Invitational Research Symposium on
Through-Course
Summative Assessments

THEORY OF ACTION AND VALIDITY ARGUMENT IN THE CONTEXT OF THROUGH-COURSE SUMMATIVE ASSESSMENT

Randy Elliot Bennett, Michael Kane, and Brent Bridgeman

Educational Testing Service

March 2011

Center for K – 12 Assessment
& Performance Management at ETS

Theory of Action and Validity Argument in the Context of Through-Course Summative Assessment

Randy Elliot Bennett, Michael Kane, and Brent Bridgeman
Educational Testing Service

Executive Summary

To validate an assessment program, it is necessary to be clear about what is being claimed by the interpretations and uses of the test scores, and in cases where the assessment is expected to produce certain outcomes, what is being claimed by the theory of action. To the extent that these claims are justified, the program can be considered valid.

Through-course assessments may have several potential benefits. First, they should make it possible to assess the Common Core State Standards more thoroughly and more representatively. Second, the assessments should be able to inform students and teachers about the content of the Common Core State Standards, demonstrate the characteristics of proficient performance, and model good teaching and learning practice. Third, by including a larger number of performances of different kinds, through-course assessments should help to control random errors of measurement and some kinds of systematic sampling errors and therefore should enhance the generalizability and validity of the proposed interpretation.

In conceptualizing the validation of a through-course assessment system, we suggest a two-part interpretive argument consisting of measurement-argument and theory-of-action components. Assuming the through-course assessment is representative of the Common Core State Standards and provides unbiased and adequately precise measure of overall performance in it, the *measurement argument* claims that it would be reasonable to interpret and use scores on the through-course assessment as an indicator of overall standing on the Common Core State Standards. This part of the interpretive argument takes us from observations of

performance on the through-course assessment to claims about expected performance on the Common Core State Standards. It is similar to the interpretive arguments for other achievement tests, with two complications: (a) the through-course assessment is not administered as a single assessment at the end of the course but, rather, is administered at various points during the course, and (b) the through-course assessment will include substantially different kinds of tasks. Both of these complications can be managed if the assessment program is designed carefully.

The *theory-of-action* part of the interpretive argument focuses on the use of the assessments to enhance individual (e.g., students, teachers, administrators) or institutional (e.g., schools, districts) performance. Some of the intended effects of through-course assessments will follow the traditional measurement-based model and depend on how the scores are used (e.g., for accountability, feedback, placement of students and teachers). In addition, both the Partnership for Assessment of Readiness for College and Careers and the SMARTER Balanced Assessment Consortium proposals make claims to the effect that the implementation of their assessment systems as such will improve the performance of individuals and instructional programs through mechanisms that are laid out in the theory of action. For these claims to be plausible, it is necessary that the through-course assessments function well as assessments, but it is also necessary that they function effectively as part of the educational programs in which they are embedded.

Recommendations

For the design and validation of through-course assessment systems, we offer the following recommendations for consideration by the consortia.

Recommendation 1

Focus the validity argument not only on the assessments but also on the theory of action.

Recommendation 2

Be as explicit as possible in the statement of the theory of action, especially in the statement of action mechanisms and intended effects, so as to allow for meaningful evaluation.

Recommendation 3

Take advantage of the flexibility inherent in within-course activities by focusing some components in the through-course assessment on achievement standards that cannot generally be included in the end-of-course examination.

Recommendation 4

In assigning weights to the different components of the through-course assessment system, consider for each component the relative importance of the standards measured, psychometric quality, and proximity to the end of the school year.

Recommendation 5

Collect data from key stakeholders (students, parents, teachers, and administrators) documenting how assessment results are used, noting both intended and unintended consequences of score use.

Recommendation 6

To help in making the case for changes in teaching and learning practice postulated by the theory of action, begin collecting data *now* so that existing practices can be documented.

Recommendation 7

Be sure to evaluate (and attempt to control) threats to validity, including potential differences in the comparability of scores across tasks, raters, and test forms, any of which could undermine the meaning and (especially the consequential) use of scores from through-course assessment systems.

Theory of Action and Validity Argument in the Context of Through-Course Summative Assessment

Randy Elliot Bennett, Michael Kane, and Brent Bridgeman[1]
Educational Testing Service

Kane (2006) offered a detailed elaboration of a two-part, argument-based approach to validation. In that approach the interpretive argument first "… specifies the proposed interpretations and uses of *test results* by laying out a chain or network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the *performances*" (p. 23, emphasis added). The validity argument then provides an evaluation of this interpretive argument.

In the current K-12 context, the explicit focus in Kane's formulation on test results and their uses may not fully address the intended effects of testing programs, for among the primary goals of the *Race to the Top Assessment Program* is positive impact of the assessment procedures, per se, on individuals and institutions (Department of Education, 2010). Thus, in addition to evaluations of score meaning and score-based use, the intended impact of the assessment implementation and its evaluation must, at least in this context, play a major role in validating testing programs.

In response to the need for a broad view of validation, Bennett (2010) suggested that "theory of action" might be a convenient vehicle for explicitly bringing together an assessment program's impact claims *and* its claims for score meaning and use. In this conceptualization, the theory of action included:

[1] We are grateful to Isaac Bejar, Dan Eignor, and Wendy Yen for their helpful comments on an earlier draft of this paper. All positions expressed in this paper are those of the authors and not necessarily those of the reviewers or ETS.

- The components of the assessment system and a logical and coherent rationale for each component, including backing for that rationale in research and theory
- The interpretive claims that will be made from assessment results
- The intended effects of the assessment system
- The action mechanisms designed to cause the intended effects
- Potential unintended negative effects and what will be done to mitigate them

In the current paper, we use a formulation that elaborates Kane's (2006) notion of interpretative argument to encapsulate a theory of action:

> The interpretive argument specifies the proposed interpretations and uses of scores by laying out a chain or network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the scores. To the extent that the assessment itself is intended to change individuals or institutions, the interpretive argument specifies a theory of action for the assessment, including the impact claims and the mechanisms though which that impact is expected to occur.

We apply this formulation to the assessment design of the Partnership for Assessment of Readiness for College and Careers (PARCC) and to that of the SMARTER Balanced Assessment Consortium (SBAC), the two entities funded under the *Race to the Top Assessment Program*. Both designs include what might be considered a version of *through-course assessment*, which the U.S. Department of Education defined as follows:

> *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. (Department of Education, 2010, p. 18178)

Key to this definition are that (a) each individual takes multiple assessments within a subject domain, (b) that these assessments are distributed across the school year, rather than being administered at a single point or within a short time window, and (c) that the results are aggregated to form a single summary judgment for each individual. As will be described, the PARCC design fits this definition well, whereas the proposed SBAC design represents more of a minimal case.

With the basic definition in mind, we next describe the assessment systems envisioned in the PARCC and SBAC winning proposals, including the theory of action underlying each system. Where needed, we supplement the PARCC and SBAC descriptions with our own thinking so as to fashion as complete a picture as possible. Following that background, we give our thoughts for validating through-course assessment systems and outline key components for the requisite interpretive argument, encapsulating the theory of action.

The Proposed PARCC Assessment System

System Components

The PARCC assessment design consists of two main summative components denoted as *through-course assessment* and *end-of-year assessmen*t (PARCC, 2010), with each component to be grounded in the Common Core State Standards (CCSS). For purposes of the interpretative argument, we view the end-of-year assessment as the last in the series of through-course assessments. We also view the CCSS as integral to the series. This treatment is consistent with the PARCC designers' intent to create a coherent assessment system and allows for a common set of intended effects and interpretive claims.

In English language arts/literacy (ELA) and mathematics in grades three through high school, the PARCC design calls for the administration of *focused assessments* after roughly 25% and 50% of instructional time. The assessments are focused in the sense that they target specific content intended to be taught during those portions of the school year. After roughly

75% of instructional time, the design calls for administration of an extended performance task. Finally, students are to take a streamlined computer-based assessment after approximately 90% of instruction is completed. Results from all four of these required components are to be aggregated to form a summary score for each subject (contingent on research and evaluation conducted during the pilot and field test phase around how to do the score aggregation).

PARCC distributes summative assessment across the school year so that "… assessment of learning can take place closer in time to when key skills and concepts are taught and [so] states can provide teachers with actionable information more frequently" (PARCC, 2010, p. 7). The proposal notes that through-course assessment "… will also allow school and district leaders to make any adjustments to instructional strategies and resource allocations throughout the year based on students' progress toward readiness" (PARCC, 2010, p. 7). In this sense, through-course assessment replaces the interim assessments currently used by many school districts.

The PARCC design suggests additional benefits that underlie factoring the through-course results into the overall summative judgment. First, the distributed assessment allows for greater depth and breadth of coverage of the CCSS because more time is allocated for assessment than a single end-of-year test could afford. Second, because of their frequency, these assessments have the potential to keep teacher and student attention more continuously focused on the CCSS, as well as on representations of target performances modeled by the tests that have utility for classroom practice (PARCC, 2010, pp. 38-39). Finally, the design reduces the impact of a single, end-of-year test performance because summative decisions are based on multiple pieces of evidence collected at different points in time.

