on how and when students actually learn material in order to fine-tune the aggregation methods Wise explored. For topics or skills taught at a particular time, the consortia can explore how much additional learning or forgetting occurs for students between the time the topic is taught and the end of the year. For topics or skills taught throughout the year, the consortia can explore how well performance at each point in time predicts end-of-year performance for groups of students.

Scaling, Linking, and Reporting

In their paper for the symposium, Rebecca Zwick and Robert Mislevy of ETS emphasized "the multitude of inferential demands" through-course assessments must accommodate and the corresponding "host of psychometric challenges" they present. They stressed the importance of recognizing the tradeoffs between the number of purposes to be served by the assessment system and the resulting complexity of the analysis model. Importantly, they also

demonstrated that a sophisticated enough model can reduce constraints on local curriculum sequencing that constitute one of the most significant practical challenges of a summative through-course design.

Zwick began her presentation of the paper by describing the many purposes the common assessment systems will have to serve. The assessments must: (a) estimate individual student proficiencies and proficiency distributions, as well as means for student groups; (b) base their estimates on a wide variety of item types, including complex performance tasks; (c) provide timely *actionable* data for teachers after each through-course component; (d) summarize through-course results across a year, accommodating students who receive different patterns of instruction; (e) measure growth across an academic year; and (f) provide results that are comparable across classrooms, schools, districts, and states.

A through-course approach that distributes summative assessment components throughout the year will impose especially heavy psychometric demands, according to Zwick. While states now must create different test forms each year (for example, this year's fourth grade test must be based on a different set of test questions than last year's fourth grade test), through-course components will require multiple forms *within* each year, as students in different states may be administered these highly memorable tasks on different dates. And because through-course components are intended to provide actionable results, the items they administer must be sensitive to the content and timing of instruction that students receive.

As a result, any simple approach to analyzing student responses in order to generate scores and other information simply will not suffice. Instead, the consortia will need to recognize a fundamental tradeoff. "The more demands that are made of the scaling and reporting model," contended Zwick and Mislevy, "the more complex the model needs to be" and the more difficult it becomes to explain to nontechnical audiences.

However, Zwick and Mislevy also demonstrated that a sophisticated model also can have great advantages. For the consortia's consideration, they proposed a specific model capable of supporting the extensive set of purposes stated in the consortia's proposals, while giving individual states or local education agencies much more flexibility to decide when and in what order to administer through-course components. The model would employ a "multidimensional item response theory (MIRT)" approach that builds on lessons from the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS).

Their proposed model would have the advantage of relying on an existing base of research conducted to support those large-scale assessments. It also would allow the consortia to carry out pilot studies and simulations to determine what kind of testing designs would be required to support the many purposes of the common assessment systems, as well as to explore when simpler procedures might suffice.

Garron Gianopulos, a psychometrician for North Carolina Public Schools, said the model might help address some of the concerns his colleagues have raised about through-course

summative assessments when the SBAC was considering them last summer. For example, "many people objected to the idea of having a single pacing guide," said Gianopulos. "At the time, that was the assumption: that we'd need to have classrooms and teachers moving in lock step through the curriculum in order to allow a summative through-course system to function properly."

However, said Gianopulos, Zwick's and Mislevy's proposed model might offer:

> …a way that we can allow greater flexibility so that, at least to some extent, we can allow teachers to spend as much time as they need to within reason, and allow even possibly different sequences of instruction and curriculum.

He concluded that, "I feel like one of the objections I've heard expressed about through-course assessments has been overcome, potentially at least, through this model."

While she praised the functionality of the model, Kit Viator of the Bill & Melinda Gates Foundation worried that the model's complexity could have serious practical drawbacks. "Whatever models are adopted, it's going to be very important to try to break down 'what is under the hood' in terms of analysis and equating," said Viator. She argued that a lack of transparency could hinder state efforts to build public confidence and investment in the new common assessment systems.

In response, Zwick suggested it might be possible to simplify the model to some extent and offered several ideas for doing so that the consortia could investigate during piloting and field testing. For example, the scaling, linking, and reporting process could be streamlined greatly by adopting a dual strategy that relies on different test forms for different purposes. A random sample of students would be given more conventional test forms, including only machine-scorable items, the results of which would be used for accountability and school, district, and state reporting purposes. Schools would then administer relatively unconstrained test forms, including complex performance tasks, for the purpose of informing instruction.