For ELA, the assessment components are given in Table 1. Note that ELA-5, the speaking and listening through-course assessment, is not included in the aggregation of summative results but, rather, is intended for teacher use in generating grades for report cards at the

Table 1. ELA Assessment Design Components

Assessment	Content focus	Administration	Task types	Purpose
ELA-1 Focused Literacy Assessments: Writing from Sources.	Ability to read increasingly complex texts, draw evidence from them, draw logical conclusions, and present analysis in writing.	After ~25% of instructional time completed. Prompts/items delivered to schools online. Items administered and responses recorded on paper for grades 3-5, and online for grades 6-11.	1-2 extended constructed response; on-demand writing in response to text. Students will be asked to read an appropri-ately complex text and to present analysis in writing.	Monitor student progress toward CCR (college and career readiness). Inform instructional decisions. Signal interventions for individual students. Assess teacher/curriculum effectiveness.
ELA-2 Focused Literacy Assessments: Writing from Sources.	*Same as ELA-1.*	After ~50% of instructional time completed. *Delivery same as ELA-1.*	*Same as ELA-1.*	*Same as ELA-1.*
ELA-3. Extended Research/ Writing Assessment.	Research and analysis skills.	After ~75% of instructional time completed. *Delivery same as ELA-1.*	1 multi-day extended performance task. Over several sessions, the assessment could provide a set of sources and require students to analyze them, evaluate their credibility and relevance, and draw on them as necessary in order to compose a coherent account of a subject or to take and defend a position on a controversial topic.	*Same as ELA-1.*

Assessment	Content focus	Administration	Task types	Purpose
ELA-4: End-of-year assessment	Reading comprehension, vocabulary, language use.	After ~90% of instructional time has been completed Online for all grades.	45-55 computer-scored selected response and "computer-enhanced" items. Draw on higher-order skills (e.g., critical thinking and analysis); assess hard-to-measure skills; use digital technologies to include video or audio clips; and include innovative item types, such as drag-and-drop input to populate lists and highlighting of text in passages. Roughly half of items will be based on long reading passages and between a third and half of the items will be from short passages/speeches. Remaining questions will likely be on vocabulary, language use.	Determine whether students are on-track and have met the standards for the year. Calculate into final grades or student report cards if desired. Data aggregated on a class/school/ district level to assess teacher/ curriculum effectiveness.
ELA-5: Speaking and Listening	Speaking and listening skills, including sharing of findings and evidence; ability to interact with audiences by responding to questions or engaging in discussion or debate; ability to speak clearly; and ability to present information in a logical manner.	After completion of the ELA-3 through-course component. Oral presentation.	1 speaking/ presentation task. Students will be required to present their findings or position from ELA-3, answer questions from others, and use evidence to make points.	Determine whether students have met the speaking and listening standards. Inform instructional decisions as needed. Calculate into student grade or report cards if desired. Not included in annual combined summative ELA literacy score.

classroom level and for parents and teachers to track progress of students' speaking and listening skills.

For math, the assessment components are given in Table 2.

Table 2. Math Assessment Design Components

Assessment	Content focus	Administration	Task Types	Purpose
Math-1	1-2 essential topics that emphasize standards or clusters of standards from the CCSS that play a central role during the first stages of mathematics instruction over the school year.	After ~25% of instructional time completed. Prompts/items delivered to schools online. Items administered and responses recorded on paper for grades 3-5 and online for grades 6-11.	Approximately 2 brief constructed responses per topic and 1 extended constructed response per topic.	Monitor student progress toward CCR. Inform instructional decisions. Signal interventions for individual students. Assess teacher/curriculum effectiveness.
Math-2	*Same as Math-1.*	After ~50% of instructional time completed. *Delivery same as Math-1*	*Same as Math-1.*	*Same as Math-1.*
Math-3	Conceptual understanding of key mathematical concepts and ability to apply math to real-world problems.	After ~75% of instructional time completed. *Delivery same as Math-1*.	1 extended performance task.	*Same as Math-1.*
Math-4: End-of-year assessment	Important knowledge and skills from a range of important topics for the year (and previous year), including the upper and lower ends of the performance continuum. Items will blend conceptual understanding, procedural fluency and problem solving.	After ~90% of instructional time has been completed. Online for all grades.	40-45 computer-scored items, including next-generation selected-response and innovative, computer-enhanced items, such as drag-and-drop input to populate lists or drawing functionality for graphing items.	Determine whether students are on-track and have met the standards for the year. Calculate into final grades or student report cards if desired. Data aggregated on a class/school/ district level to assess teacher/ curriculum effectiveness.

Tables 1 and 2 suggest that the following are among the key attributes of the through-course assessment system proposed by PARCC:

- In both ELA and math, students in all grades 3-11 take the end-of-year test (#4) on computer. The same is true for tests #1-3 for students in grades 5-11. For students in grades 3-5, tests #1-3 will "likely" be delivered on computer but responded to on paper (until such time as it is determined that these students can effectively respond on computer) (PARCC, 2010, pp. 46-48).
- In both ELA and math, tests #1 and #2 consist entirely of a small number of constructed-response (CR) items. The two ELA tests focus on writing based on the reading of complex texts. Each test consists of 1-2 constructed-response questions. In math, these two tests each focus on 1-2 essential topics and have 2 brief constructed-response items per topic and one extended constructed-response item per topic, for a total of 3-6 items, depending on the number of topics included.
- In both ELA and math, test #3 is a performance event. In ELA, the event is described as a single, multi-day task. In math, it is also a single extended, performance task with no time specified.
- The end-of-year test (#4) is a linear, machine-scored assessment having between 40 and 60 items, with results reported in approximately one week. In ELA, this test is intended to cover reading comprehension, vocabulary, and language use (though apparently not writing). In math, this test is intended to be a sampling of important topics from the current year, as well as from the previous year.
- For each student, scores from all assessments (except ELA #5, speaking and listening), are aggregated within domain to form the annual summative result.
- The ELA speaking and listening component (#5) is a live performance task that will be administered by the student's teacher for classroom purposes.

Intended Effects and Hypothesized Action Mechanisms

The PARCC proposal identifies four impacts intended to result from use of the Partnership's assessment system (PARCC, 2010, pp. 37-38). The four impacts are, in turn, intended to increase the rates at which students graduate from high school prepared for success in college or the workforce. The four impacts as stated in the proposal are given below.

1. Reporting achievement results based on a clear definition of college and career readiness will improve outcomes for students.
2. The common assessment system will help make accountability policies better drivers of improvement.
3. Classroom teachers will have an assessment system that provides as much for them as it asks from them--one that functions as an integrated element in a larger system of standards; curriculum; and ongoing collaborative, professional work.
4. The common assessment system will help education leaders and policymakers make the case for improvement and for sustaining education reforms.

Because these impacts appear to be too generally stated for purposes of creating a workable theory of action (e.g., "Classroom teachers will have an assessment system that provides as much for them as it asks from them …"), we have restated the impacts based on the proposal's explications, as well as on other statements distributed throughout that document. Our restatements are given below, segmented into intermediate and ultimate effects. A careful reading of the restated effects should indicate that they capture, albeit at a more specific level, the intent behind the original four impacts. (See Table A-1 for the proposal text upon which our restatements are based.)

- Intermediate effects
  - Teacher has working understanding of the CCSS (including the types of behaviors that indicate whether a student is on track for college and career readiness).
  - Focusing of classroom instruction on the CCSS, including adoption of the curricular structure and sequence of the math CCSS.
  - Frequent modeling of classroom instruction on the "rich and rigorous" performance tasks included on the through-course assessments (e.g., tasks that ask students to write in response to "rich" literary texts or informational sources).
  - Routine use of through-course-assessment results to adjust instruction, including for students who are struggling, meeting expectations, or excelling.
  - Routine use of assessment results to hold educators accountable for achievement of the CCSS.
  - Routine use of through-course-component results to identify mid-year professional-development and support needs for individual teachers and groups of teachers.
  - Routine use of student-growth data to help recruit and hire teachers, as well as to assign teachers to work with students needing the most help.
  - Better understanding of the effectiveness of the education system in each Partnership state relative to other states and countries.
  - Support for needed education reforms.
- Ultimate effects
  - Greater achievement by the end of each year.
  - Increase in the number of students achieving college and career readiness (CCR) by the time they graduate from high school.

The proposal text found in Table A-1 suggests the following types of action mechanisms as leading to these intended effects:

By virtue of being held accountable for their students' achievement of the CCSS, teachers will review and analyze the standards and communicate them to students. As a consequence of this review, analysis, and communication, teachers will develop a better working knowledge of the standards and of the behaviors that signal whether students are on track toward achieving them. Additionally, teachers will focus classroom instruction toward that achievement. Teachers will use lesson plans, assignments, and formative tasks from the Partnership Resource Center (PRC), and in ELA the "text complexity diagnostic tool" (PARCC, 2010, p. 58), as aids in focusing instruction. In math, teachers will use the standards to specify a curricular sequence so that their instruction is tightly coordinated with the topics of the first two through-course assessments.[2]

In both math and ELA, teachers will not only target classroom instruction toward the standards but model instructional activities after the rich and rigorous performance tasks students encounter particularly, we assume, on assessment #3 (the multi-day performance). After each through-course assessment, teachers will use the *Periodic Feedback Reports* and the *Interactive Data Tool* "to understand how their class and individual students are performing on the assessment components relative to the assessed CCSS, at an overall level and by standard, during the school year" (PARCC, 2010, p. 68). Teachers would then use this understanding, supplemented by other data, to adjust instruction for individuals as well as for groups.

School, district, and state officials will review and analyze the CCSS. They will communicate to educators the importance of the standards and of aligning instructional practice with them. Further, those officials will routinely review the *Periodic Feedback Reports* and employ the *Interactive Data Tool,* using assessment results from those sources to hold teachers and principals accountable for student achievement of the standards. Officials will use

[2] No such sequence is apparent that is to be similarly enforced by the ELA through-course assessments.

the through-course-assessment results to provide mid-year professional development to individuals and groups. Officials will also use student-growth data to help recruit and hire teachers, as well as to make teacher assignments.[3]

PAARC, in collaboration with state and local officials, will disseminate assessment data to help policymakers, students, parents, and the public better understand the effectiveness of the education system in each Partnership state relative to other states and countries. In addition, those data will help policymakers build support for needed educational reforms.

Collectively, this set of school-official, teacher, student, parent, and policymaker intermediate effects would be expected to contribute to the ultimate effects of increased student achievement and readiness for college and careers.

The logic diagram given in Figure 1 summarizes the theory of action for through-course assessment in PARCC, as interpreted by us. It includes the PARC through-course assessment components, the hypothesized action mechanisms, and the intermediate and ultimate effects.

The Proposed SBAC Assessment System

System Components

In contrast to PARCC, SBAC does not include through-course assessment as a key aspect of its proposed summative design.[4] While SBAC does not present a through-course design, the

[3] The use of student-growth data in recruitment and hiring would appear to assume that teacher candidates apply for positions with such data in hand and that those data are comparable across teacher candidates.

[4] The Consortium commits only to conducting research on an optional version of the concept, as described in the following passage (SBAC, 2010, pp. 43-44):

"… the Consortium is committed to a comprehensive research agenda to ensure that options for administration of the summative assessment adhere to three key principles--comparability, technical adequacy, and fairness. For example, the Consortium will be investigating the reliability and validity of offering States an optional distributed summative assessment as an alternative for States to the administration of the summative assessment within the fixed 12-week testing window. The distributed summative assessment will be developed 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. The scores of these distributed assessments would be rolled up (along with the students' scores on the performance events) to make the overall decision about students' achievement with respect to college- and career-readiness."

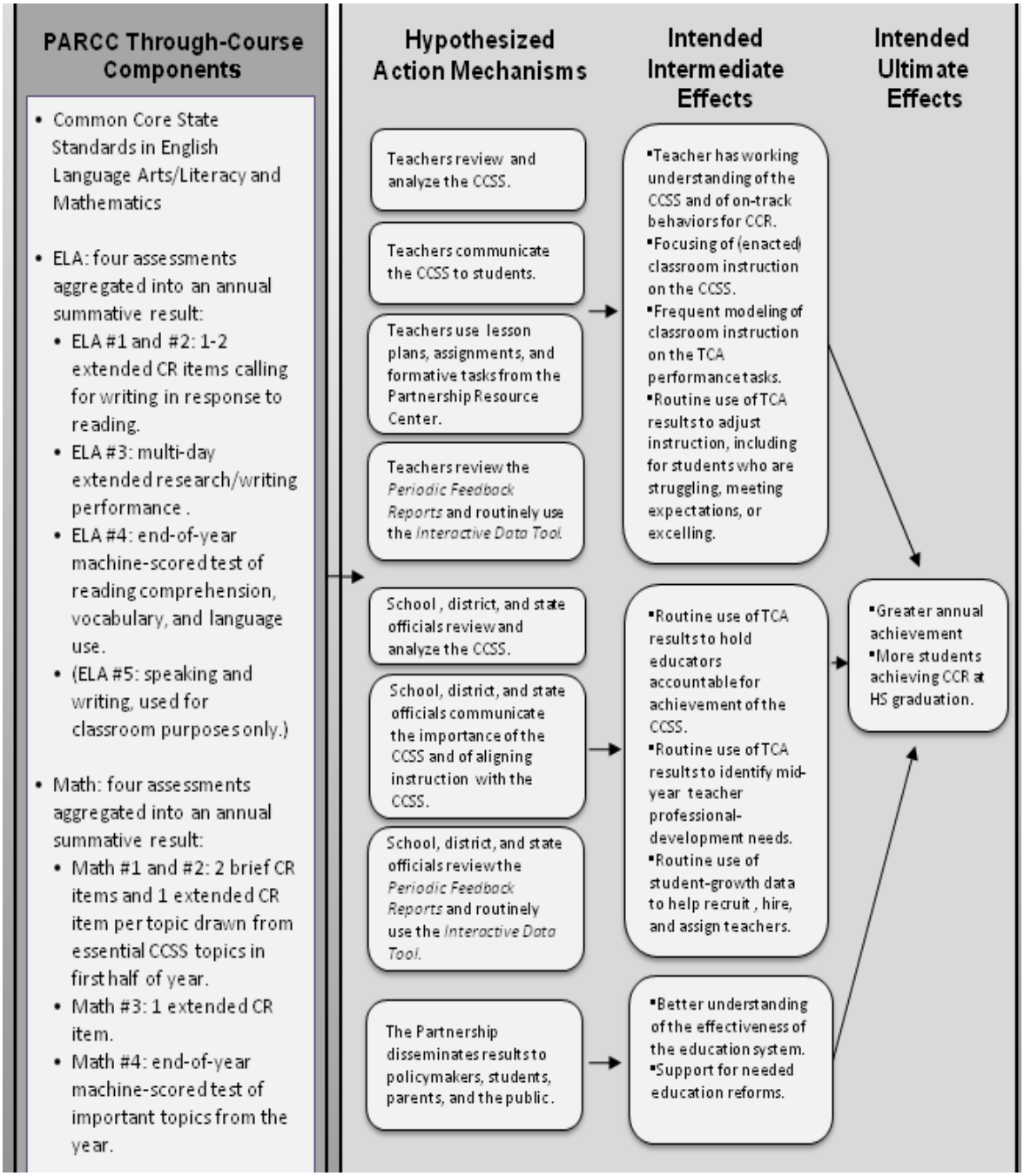

Figure 1. A logic diagram summarizing the PARCC theory of action for through-course assessment (TCA).

Note. CCSS = Common Core State Standards; CCR = College and Career Readiness; CR = constructed response; ELA = English language arts; TCA = through-course assessment.

summative design it does present can be regarded as a special, albeit limited, case of through-course assessment.

The SBAC assessment design consists of two main summative components, with each component separately implemented in ELA and in math. The components are a computer adaptive test (CAT) and a computer-delivered performance event. Each component is based upon the CCSS and is administered toward the end of the school year. Within each of ELA and math, results are aggregated for accountability purposes across the CAT and the performance event.

The SBAC summative assessment components are given in Table 3.

Table 3. Summative Assessment Design Components

Assessment	Content focus	Administration	Task Types	Purpose
ELA computerized adaptive test (CAT): Reading	Important knowledge and skills consistent with CCSS, including complex thinking skills; proposing causes and effects; identifying patterns or conflicting points of view; categorizing, summarizing, or interpreting information; developing generalizations, explanations, justifications, or evidence-based conclusions.	Administered within 12 weeks of end of school year. Grades 3-8 and 11. 1-2 testing opportunities for each student per year.	30 Selected-response items; 3 extended constructed-response items; 7 technology-enhanced items.[a]	Used for measuring achievement and growth for Title I accountability (on track to being CCR).
ELA Performance event: Reading	Ability to integrate skills across multiple standards from the CCSS, depth of understanding, research skills, complex analysis.	Administered within 12 weeks of end of school year. Delivered by computer. 1-2 testing opportunities for each student.	1 performance event in each of grades 3-8. Up to 3 events for all of HS to assess ELA content in the context of science or social studies. Each event takes 1-2 class periods.	Same as above.

Assessment	Content focus	Administration	Task Types	Purpose
ELA Computerized adaptive test (CAT): Writing, Listening and Speaking, and Language	Important knowledge and skills consistent with CCSS, including complex thinking skills; proposing causes and effects; identifying patterns or conflicting points of view; categorizing, summarizing, or interpreting information; developing generalizations, explanations, justifications, or evidence-based conclusions.	Administered within 12 weeks of end of school year. Grades 3-8 and 11. 1-2 testing opportunities for each student per year.	Writing/Language: 6 selected response items; 6 technology-enhanced items; 2 on-demand writing prompts.[a] Speaking/Listening: 2 technology-enhanced items and 2 technology-enhanced constructed-response items (tied to writing performance event in grades 3-8).[a]	Same as above.
ELA Performance event: Writing, Listening and Speaking, and Language	Ability to integrate skills across multiple standards from the CCSS, depth of understanding, research skills, complex analysis.	Administered within 12 weeks of end of school year. Delivered by computer. 1-2 testing opportunities for each student.	Writing/Language: 1 performance event (involving prewriting, drafts, and edits over time) in each of grades 3-8. Up to 3 events for all of HS to assess ELA content in the context of science or social studies. Each event takes 1-2 class periods.	Same as above.
Math Computerized adaptive test (CAT)	Important knowledge and skills consistent with CCSS, including complex thinking skills; identifying patterns; interpreting information; developing explanations, justifications, or evidence-based conclusions.	Administered within 12 weeks of end of school year. Grades 3-8 and 11. 1-2 testing opportunities for each student per year.	19 selected-response items; 3 extended constructed-response items; 18 technology-enhanced items.[a]	Same as above.

Assessment	Content focus	Administration	Task Types	Purpose
Math Performance event	Ability to integrate skills across multiple standards from the CCSS, depth of understanding, research skills, complex analysis.	Administered within 12 weeks of end of school year. Delivered by computer. 1-2 testing opportunities for each student in grades 3-8; 1 opportunity per year in HS.	2 performance events per year in grades 3-8. Up to 6 performance events for HS by the end of grade 11.	Same as above.

Note. Based in part on SBAC (2010) pp. 59-71.
[a] All constructed-response items included in the adaptive test are to be 100% computer scored.