Fortunately, a complex model for *analyzing* student performance need not entail a complex model for *reporting* it. To that end, Zwick and Mislevy recommended that the consortia consider a "market basket" approach to reporting results, so called because it reports student performance in terms of a set of specific items and tasks that can be released publicly to provide tangible examples of what students know and can do. "The 'behind-the-scenes' machinery is complex," Zwick explained, "but the resulting scores are simple and look like ordinary test scores."

Analyzing Student Growth

In his presentation at the symposium, Andrew Ho of the Harvard Graduate School of Education contrasted several kinds of student growth models that states currently are

implementing under the federal Growth Model Pilot Program (GMPP) announced in 2005. The GMPP provides states with flexibility to consider student growth measures in their policies for determining whether schools are making Adequate Yearly Progress (AYP) under the Elementary and Secondary Education Act (ESEA). These models serve as prototypes that the consortia may incorporate into the through-course assessment context.

Ho focused on the two main types of growth models approved for use under the GMPP, trajectory models and projection models. Both are examples of so-called *growth-to-proficiency* or *growth-to-standard* models because they make predictions about whether students are likely to reach a given proficiency cutoff on state assessments within a certain period of time. Although the Race to the Top Assessment Program does not require PARCC and SBAC to use growth-to-standard models, Ho believes they are useful exemplars, given both consortia's clear focus on ensuring that students reach career and college readiness.

For example, the consortia could use growth-to-standard models to generate predictions of whether students are "on track" to meet career and college readiness by the end of high school or to meet aligned proficiency expectations in earlier grades.[4] consortia or individual states could report such predictions *during* the school year based on the results of through-course components, at the *end* of each year based on combined annual scores, or both.

In their simplest form, *trajectory models* draw a line through a student's past test scores and extend that line into the future to anticipate whether the student will meet a proficiency cutoff within a certain timeframe *assuming* he or she remains on his or her current course or, to phrase it slightly differently, maintains his or her current *momentum*.

In contrast, *projection models* use regression techniques to analyze what happened for a *previous* cohort of students in order to predict success for students in a current cohort. In essence, such models answer the question: Based on what happened to previous students who had a test score history similar to this student's (and, depending on the specific model, who may have shared other characteristics in common with the student), is this student likely to reach proficiency within the given time frame?

Ho's analysis revealed stark contrasts and tradeoffs between the two models. For example, trajectory models make more intuitive sense to people as a measure of *growth* than do projection models. In fact, projection models do not consider the order of a student's prior test scores, or even whether scores are trending downward or upward.

---

[4] Confusingly, the federal Race to the Top program defined the term "on track" quite differently from how the term "on track" is used in the federal GMPP pilot. The Race to the Top application defined "on track to college and career readiness" as a standardized proficiency cutoff to be used in earlier grades as a replacement for "proficient" under the Elementary and Secondary Education Act. Therefore, Consortia or states that decide to report growth-to-standard measures would need to use a different term than "on track" to describe such measures. Otherwise, a student might be considered "on track" according to a predictive growth measure even if he or she fails to meet the "on track" benchmark for his or her grade level.

Another reason trajectory models make more intuitive sense as measuring *growth* is that they require assessments to incorporate a vertical scale. Although vertical scales are costly to develop, they are useful for making inferences about student growth because they place the scores from all grade level assessments onto a single continuous scale. Scores based on the scale can then be used to monitor growth in achievement, much as one uses a yardstick to measure the physical growth of a child over time.

Nevertheless, studies have shown that projection models are consistently more accurate than trajectory models when it comes to actually predicting students' future achievement, which might be an important consideration for accountability purposes.[5]

The models also present tradeoffs in terms of potential negative effects and unintended consequences through the incentives they create. On the one hand, trajectory models could encourage teachers and students to try to "game" the measure by deflating early test scores. This strategy would increase the slope of the trajectory line and thereby inflate predictions of future scores. (See Figure 3.)

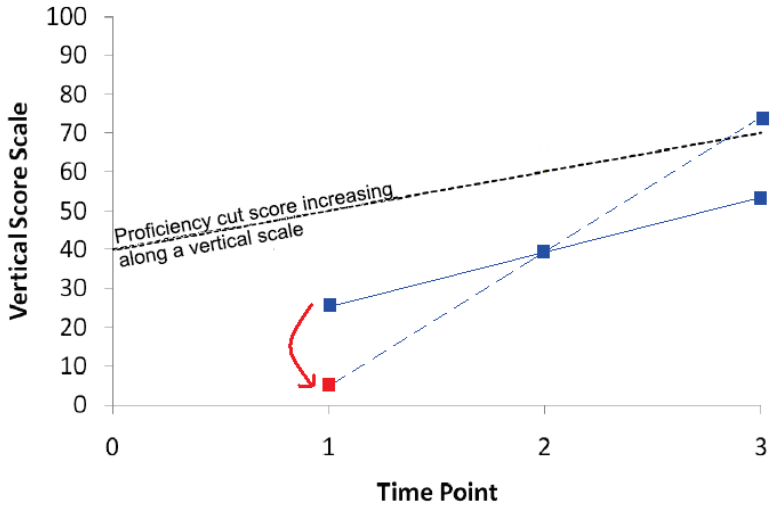

Figure 3. Gaming trajectory models.

[5] See for example Hoffer, et al. (2011). Predictive accuracy also would be a highly important consideration for states planning to use such growth measures as part of an "early warning system" to target intensive interventions to students most at risk of not graduating career and college ready, especially given the high economic and opportunity costs associated with false positives and false negatives in such systems. See for example Jerald (2006). In fact, considering both a trajectory measure and a projection measure at the end of each year might help prioritize intensive interventions for individual students, thereby maximizing their effectiveness and cost-effectiveness.

On the other hand, projection models suffer from what Ho calls "inertial effects." Low scores tend to predict low scores and high scores tend to predict high scores—no matter what a student's longitudinal trajectory may be. Students with a history of very high scores could actually decline substantially on one or more subsequent test administrations and still be considered "on track" according to the model. And students with a history of low scores would have a very difficult time earning an "on track" prediction even if they are making significant improvements in achievement, which could be very discouraging both for low-performing students and their teachers.

Garron Gianopulos of the North Carolina Department of Education observed that the "inertial effect" Ho described is due to the very problem that teachers would presumably be working very hard to fix—the fact that in our current education system, low achievement in the past predicts low achievement in the future for too many students. "Can we change the educational system so that the achievement data from the past is no longer applicable for predicting the future?" Gianopulos wondered. "And if we're making those educational changes, couldn't we update our projection models to reflect the new reality so that eventually one day low scores will not predict more low scores?"

Ho responded that while it is possible to conceive of that happening over time, it is important to keep in mind the incentive structure at any given point in time. Because a regression-based model makes predictions based on the reality for a previous cohort of students, the students who are benefitting from better instruction and support—and the teachers providing it—would not see those improvements reflected in growth model predictions. "To a certain extent, if you use a projection model, it fixes that incentive structure for students who might be benefitting from such changes," Ho observed.

Therefore, based on what is currently possible, Ho recommended that the consortia honestly recognize the tradeoffs among alternative growth models, not only in terms of the inferences they support concerning growth, but also in terms of the incentives they create. Furthermore, if a state or consortium opts to use a projection model for accountability purposes (perhaps due to its greater predictive accuracy), Ho recommended not calling it a "growth measure" and not reporting individual student scores to teachers, since they do not offer pedagogically useful information about the progress of individual students.

Finally, Ho believes that a *hybrid* model might prove best of all by mitigating the worst features of both models. "Maximizing accurate prediction of future status results in inertial tendencies that decrease the incentives to teach initially low- or high-scorers, but embracing intuitive trajectory models leads to incentives to decrease early scores," he told participants in a follow-up webinar.[6] "Therefore, I would argue for a transparent hybrid with positive but

[6] The webinar may be viewed at http://media.all4ed.org/webinar-mar-10.

lower earlier weights that would sacrifice some accuracy of prediction and true growth-over-time interpretations but that may, in fact, be the optimal design."

Overarching Themes and Lessons

Although the symposium filled in some critical gaps in our knowledge about through-course summative assessments, its greater value was in clarifying the questions, challenges, and trade-offs that lie ahead of these consortia. Answering those questions will require both consortia to complete a truly *ambitious research and development agenda*, much of which will need to be accomplished during a very short timeframe, since the assessments must be ready for wide-scale administration by the 2014-201 5 school year. Other critical questions, such as those having to do with the impact of these new systems on the quality of instruction and readiness of students for college and careers, can only be answered after the assessments are administered in 2014-2015.