In addition to the summative component, the SBAC proposal gives considerable attention to interim and formative components which are viewed as integral to the assessment system. Table 4 briefly summarizes these components, which include adaptive interim/benchmark assessments and formative tools, processes, and practices.

Table 4. Interim and Formative Design Components in ELA and Math

Assessment	Content focus	Administration	Task types	Purpose
Adaptive interim/ benchmark assessments and performance event bank	Based on learning progressions and/or CCSS content clusters in reading, writing/ language, speaking/ listening, and math.	All computer delivered. Open testing window which States may restrict according to their policies.	*Reading Comprehensive Assessment*: 30 selected response (SR), 3 extended constructed response (ECR), 7 technology-enhanced (TE), 1 performance event (PE). *Reading Cluster Assessment*: 15 SR, 1 ECR, 3 TE, 1 PE	Used to give more in-depth, fine-grained assessment of what students know and can do with respect to a learning progression or CCSS content cluster so as to help adjust instruction.

Assessment	Content focus	Administration	Task types	Purpose
			Writing, Speaking and Listening, and Language Comprehensive Assessment: Writing/Language: 6 SR, 6 TE, 2 writing prompts (WP), 1 PE. Speaking/Listening: 2 TE, 2 ECR *Writing, Speaking and Listening, and Language Cluster Assessment:* Writing/Language: 6 SR, 6 TE, 2 WP, 1 PE Speaking/Listening: 2 TE, 2 ECR *Math Comprehensive Assessment*: 19 SR, 3 ECR, 18 TE, 2 PE *Math Cluster Assessment*: 15 SR, 1 ECR, 3 TE, 1 PE	
Formative tools, process, and practices	CCSS achievement clusters.	At teacher discretion. All computer-delivered except for classroom exercises.	Adaptive test consisting of SR, ECR, and TE items. Teacher-administered performance event and TE items. Teacher-administered classroom exercises.	Build teacher capacity to collect evidence during instruction useful in diagnosing students' learning needs. Measure cluster-level achievement and growth.

Note. Cluster assessments are assumed to target a small group of related standards. Comprehensive assessments are assumed to traverse a broad array of standards.

Tables 3 and 4 suggest that the following are among the key attributes of the assessment system proposed by SBAC:

- Students in grades 3-8 and 11 take a computerized adaptive test and a computer-delivered performance event in reading; writing, listening and speaking, and language; and math.
- All measures are administered toward the end of the school year.
- The performance event will target the integration of skills across multiple CCSS standards, depth of understanding, research skills, and complex analysis. The adaptive test will presumably have a mapping of items to standards that is largely one-to-one, focusing more on breadth than on depth of coverage.
- For each student, results are aggregated for annual summative purposes across the CAT and performance event within ELA and math separately.
- Interim/benchmark assessments, including adaptive measures and performance events, are available for helping teachers gather information relevant to adjusting instruction.
- Several formative tools are provided to help build teacher capacity to collect evidence for diagnosing students' learning needs and to measure cluster-level achievement and growth.

Intended Effects and Hypothesized Action Mechanisms

The SBAC proposal suggests certain effects for the assessment system (SBAC, 2010, p. 37). For the summative assessment components, the intended effects are stated as follows:

- Signal high expectations to students, parents, teachers, administrators, and policymakers.
- Provide efficient, reliable, and valid information across the full range of achievement.

- Engage institutions of higher education (IHEs) at the high school level to ensure that assessments truly reflect a measure of readiness for college and careers.
- Provide explicit measures of student progress toward college-and career-readiness through growth models and criterion-validity studies.
- Promote policy alignment by establishing internationally benchmarked achievement standards that are common across consortium states and that are comparable across multiple consortia.

The proposal also says, "Data from the summative assessments will be used to measure annual achievement and growth, to inform evaluations of teacher and principal effectiveness, and for Title I accountability purposes" (SBAC, 2010, p. 90).

Other statements pertinent to the SBAC assessment system's intended effects are dispersed throughout the proposal narrative and appendices. As for PARCC, we have restated the impacts based on the proposal's explications. A careful reading of the restated effects should indicate that they generally capture the intent behind the original statements. (See Appendix Table A-2 for the proposal text upon which our restatements are based.)

The restated intended effects are as follows:

- Intermediate Effects
    - High learning expectations for all students
    - Teacher and student understanding of the CCSS and of the features of high-quality work that exemplify those standards
    - Routine diagnosis of learning needs and adjustments in instruction
    - Students encouraged to routinely adjust their learning efforts meta-cognitively (e.g., by reviewing progress and acting on the results of that review)
    - Improved teaching practice

  - Better understanding by school, district, and state officials of what students know and can do
  - Routine use of summative results to guide curriculum and professional-development decisions
  - Common policy across states and consortia based on internationally benchmarked standards
  - Support for education reform
- Ultimate Effects
  - Increased student learning
  - All students leave high school ready for college or careers

The proposal text found in Table A-2 suggests the following types of action mechanisms as leading to the intended effects:

At the classroom level, teachers will use curriculum materials and get professional development to support teaching and assessing the CCSS; participate in the design and scoring of test items aligned to the CCSS; review summative results via the consortium's *System Portal* and the *Data Mining Tool*; use adaptive interim/benchmark assessments; and use the formative tools and processes. These action mechanisms will lead to the intermediate effects of teachers having high learning expectations for all students; understanding the CCSS and the features of high-quality work that exemplify those standards; routinely diagnosing learning needs and adjusting instruction; and improving teaching practice. The action mechanisms will also cause students to understand the CCSS and the features of high-quality work, and to routinely adjust their learning efforts meta-cognitively.

At the school/district/state level, school and district officials will receive professional development to become more informed users of assessment data. School, district, and state officials will use the *System Portal* to access information about the CCSS, SBAC activities, and

assessment results; will review summative assessment reports and use the *Data Mining Tool*; and will implement the consortium's communication plan to educate stakeholders about college and career readiness expectations. These action mechanisms will cause several intermediate effects. One such effect is that officials will develop a better understanding of what students know and can do. In addition, officials will routinely use summative results to guide curriculum and professional-development decisions, and to inform evaluations of educator effectiveness.

Finally, two action mechanisms are postulated at the consortium level. These mechanisms are that the consortium communicates the CCSS, policies, and practices to schools, districts, and policymakers; and that the consortium works with institutions of higher education to define college and career readiness. The intermediate effects that these mechanisms produce are a common policy across states and consortia based on internationally benchmarked standards, and public support for education reform.

Collectively, the intermediate effects caused by the action mechanisms will contribute to the ultimate effects of increased student learning and all students leaving high school ready for college or careers.

The logic diagram in Figure 2 summarizes SBAC's theory of action for its assessment system as we interpret it. Note that, although it is not suggested in the SBAC proposal, a through-course component could be created by substituting the adaptive interim "cluster" assessments in ELA and math for the currently proposed summative adaptive test. Such a design might differ from the proposed PARCC design primarily in the absence of an end-of-year test that attempted to cover a given year's standards comprehensively. In considering such a substitution, the potential impact of that absence on the system's intended intermediate and ultimate effects would need to be carefully weighed.

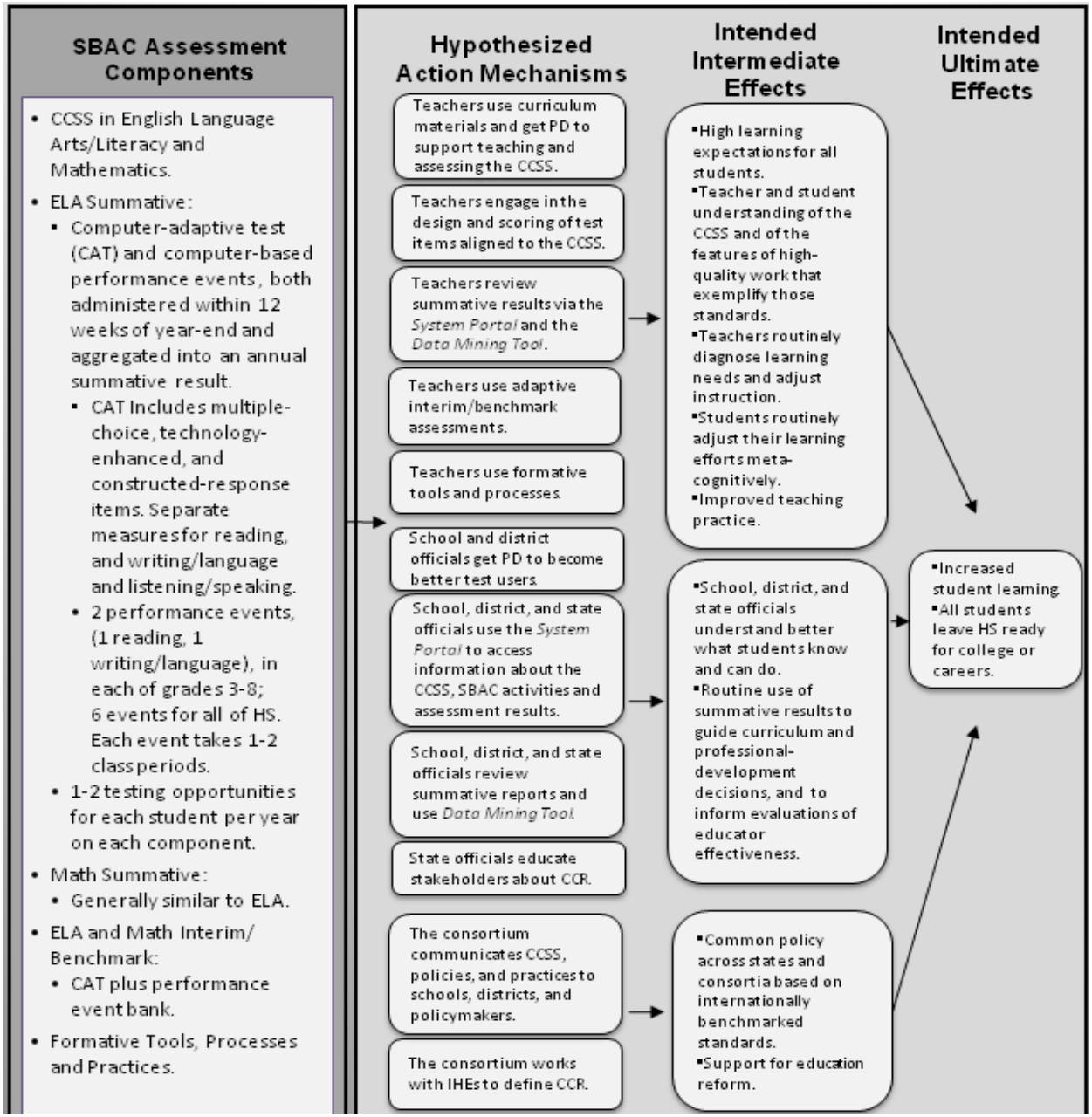

Figure 2. A logic diagram summarizing the SBAC theory of action.
Note. PD = professional development; CCSS = Common Core State Standards; CCR = College and Career Readiness; HS = high school; ELA = English language arts.

Validating Through-Course Assessments

To validate a proposed interpretation and/or use, it is necessary to be clear about what is being claimed by the interpretation and use, and in cases where the assessment is expected to produce certain outcomes, what is being claimed by the theory of action. One way to make the claims explicit is to outline the proposed interpretation and use in the form of an explicit argument, leading from the observations included in the assessment system to the intermediate and final claims being made.

In the context of assessment systems that are also intended to promote the achievement of certain goals (e.g., some teacher practices and student learning), the interpretation and use of scores is likely to be entwined with an instructional program, and the validation of the assessment program will require an evaluation of the effectiveness of the theory of action describing the intended instructional practices and expected outcomes based on the assessment results. In these cases, the validation of the intended use will take on many of the characteristics of a program evaluation.

To the extent that these claims are justified rationally or empirically, the proposed interpretation and use can be considered valid. That is, if the claims being made are adequately supported by evidence, these claims can be accepted, and the proposed interpretation and use can be considered valid.

Rationale for Through-Course Assessments

As indicated earlier, through-course assessments have two structural characteristics that differentiate them from traditional standardized, end-of-course tests. First, because they do not have to fit into the restricted time limits usually specified for standardized tests and because they are not necessarily subject to all of the contextual constraints associated with standardized testing, the “focused assessments” and longer performance tasks (e.g., a research project involving use of the internet and a library) can expand the range of content standards in

the CCSS that can be directly assessed. Combining the results of the focused assessments and performance tasks with scores on the end-of-course assessment can enhance the validity of the final scores based on all of the components in the through-course assessment as measures of achievement in the CCSS. This use of different assessments to measure achievement on different content standards does not necessarily introduce any fundamental measurement problems (Kane & Case, 2004).

Second, the assessments are distributed throughout the course, which implies that our conclusions about a student's achievement in the course are based in part on performance throughout the course and not only on a student's level of achievement at the end of the course. This feature has several desirable characteristics (e.g., modeling desired performances for students, and making it possible to provide some useful feedback to teachers, students, and administrators during the course), but it also has the drawback of making the interpretation of the results as indications of end-of-course achievement a bit murkier than it would otherwise be. As noted below, there are precedents for employing the results of assessments administered throughout the course in a summative score, and to the extent that an interpretive problem does exist, it can be alleviated by judicious choice of weights for the various components. For skill areas in which there is a clear learning progression such that performance on the later tasks assumes and requires mastery of prerequisite skills assessed earlier, it may be appropriate to give the early assessments relatively low weight in the final summative aggregation of results. On the other hand, where skills in a domain do not necessarily build on one another in a clear way (e.g., knowledge of sentence punctuation and knowledge of apostrophe rules) there would be no reason to give more weight to the skill learned later. Note, however, that this does not imply that the separate skills assessed in different through-course assessments be equally weighted, as one set of skills might be more important for ultimate college and career readiness.

Assessment systems akin to through-course assessments have a long history and are widely used. In both high schools and colleges, a commonly used summative measure is the grade-point average, typically computed as a weighted average of grades in all of the courses taken at the institution. The working assumption is that the GPA provides a good overall measure of the student's performance at the institution and his or her level of achievement at the end of the program.

The GPA has at least two virtues. First, it is comprehensive and reflects the student's overall performance, and therefore, it can have the desirable effect of encouraging students to be reasonably diligent in all of their courses. Second, the GPA includes a fairly large number of separate grades (independent observations) involving different relevant content and different graders, and therefore, tends to be more reliable (or generalizable) than the grade in any particular course or small sample of courses. Admittedly, at least for some purposes, an argument could be made for focusing attention on the most recent courses, or on upper-level courses, or on courses in the student's major, but all of these potential choices involve tradeoffs.

A more directly relevant analogy would be to course grades in high school and college, where the final grade is some weighted average over a final exam, a midterm, and research papers and projects. The argument that can be made for these grading practices is threefold. First, the inclusion of different kinds of tasks (particularly extended 'authentic' performances), makes it possible to directly assess more of the full range of competencies (e.g., those needed for argumentation and writing, laboratory and archival research, or extended performances on a musical instrument) developed in the course than could be the case on a single test (of two or three hours). From a psychometric point of view, it can be argued that the inclusion of multiple sources of evidence relevant to the competencies developed in the course provides a stronger basis for interpreting the scores in terms of these competencies than would be the case for scores based on a typical final examination.

Second, the assessment system can be used to draw the students' attention to certain activities, and/or to guide them in prioritizing and scheduling their work. The grading of laboratory work in science courses is, in many cases, clearly designed to focus attention on and engagement with empirical methodology and with observable phenomena, in addition to the more formal treatments encountered in textbooks and other printed materials. Regularly scheduled quizzes, lab reports, and papers provide information to the instructor, discourage procrastination, and provide the students with signposts about activities and performances that are worthy of their attention. From this point of view, the intent is to promote positive outcomes, or at least to discourage the negative outcomes that would result from students' focusing on the narrow range of competencies included in the final exam.

Third, by including information about a larger number of performances, the weighted average over several examinations, projects, and reports can provide a hedge against the possibility that a student's performance on any small set of assessment tasks might not be representative of their overall performance on the full range of performances that are of interest. So a broad sampling of relevant performances (over tasks, over types of tasks, over raters, and over occasions) provides protection against random errors and some kinds of systematic errors, and therefore can enhance the reliability/generalizability and validity of the proposed interpretation.

Interpretive Arguments for Summative Through-Course Assessments in K-12

In this section we will develop a generic version of an interpretive argument for through-course summative assessment in K-12 education. In practice, the interpretive argument for the particular assessment program, as implemented in particular contexts and with particular populations of students, should be examined. As noted above, the assessment designs and the goals for the PARCC and SBAC proposals are different from each other, and both are likely to change as the programs are developed and implemented. So, at this point it seems most appropriate to consider the general characteristics of such through-course

assessments, and to analyze some of the problems that are likely to be encountered in employing them in the K-12 context.

In developing the interpretive argument, it is helpful to work backwards from the intended goals of a testing program to the kinds of decisions to be made, to the interpretation supporting these decisions, and to the inferences and assumptions implicit in the intended interpretations and uses. The interpretive argument can then be played forward in interpreting and using the scores from the assessments.

In addition to providing dependable information on student achievement, the through-course assessment program is intended to have a positive impact on teacher understanding of the CCSS and on teaching practices, and ultimately, on student achievement. By representing the CCSS in concrete form, the through-course assessments are expected to make clear to the teachers and students what is expected in terms of student competencies, and thereby to help focus the instructional program on the CCSS

At the end of the process the results of the summative assessment are to be used to make decisions about students (e.g., whether they have attained some achievement level, in various content areas), and possibly about the effectiveness of educational programs (e.g., classes, schools, districts, or states). In particular, a major goal is to estimate student achievement against a set of broadly defined and ambitious content standards, the CCSS. Presumably, we would like the assessment to provide a reliable and accurate indication of level of achievement in the domain defined by the content standards.

To achieve these goals, the assessment needs to cover the domain as completely as possible. To the extent that the assessment tasks are representative of the domain as a whole, inferences from assessment scores to conclusions about level of achievement in the domain can be treated as simple generalizations from a sample of performances to the domain from which the performances are sampled.

If the assessment omits substantial parts of the domain, the inferences become more complicated and therefore more suspect, because they require conclusions about types of performances that have not been observed as well as the performances that have been sampled. This problem tends to be exacerbated if high stakes for students, teachers, or schools are associated with the assessment results because the stakes will naturally focus attention on those parts of the content domain that are assessed, and thereby, draw attention away from those parts that are not assessed. To the extent that this shift in focus occurs, inferences from the assessment results to conclusions about achievement in the domain can be compromised. In addition, the quality of instruction and student learning may be seriously undermined, especially if the parts of the domain that are not being assessed are considered to be especially important.