Therefore, the consortia will need to flesh out detailed research and development plans as soon as possible, including, but not limited to, specific studies suggested by presenters at the symposium. While working collaboratively as consortia will enable states to accomplish much more than they could tackle as individual states, the amount of remaining research is so significant and the timeframe so short that the consortia will need to *reach out to the broader community of assessment experts*—university researchers, national research associations, and federal agencies—to enlist their support. John Easton, director of the federal Institute of Educational Sciences, punctuated the need for a comprehensive, robust, and integrated research and evaluation agenda in order to learn as much as possible about system and school improvement from this "major education intervention" of new assessment systems and related tools and resources.

Some participants at the symposium argued that outside experts, particularly those in the psychometric community, will need to "*stretch beyond our current comfort zone*" to be of maximum assistance over the next three years. Historically, psychometricians have tended to prioritize measurement goals such as reliability over the kinds of practical utility that policymakers and teachers increasingly demand from tests. Some participants argued that policymakers need to scale back demands on the new assessment systems. However, others said that is very unlikely and instead urged measurement experts to help the consortia find ways to deal with such demands proactively, pragmatically, and creatively.

"In the measurement community, we have to stop saying only 'No, you can't do that'—because it's going be done anyway—and instead figure out new methods to get beyond those obstacles," said Joan Herman, director of the National Center for Research on Evaluation, Standards, and Student Testing at UCLA. "We need to explore new approaches to establishing the accuracy and reliability of our measures."

That kind of assistance will be especially important as the consortia *navigate many difficult tradeoffs* among competing accountability goals, instructional goals, and measurement

goals over the next three years. Not only will the consortia often face tensions between different types of goals (for example, between greater reliability of scores and greater utility of tests for improving classroom instruction), they sometimes will face difficult tradeoffs between two goals of the same type.

For example, Michael Cohen, president of Achieve, the project management partner for PARCC, described how PARCC has been struggling to manage tension between two important instructional goals. Through-course performance tasks that require students to "read a text of an appropriate level of complexity, whether literary or informational, and write in response to a prompt that requires finding evidence in the text and make a logical, coherent argument about it in writing," Cohen said, will encourage teachers to incorporate such projects into their ongoing instruction, especially if those tasks count for summative scores and, thus, accountability purposes.

However, "that may not be the form of assessment that will provide the most valuable diagnostic information," Cohen explained, "that will, on its own, allow teachers to know exactly what to do when they get the results back in terms of adjusting instruction." One solution might be to add additional kinds of items that can measure discrete subskills (which also would address concerns about reliability raised by Kolen). But that solution carries a risk too. Through-course components might begin to look so much like current end-of-year tests that teachers reject them as "just more testing to deal with."

Many experts at the symposium stressed that the consortia will be better equipped to recognize and navigate such difficult design tradeoffs if they each develop a *clear and specific theory of action*. As explained in the Meeting Major Measurement Challenges section, a theory of action begins by describing the explicit goals the assessment system is intended to achieve and then works backward to construct a detailed, logical explanation of how each feature of the assessment system will help achieve those goals.

Developing a clear theory of action also could help the consortia communicate with stakeholders about the many advantages and benefits of the new assessment systems. "I think we have to be very explicit in our theory of action to be clear about how this is going to benefit kids," said Massachusetts Commissioner Mitchell Chester. "If you can't bring it back to how kids are going to benefit, and you can't put kids squarely in the middle of these conversations, then you will have a hard time making the case."

At the same time, the consortia should be as *transparent and honest as possible* in communicating with stakeholders about what the new assessment systems are intended to accomplish and what they cannot accomplish, including any necessary tradeoffs among competing goals.

For example, parents should be told exactly what kinds of questions the new assessment systems can help answer and to what extent (for example, "Is my child progressing adequately toward college and career readiness?"; "Is my child's teacher effective?"; "Is this school performing adequately?"). However, participants also urged the consortia to proactively

discourage stakeholders from making *inappropriate* interferences about students, teachers, or schools based on assessment results—rather than simply bemoaning such unintended consequences after the fact. "Amongst ourselves, we can comfortably talk about things like reliability and validity and what that really means for a test," observed Gilbert Andrada of the Connecticut Department of Education. "But we sometimes lose control of how those numbers get used when they pop out of the system, and we'll read in newspapers things about our data that we never intended."