Typically, some of the competencies included in content standards (e.g., cognitive, analytic competencies and discrete performances) can be evaluated in a standardized, end-of-course assessment. Assuming that it is well designed and carefully implemented, an assessment consisting of objective items and assessment tasks involving brief responses can provide direct measures of many competencies (e.g., skill in solving arithmetic word problems, knowledge of vocabulary, skill in applying rules of grammar) and can provide indirect measures of some additional competencies (e.g., writing skill as indicated by responses to editing tasks). The standardized, end-of-course assessment can also include some extended-response tasks, but because of time constraints and context constraints, these tasks will necessarily be limited in number and scope.

For competencies involving an extended performance (e. g., designing or conducting an experiment, writing an extended essay or story, producing a work of art), it may not be feasible to include direct assessments in a standardized end-of-course assessment. However, it may be possible to conduct direct performance assessments of these competencies during the course. We can, for example, have students design and conduct an experiment over the course of days

or weeks in a science class, or have students research, write, and revise an extended essay in an English class. The plausibility of an interpretation of the overall assessment scores as measures of achievement in the content standards could be enhanced by including the performance assessment scores as part of the overall summative assessment result.

A relatively full and representative coverage of the CCSS is essential for many of the intended outcomes of the PARCC and SBAC assessment designs. For example, if the assessment results are to be used to hold teachers accountable for achievement of the CCSS and to accurately identify professional development needs for teachers (see Figure 1), it is clearly necessary that the assessment reflect the content of the CCSS. To the extent that the assessment is to help teachers, administrators, and students to understand the CCSS (see Figures 1 and 2), it is necessary that the assessment be representative of the full range of competencies in the CCSS. If important sets of competencies are omitted or get little attention, a distorted view of the CCSS will be communicated. To the extent that the assessments are not representative of the CCSS, they may push teacher practices and student learning in directions that are not consistent with the CCSS.

In addition to the opportunity for a more complete and representative sampling of the competencies, and therefore, a more accurate indication of achievement in the domain, the inclusion of a full range of target-domain tasks should help to avoid the potential problems that arise if students, teachers and schools are encouraged to focus on a narrow range of tasks in the final examination rather than the full target domain associated with content standards.

These considerations suggest that the assessments should provide direct assessments of as much of the performance domain specified by the CCSS as possible, and this conclusion suggests that the early assessments (i.e., those administered before the end-of-course test) should focus on performances that cannot feasibly be included in a single end-of-course assessment. To the extent that the through-course assessment system (including the early assessments and the end-of-course examination) is representative of the full range of

competencies in the CCSS (and the components meet appropriate psychometric standards), it can more defensibly support the kinds of claims summarized in Figures 1 and 2.

Support for such claims is expressed in the form of an interpretive argument. Interpretive arguments for summative through-course assessments have two parts, a *measurement part* and a *theory-of-action part*. These two parts make different kinds of assumptions and claims.

The *measurement part* of the interpretive argument focuses on the interpretation of the scores in terms of expected performance on the CCSS. Assuming that the through-course assessment is representative of the CCSS and provides an unbiased and adequately precise indicator of overall performance in the domain of competencies specified by the CCSS, the measurement argument claims that it would be reasonable to interpret and use scores on the through-course assessment to represent performance on the CCSS. This part of the interpretive argument takes us from observations of performance on the through-course assessment to claims about expected performance on the CCSS; it is similar to the interpretive arguments for other achievement tests, with two additional wrinkles. The through-course assessment is not administered as a single assessment at the end of the course, but is administered at various points during the course (including at the end of the course), and the through-course assessment will include substantially different kinds of tasks.

The *theory-of-action part* of the interpretive argument focuses on the use of the assessments to enhance performance on the CCSS. Many of the intended effects of through-course assessments will follow the traditional measurement-based model and depend on how the scores are used (e.g., for accountability, feedback, placement of students and teachers).

In addition, both the PARCC and SBAC programs make claims to the effect that the implementation of the through-course assessment, as such, will improve the performance of the instructional programs through mechanisms that are laid out in the theory of action. This part of the overall interpretive argument for the through-course assessment makes claims that

are based more on the content, structure, and format of the through-course assessment than on its measurement properties. For these claims to be plausible, it is necessary that the through-course assessments be representative of the CCSS, and that they function fairly well as measurements (i.e., have reasonable reliability, and be free of any substantial sources of systematic error, or bias), but it might not be necessary that each individual through-course component meet the same level of psychometric rigor as applied to high-stakes tests. For purposes of supporting the theory of action, it might be sufficient for any given component to meet the more relaxed psychometric standards typically applied to formative diagnostic tests--unless, of course, the scores on the individual through-course assessments are also to be used for consequential decisions, like judging interim teacher performance. High levels of technical quality would, however, be necessary for the aggregated summative result (on which high-stakes decisions would more likely be based), and this high level of technical quality can be achieved even if some individual components do not meet rigorous psychometric standards (e.g., reliability over 0.9).

Assuming that the theory of action is plausible, it would be reasonable to expect that integrating a through-course assessment that is fully representative of the CCSS with an instructional program that is focused on the competencies outlined in the CCSS, would enhance the effectiveness of the instructional program and improve student achievement on the CCSS. Whether the theory of action works as intended could be evaluated empirically by examining whether the expected changes in teacher behaviors and school climate occur, and whether student achievement improves.

The measurement argument. A typical achievement test-score interpretation can be summarized in terms of a sequence of inferences, and the measurement part of the interpretive argument for the through-course assessment can be summarized in terms of this sequence. The specific inference can be more complicated and may require more support for through-course assessments, but the pattern is basically the same.

Scoring. The first inference derives an observed score for a student by evaluating the student's performance on the assessment using some set of scoring rules. For objective tests, the scoring rule will be very easy to implement; for a multiple-choice test, we basically count the number of responses that agree with the scoring key. For performance tests and essay questions, the scoring rule can be more complicated and more subjective, but its development does not introduce any insurmountable problems.

The scoring rule for the through-course assessment will tend to be more complicated than the scoring rules for most standardized assessments in two ways. First, the through-course assessment will include different kinds of tasks, and therefore, will require different scoring rules for each component. Second the scores on the different components will have to be combined in some way.

One approach to combining the scores on the different components would involve a linear composite of the end-of-course test and the earlier assessments. As part of the development of the through-course assessment and its interpretive argument, it would be necessary to specify the weights in this composite, an issue discussed more fully below.

Generalization. For a test made up of one kind of task (e.g., objective questions, performance tasks), the generalization inference extends the interpretation from the observed set of performances to the expected score for a universe of performances that would be considered acceptable, or exchangeable, given the definition of the testing procedure. In *G* theory, the universe of acceptable observations is referred to as the universe of generalization, the expected value over the universe of generalization is the universe score, and evidence for the generalization inference is collected through *G* studies (Brennan, 2001).

For a weighted-composite through-course assessment, this inference is more complicated than it is for a more homogeneous test, but viable approaches are readily available. One possibility is to evaluate the generalizability of each of the components and then use classical test theory with the estimates of the component generalizabilities to project the

generalizability of the weighted composite (Kane & Case, 2004). A second alternative would use multivariate *G* theory to estimate the generalizability of the weighted composite, with fixed weights, fixed test components, and random tasks/items nested within components.

In either case, the generalizability of the composite scores will depend in part on the weights assigned to the separate components. Generally, the reliability will increase as the weights assigned to the more reliable components increases. However, as discussed in the next subsection, giving too much weight to the most reliable components (probably the end-of-course assessment) will tend to undermine the effectiveness of the theory-of-action part of the argument. Giving too much weight to some components will also tend to undermine the plausibility of extrapolations from the composite universe score to the CCSS.

Extrapolation. Extrapolation extends the inference from the universe score for the measurement procedure to the larger universe of performances that is of interest. For interpretations in terms of broadly defined standards (e.g., the CCSS) and on assessments that are made up of a restricted kind of task/item (e.g., answering objective questions, or writing short essays), extrapolation tends to be questionable. Therefore, extrapolation needs empirical support, because the universe of generalization is at best a small subset of the target universe of performances that is of interest.

For through-course assessments, the extrapolation inference can be strengthened by the inclusion of a sample of performances that is relatively representative of the CCSS and the use of a range of performance contexts and formats. With an assessment based on a representative sample of performances from the CCSS, the extrapolation (or leap) from the universe of generalization to the CCSS-based target universe is potentially more plausible than it would be for a more narrowly defined assessment.

The through-course assessments could achieve good coverage of the target domain by supplementing information from the end-of-course standardized assessment with scores from earlier assessments, which would be given (perhaps at more-or-less fixed times) during the

course. The earlier assessments would be most useful in improving the validity of the overall summative assessment as a measure of achievement in the CCSS domain, to the extent that they focused on topics and skills that are not covered in the end-of-course standardized assessment or that are not adequately covered in that assessment.

Assuming that each of these three inferences--scoring, generalization, and extrapolation-- is warranted, the measurement part of the argument would be warranted. The interpretation would involve an evaluation of a student's observed performance, its generalization to a universe score of the assessment procedures, and an extrapolation to the CCSS.

To the extent that the scores are aggregated to provide indications of class or school-level effects, these hierarchical interpretations will also need to be evaluated. In particular, the reliability of class or school means can be quite different from the reliability of individual student scores (much smaller or much larger). In addition, any causal inferences (e.g., that the teacher is the dominant factor in determining the class mean scores) will require close scrutiny.

In warranting the measurement argument, it is critical to evaluate, and attempt to control, threats to validity, which may undermine the meaning and use of scores from through-course assessment. Of particular concern with reference to the assessment systems proposed by PARCC and SBAC are the comparability of scores across tasks, raters, and test forms. If not controlled, or otherwise accounted for, such effects may manifest themselves in erroneous results, including larger or smaller differences in performance between students, teachers, schools, or years than are really the case.

The theory-of-action argument. As indicated earlier (see Figures 1 and 2), the theories of action associated with both the PARCC and SBAC proposals assume that the use of the through-course assessments as an integral part of the instructional programs will lead to certain actions, or action mechanisms, on the part of teachers and students (e.g., teachers engage in the design and scoring of assessment tasks linked to various parts of the CCSS), that

these actions will lead to certain intermediate effects (e.g., greater teacher and student understanding of the CCSS), and that these intermediate effects will lead to certain ultimate effects, or outcomes (e.g., greater student achievement on the CCSS).

For certain aspects of the theory of action to work well, the measurement properties of the through-course assessments are not a major concern. An individual through-course assessment that adequately samples from specific portions of the CCSS may be very useful in focusing teacher (and by extension student) attention on the sampled standards even if the ultimate measurement of those particular standards is not very reliable. On the other hand, other aspects of the theory of action set a high psychometric bar. "Routine use of [through-course assessment] results to adjust instruction, including for students who are struggling, meeting expectations, or excelling," requires a very reliable indicator to appropriately identify these struggling or excelling students. The same can be said for "routine use of through-course assessment results to hold educators accountable for student achievement of the CCSS."

In order for this theory of action to work well, the measurement part of the interpretive argument has to work reasonably well, at least well enough for the through-course assessments to be accepted as credible and used by the teachers, administrators, students, parents, and the public. However, the theory of action does not rely heavily on having strong support for some of the more technical aspects of the *earlier* through-course assessments (e.g. their generalizability/reliability). Rather, the theory of action depends on the assumptions that the through-course assessments collectively will generate a sample of performances that is highly representative of the CCSS; that the teachers, administrators and students will be very familiar with these assessments; and that aggregated results that are used for high-stakes purposes have technical quality adequate to those purposes.

***Hypothesized action mechanisms*.** Some of the hypothesized action mechanisms (e.g. use of periodic feedback reports) that have been suggested depend on and make use of the outcomes of the through-course assessment, and some of the hypothesized action mechanisms

(e.g., the availability of CCSS-based resources and materials) will operate in conjunction with the through-course assessment, but do not depend directly on through-course assessment scores. These different mechanisms are expected to work in a coherent way to promote the intended intermediate effects, which are, in turn, expected to generate the intended ultimate effects.

The questions to be addressed in evaluating this part of the overall interpretive argument focus on whether the hypothesized action mechanisms are actually implemented as intended, and in this implementation, whether the through-course assessments promote the implementation or retard it. These questions about program implementation have gotten a fair amount of attention in program evaluation, where a basic issue is the extent to which the intended program was actually carried out as designed. The wrinkle in this context is the potentially large role played by the assessments and the interactions between the assessments and the broader set of hypothesized action mechanisms.

***Intermediate effects*.** The intermediate effects (e.g., focusing of instruction on the CCSS, teacher understanding of the CCSS) are assumed to follow from the hypothesized action mechanisms, which include the through-course assessments as integral components. The issues to be addressed here also have roots in program evaluation; they address the question of whether the program as implemented, including the through-course assessments, has had the expected intermediate effects.

The extent to which many of the intermediate effects are realized will depend strongly on the extent to which the through-course assessment is representative of the CCSS, and the coverage of the CCSS target domain on the through-course assessment is enhanced by supplementing the information obtained from the end-of-course test with that obtained from the earlier assessments, which could address standards that are not easily included in end-of-course tests. For example, if the through-course assessment system does include the full range of competencies in the CCSS, and these performances are used to model effective instruction, it

would be reasonable to expect that teachers and students will be familiar with the full range of competencies in the CCSS, and that high learning expectations will be promoted. On the other hand, the goal of generating high learning expectations for all students is not likely to be promoted by the use of the through-course assessment if this assessment does not cover the more ambitious standards in the CCSS.

The claims inherent in the intended intermediate effects could be examined in a variety of ways, perhaps most directly through observational studies of teacher performance, and through qualitative studies of how schools and classrooms change in response to introduction of the through-course assessments. In order to accurately document these changes, studies should begin prior to the implementation of the Common Core State Assessments and continue for at least three years after implementation, as it may take several years for administrators, teachers, and students to adjust to the new system. It is difficult to predict the changes during this adjustment period; they could be increasingly positive as the system became better understood, or initial enthusiasm could wane and apparent positive effects may not be sustained.

***Ultimate effects or outcomes*.** The intended ultimate effects (e.g., increased student learning) are assumed to follow from the intended intermediate effects. The extent to which a program achieves its ultimate goals is also a focus of program evaluation. In the case of through-course assessments, the question is whether the program as implemented, including the through-course assessments, has had the intended ultimate effects and whether these effects are sustainable in the long term.

The extent to which the intended ultimate effects are realized will also depend strongly on the extent to which the through-course assessment system is representative of the CCSS. To the extent that the through-course assessments do include the full range of competencies in the CCSS, and these performances are used to model effective instruction, and thereby focus

more attention on these competencies, it would be reasonable to expect that student performance on the full range of competencies might improve.

This improvement assumption could also, obviously, be examined empirically by evaluating whether performance on the through-course assessment (and possible other, external assessments) changes as a result of the action mechanisms.

Use of Through-Course Assessments to Make Claims About Achievement

One concern in using through-course assessments arises from the fact that the course grade is intended to represent the student's level of achievement after the completion of the course, though an assessment completed during the course would not reflect any learning that occurs after the assessment is completed. This concern certainly has merit, but at least in some cases, it would not be a serious problem for several reasons.

First, we tend to interpret the grade in a course as a plausible indicator of achievement in the course for some time after the completion of the course. Any learning or forgetting that occurs during the interval between the end of the course and the point at which the grade is subsequently interpreted and used (e.g., in placing the student into a course the following year) is likely to be considered acceptable unless there is evidence of a problem in doing so.

Second, the earlier assessments included in the through-course series would presumably be administered after most of the instruction on the content covered by each such assessment had been completed. Assessments that are given very early in the course and/or assessments that are followed by substantial instruction on the competencies covered by the test could also be given relatively low weights in the composite.

Third, as indicated above, the gain inherent in using through-course assessments to more fully cover the CCSS can be substantial, and therefore any uncertainty introduced by conducting an assessment component during the course, rather than at the end of the course, may be relatively minimal. The uncertainty introduced by the fact that earlier assessments are

followed by additional instruction and learning can be kept small, in part, by not giving these earlier assessments a lot of weight in the overall summative score.

Dangers and Unintended Consequences

An evaluation of the utility of through-course assessments must focus on potential dangers to the educational system that could be introduced and on unintended as well as intended consequences.

Construct underrepresentation. One of the classic issues in test validation is when the test fails to measure some important aspect of the intended construct. In their proposal executive summaries, both PARCC and SBAC state a clear goal of preparing students for college and the workforce: "The overarching goal of the SBAC is to ensure that all students leave high school prepared for postsecondary success in college or a career…"(SBAC, 2010), and "State leaders…share one fundamental goal: building their collective capacity to dramatically increase the rates at which students graduate from high school ready for academic success" (PARCC, 2010). Nevertheless, the proposed assessments and related theories of action address only skills in English language arts (ELA) and mathematics. Aside from some ELA assessments "in the context of science," scientific knowledge and scientific thinking, for example, are ignored, but it would be hard to argue that these skills are unimportant for success in 21st century jobs. Although the proposed assessments are no worse than the current *NCLB* tests in this respect, this underrepresentation does not mean that a comprehensive effort to validate the assessments and theories of action can ignore this issue. Unfortunately, there is no clear way to address the seriousness of this problem. The state-by-state longitudinal picture provided by the National Assessment of Education Progress (NAEP) is of limited value in focusing attention on areas outside of ELA and math because baselines set in the *NCLB* era also excluded these areas. International assessments such as the Programme for International Student Assessment (PISA) may provide a better indicator of how well students in the United States are prepared in these areas compared to their international competitors in the coming "flatter" world.

Use of through-course assessments to adjust instruction. The use of through-course assessments to adjust instruction is certainly a key feature of the theory or action, but such adjustments can have negative as well as positive consequences, especially if teachers are not given very clear guidance in how to use the diagnostic information provided. A possible scenario is that a teacher is given feedback that her or his students are below average in skill A and average in skill B. In many assessments, this feedback is based on only a few questions in each skill area so that the resulting skill determinations are highly unreliable. It is obvious that making decisions about students based on very unreliable information is ill advised. In the current case, a wrong decision can lead to significant opportunity costs (e.g., time wasted remediating skills that didn't need as much attention as other skill deficiencies).

Less obvious is the desired course of action when the diagnostic scores are in fact highly reliable. If the skill-level characterizations suggested above were indeed based on a very reliable assessment, a teacher might conclude that more attention is needed on skill A than on skill B. This would be true *if and only if* skills A and B were equally important in moving students to the ultimate goal of being college and career ready. Suppose skill A concerned the appropriate use of the apostrophe with plural nouns and skill B concerned sentence fragments and run-on sentences. It might be a better use of limited time to move students from average to above-average competency on the latter skill than to move them from the below-average to average competency on the former skill. Thus, the validation of the theory of action with respect to using through-course assessment results to "adjust instruction" must evaluate not only if the results were used, but if they were used appropriately. Answering this question is likely to be difficult as there has been much more research on the reliability of diagnostic scores than on their appropriate use.