To that end, the consortia should consider taking Bridgeman's recommendation to examine potential unintended consequences as an additional step in fleshing out their respective theories of action. Doing so will help identify and clarify significant risks as early as possible before the assessments are formally administered in 2014-201 5. This would provide the consortia with plenty of time to work with assessment experts and policymakers to develop creative solutions for avoiding or mitigating those risks.

Several policy-related discussions at the symposium made clear that the consortia also will need to find ways to deal with difficult *political challenges and practical tradeoffs* over the next three years.

For example, Stan Heffner of the Ohio Department of Education warned that the lack of readily available and high-quality model curricula and professional development resources might undermine the consortia's efforts to improve teaching and learning, since new standards and assessments alone will not help teachers dramatically change their instruction to give students sufficient opportunities to practice the kinds of performance tasks the assessments will incorporate. However, Chris Cross, chairman of the consulting firm Cross & Joftus, pointed out that it may not be appropriate for the federal government to fund the kind of voluntary curriculum tools and materials the consortia plan to create using supplemental Race to the Top grants awarded last fall.[7] Speaking as a state official himself, Heffner countered that many states and districts need and welcome federal support to develop such resources, as long as they remain voluntary.

Another thorny curriculum-related issue has to do with potential constraints on local curricula that might result from a through-course assessment design. Rick Hess of the American Enterprise Institute suggested that some constituencies—particularly charter school managers and advocates—might strongly oppose through-course summative assessments if they diminish curriculum flexibility to any significant extent. However, as described above, some participants thought that Zwick's and Mislevy's model for analyzing student performance could greatly abate such worries by allowing local education agencies, including charter schools, to administer through-course components whenever and in whatever order they choose. (Of course, the consortia will need to consider additional policy and security implications before

[7] See http://blogs.edweek.org/edweek/curriculum/2011/02/can_the_federal_government_fun.html.

making decisions regarding local flexibility in timing and sequencing the administration of components.)

That conversation prompted some participants to argue that the consortia will need to *reach out to many different constituencies*—beyond the expert psychometric community—early and often, rather than simply unveiling the new assessment systems as a *fait accompli* three years from now. "How do you reach out and make sure you are talking to constituencies who have legitimate concerns or real questions—whether practical, logistical, or normative—about how these assessment designs are taking shape?" asked Hess. "How do you make sure that a 'communications strategy' is a two-way conversation rather than a one-sided lecture?"

Officials working with the state consortia responded positively to such suggestions. "I agree with you that communications is two-way," said Cohen. "This is not just about telling; this is about listening, and the listening has got to be happening right now." He suggested that just as the consortia can reach out to the broader expert assessment community for help developing assessments, they also can begin reaching out to many other national and state organizations than can help communicate about the assessments and establish forums where educators and community members can ask questions, raise concerns, and make suggestions.

Conclusion

Clearly, the symposium highlighted a number of practical steps the consortia can take, with support from the larger measurement community, to proactively address several broad challenges, including the following:

- Review and revise the theory of action to make it more clear and explicit in order to proactively identify risks and challenges, navigate difficult design tradeoffs, and communicate more effectively about how the common assessment systems will benefit students.
- Review preliminary research plans to flesh out a detailed research and development agenda in order to identify potential allies in the broader assessment community and enlist their assistance as soon as possible.
- Review and revise preliminary communications plans to ensure they represent a strategy to engage key groups and constituencies early and often in an ongoing two-way dialogue about key issues.

Even so, accomplishing everything necessary to design, develop, and pilot sophisticated new common assessments systems by 2014-2015 will be far from easy. One key to their success may be the size of the consortia. As Kris Ellington of the Florida Department of Education put it, "There may be more alligators [within these next generation assessment designs], but there are also more opportunities" that come from the ability to draw upon "the collective wisdom" among states and among the leading measurement experts in the country.

If one consistent theme emerged from the symposium, it was this: *failure simply is not an option*. Current state testing systems are far from adequate for assessing the kinds of college

and career ready skills embodied by the Common Core State Standards, which in turn are vital to securing America's ability to remain competitive in the 21st century. Moreover, if we are to *accelerate the learning of American students*, new systems are needed that yield more timely and useful information to students, educators, parents, and policymakers.

The ambitious plans laid out by the consortia *will undoubtedly drive innovation and improvements in the field of assessment*. But it will take a rare combination of pragmatism, humility, optimism, resourcefulness, and persistence—as well as broad cooperation from a wide variety of experts and stakeholders—to fully achieve their visions. The K–12 Center will continue to identify measurement challenges to be overcome, organize thoughtful exploration of options, and share the best thinking broadly in order to assist in this timely and critical national endeavor.

References

Council of Chief State School Officers, & the National Governors Association. (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, & technical subjects.* Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Darling-Hammond, L., & Pecheone, R. (2010, March). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Princeton, NJ: Center for K–12 Assessment & Performance Management at ETS.

Doorey, N. (2011, February). Finding solutions, moving forward. In *Coming together to raise achievement: New assessments for the Common Core State Standards* (pp. 15–17). Princeton, NJ: Center for K–12 Assessment & Performance Management at ETS.

Gewertz, C. (2011, April 4). Common-Standards watch: Maine makes 45. *Education Week Online*. Retrieved from http://blogs.edweek.org/edweek/curriculum/2011/04/common-standards_watch_maine_m_1.html

Hoffer, T. B., Hedberg, E. C., Brown, K. L., Halverson, M. L., Reid-Brossard, P., Ho, A. D., & Furgol, K. (2011). *Final report on the evaluation of the Growth Model Pilot Project*. Washington, DC: U.S. Department of Education.

Jerald, C. (2006). *Identifying potential dropouts: Key lessons for building an early warning data system.* Washington, DC: Achieve, Inc.

Kirsch, I. S. (2001). The International Adult Literacy Survey (IALS): Understanding what was measured (No. RR-01-25). Princeton, NJ: ETS.

SMARTER Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program application for new grants: Comprehensive assessment systems CFDA Number: 84.395B*. Retrieved from http://www.k12.wa.us/SMARTER/RTTTApplication.aspx

Valencia, S., Pearson, P. D., & Wixson, K. (2011) *Tracking progress in reading: The search for keystone elements in predicting college and career readiness.* Retrieved from http://k-12center.net/rsc/pdf/TCSA_Symposium_Final_Paper_Valencia_Pearson_Wixson.pdf

U.S. Department of Education. (2010, April 9). Race to the Top Fund Assessment Program notice inviting applications for new awards for fiscal year (FY) 2010. *Federal Register, 75*(68), p. 18178.

Appendix

Research and Resources on Through-Course Summative Assessments at www.k12center.org

Research papers and presentation slides from the K–12 Center's Invitational Research Symposium on Through-Course Summative Assessments (February 2011):

Bennett, R. E., Kane, M. T., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment.*

Heffner, S. (2011). New common assessments: Practical applications of new opportunities. (Commissioned by CCSSO)

Ho, A. D. (2011). *Supporting growth interpretations using through-course assessments*.

Kolen, M. J. (2011). *Generalizability and reliability: Approaches for through-course assessments.*

Sabatini, J. (2011). *Lessons learned from the CBAL's four year experience in developing through-course formative and summative assessments*.

Valencia, S., Pearson, P. D., & Wixson, K. (2011) *Tracking progress in reading: The search for keystone elements in predicting college and career readiness.*

Wise, L. L. (2011). *Picking up the pieces: Aggregating results from through-course assessments.*

Zwick, R., & Mislevy, R. J. (2011). *Scaling and linking through-course summative assessments.*

Videotaped Panel Discussions on the Implications, Benefits, and Challenges of Through-Course Summative Assessments (Recorded February 2011. Released online March 2011)

State Assessment Directors and Local Policy Leaders Conversation

- Rick Hess, Director of Education Policy Studies, American Enterprise Institute (moderator)
- Kris Ellington, Deputy Commissioner, Florida Department of Education
- Shelley Loving-Ryder, Assistant Superintendent, Virginia Department of Education
- William Herrera, Director of Test Development and Research, Wyoming Department of Education

Consortium and National Policy Leaders Conversation

- Rick Hess, Director of Education Policy Studies, American Enterprise Institute (moderator)
- Mitchell D. Chester, Commissioner, Massachusetts Department of Elementary and Secondary Education
- Christopher T. Cross, Chairman, Cross and Joftus, LLC
- John Easton, Director, Institute of Educational Sciences
- Angela Hinson Quick, Assistant State Superintendent, North Carolina Department of Public Instruction (SBAC)

Videotape of March 10, 2011, webinar discussion featured paper authors Lauress Wise, Rebecca Zwick, and Andrew Ho and moderated by Pascal (Pat) Forgione of the K-12 Center and Governor Bob Wise of the Alliance for Excellent Education.