In the PARCC model, periodic feedback reports, based on the through-course assessments, will be made available not only to teachers and school leaders, but also to students and parents. Similarly, the interim/benchmark reports in the SBAC model will be made

available to students and parents. Although the hypothesized mechanisms and intended intermediate effects in the theory of action emphasize actions by teachers, the actions of parents and students should also be evaluated, as an inappropriate use of scores by parents or students could undercut effective teacher strategies. Continuing the above example, the teacher might recognize the importance of emphasizing instruction on skill B, but these efforts might be unintentionally sabotaged by parents who helped their children only on the apparently weakest skills identified in the assessment.

Test security. *NCLB* tests were typically low-stakes tests for students, though they may have been perceived as higher-stakes tests by teachers and administrators. Most of the identified breaches of test security indeed were at the teacher—and administrator—level and not at the student-level. Security-related threats to the validity of *Race to the Top Assessment Program* through-course assessments may be more serious than for *NCLB* tests for two reasons. First, the stakes at the student level are higher. An explicitly stated goal is for the assessments to be useful for placement into college courses. Thus, students could potentially save thousands of dollars by doing well on high-school through-course assessments used for college placement. Second, the nature of the performance events that may require or allow work outside of the classroom may provide an opportunity for students to present the work of others as their own. These problems are addressable, and we are not suggesting that performance events should be avoided, but it is important to include attention to these issues in the overall validation scheme.

Summary

This paper reviewed the proposed assessment system designs for PARCC and SBAC. In this review, particular attention was paid to the concept of through-course assessment from the perspective of validity argument. Of note is that the paper incorporated the notion of theory-of-action into validity argument because both the PARCC and SBAC assessment designs intend, as a primary goal, that the assessments, as such, have positive effects on individuals and institutions. A key theme of this paper was that, for validation to proceed, careful explication is

needed of what the proposed effects are and of how they are to be achieved. The evaluation of those impact claims, along with more conventional claims around score meaning and use, and of potential unintended consequences of the assessment programs are all essential to a meaningful validity argument in the *Race to the Top Assessment Program* context.

Recommendations

For the design and validation of through-course assessment systems, we offer the following recommendations for consideration by the consortia.

Recommendation 1

Focus the validity argument not only on the assessments but also on the theory of action.

Recommendation 2

Be as explicit as possible in the statement of the theory of action, especially in the statement of action mechanisms and intended effects, so as to allow for meaningful evaluation.

Recommendation 3

Take advantage of the flexibility inherent in within-course activities by focusing some components in the through-course assessment on achievement standards that cannot generally be included in the end-of-course examination.

Recommendation 4

In assigning weights to the different components of the through-course assessment system, consider for each component the relative importance of the standards measured, psychometric quality, and proximity to the end of the school year.

Recommendation 5

Collect data from key stakeholders (students, parents, teachers, and administrators) documenting how assessment results are used, noting both intended and unintended consequences of score use.

Recommendation 6

To help in making the case for changes in teaching and learning practice postulated by the theory of action, begin collecting data *now* so that existing practices can be documented.

Recommendation 7

Be sure to evaluate (and attempt to control) threats to validity, including potential differences in the comparability of scores across tasks, raters, and test forms, any of which could undermine the meaning and (especially the consequential) use of scores from through-course assessment systems.

References

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8,* 70-91.

Brennan, R. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Department of Education. Overview information: Race to the Top Fund Assessment Program; Notice inviting applications for new awards for fiscal year (FY) 2010. 75 Fed. Reg., 18171-18185. (April 9, 2010).

Kane, M. (2006). Validation. In R.Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport , CT: Greenwood Publishing.

Kane, M., & Case, S. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221-240.

Partnership for Assessment of Readiness for College and Careers. (2010). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the Top Comprehensive Assessment Systems Competition*. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

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

Appendix

Table A-1. Restated Intermediate and Ultimate Intended Effects of Through-Course Assessment in PARCC and Their Bases

Restated intended effect	Proposal statement and page
Intermediate Intended Effects	
Teacher working understanding of the CCSS, including of the types of behaviors that indicate whether a student is on track for college and career readiness.	pp. 38-39 *Teachers*. PARCC's common assessments will help teachers improve their instructional practices so that all students can achieve college and career readiness by the time they graduate from high school in several important ways. Providing teachers with a clearer picture of what students should know and be able to do to demonstrate that they are ready or on track to readiness through assessments that measure the full breadth and depth of the CCSS.
Focusing of classroom instruction on the CCSS, including adoption of the curricular structure and sequence of the math CCSS.	pp. 51-53 *Math-1 and Math-2. Focused Assessments of Essential Topics.* The first two through-course components emphasize standards or clusters of standards (i.e., one to two essential topics) from the CCSS that play a central role during the first stages of mathematics instruction over the school year. These include standards that are prerequisites for others at the same grade level, as well as standards or clusters of standards for fields of study that first appear during the grade in question. Thus, instead of surveying an overly broad mathematical landscape as typical interim assessments currently do, **these components will promote the coherent curricular structure embedded in the CCSS**. (Bold emphasis added.) p. 55 *Measuring Essential Topics in Depth.* The through-course design supports the structure emphasized in the CCSS. One limitation of traditional summative tests has been that the survey approach to testing topics at the end of the year provides little guidance to help teachers plan instruction and few useful data to adapt and differentiate instruction. The introductions to each grade level of the CCSS mathematics standards set clear priorities. The Partnership will take advantage of this feature of the CCSS by focusing on a few essential topics in each of the two focused through-course components. For example, a component might focus on multiplication and area in grade 3 or both—the relationship of multiplication to rectangular area.

Restated intended effect	Proposal statement and page
Frequent modeling of classroom instruction on the "rich and rigorous" performance tasks included on the through-course assessments (e.g., tasks that ask students to write in response to "rich" literary texts or informational sources). Routine use of through-course-assessment results to adjust instruction, including for students who are struggling, meeting expectations, or excelling.	pp. 38-39 *Teachers*. PARCC's common assessments will help teachers improve their instructional practices so that all students can achieve college and career readiness by the time they graduate from high school in several important ways. Signaling what good instruction should look like through rich and rigorous through-course performance tasks that model the kinds of activities and assignments that teachers should incorporate into their classrooms throughout the year. Providing data more rapidly on students' academic strengths and weaknesses with a quick turnaround of assessment results. Helping teachers identify gaps in students' knowledge in time to adjust plans for instruction during the next quarter, provide extra support to students who are struggling, or provide academic stretch to those students meeting or exceeding readiness targets. p. 207 Have the performance tasks in the through-course assessments inspired teachers to use similar instructional strategies during the rest of the year? For example, to what extent are ELA/literacy teachers asking students to produce writing in response to rich literary texts or informational sources as students are asked to do on the short and long through-course tasks?
Routine use of assessment results to hold educators accountable for achievement of the CCSS. Routine use of through-course-component results to identify mid-year professional-development and support needs for individual teachers and groups of teachers. Routine use student-growth data to help recruit and hire teachers, as well as to assign teachers to work with students needing the most help.	p. 40 This means education leaders can use the assessment system data to *hold school professionals accountable for outcomes as well as inputs*, *more strategically manage human resources and make much better-informed personnel decisions*. Because Partnership assessments will model the kinds of activities that students should be engaged in throughout the school year, school and district leaders can motivate educators to focus their efforts on instructional practices that help students learn the knowledge and skills most important for postsecondary success when they value the assessment results in their evaluation systems. Results from the through-course components can also help school officials identify mid-year professional development and support needs for individual educators, teachers in a particular grade or educators across a whole school. Likewise, data on student growth can help principals and administrators recruit, select and hire experienced teachers based on those who help students achieve the most growth toward college and career readiness, as well as for assigning the most effective teachers to work with students who need the most help getting on track or staying on track.

Restated intended effect	Proposal statement and page
Better understanding of the effectiveness of the education system in each Partnership state relative to other states and countries. Support for needed education reforms.	p. 38 Readiness and on-track scores on the Partnership's assessment system will, for the first time, allow parents, students, policymakers and the public to compare students' performance against students in all 26 Partnership states and in other countries—and against a widely shared benchmark of postsecondary readiness. These kinds of unprecedented comparisons will powerfully demonstrate the progress students are making toward the Partnership-wide goal, and those comparisons will help policymakers make the case for reforms needed to raise student achievement in their states. This will be buttressed by the Partnership's work to ensure our assessments produce internationally benchmarked results, allowing student performance to be compared with the performance of students in high-performing countries. International comparisons are particularly important to business leaders and governors, but will help parents, educators and the public at large better understand the performance and progress of the education system in each state.
Ultimate effects	
Greater achievement by the end of each year. More students than before achieving college and career readiness by the time they graduate from high school.	pp. 38-39 Through these mechanisms, the Partnership's assessment components will serve as the foundation for a system of continuous improvement in classrooms and schools. As a result, instruction will improve, students will begin to learn more by the end of each year than they would have otherwise and academic achievement will rise. A higher proportion of the Partnership's more than 31 million students will get on track and stay on track to become ready for college and careers, and more students will reach the end of high school highly prepared to meet the challenges of postsecondary education and the modern workplace.

Table A-2. Restated Intermediate and Ultimate Intended Effects of the SBAC Assessment Design and Their Bases

Restated intended effect	Proposal statement and page
Intermediate intended effects	
High learning expectations for all students.	p. 37 The summative assessment will accomplish the following: 1. Signal high expectations to students, parents, teachers, administrators, and policymakers. p. 5 The overarching goal of the SBAC is to ensure that all students leave high school prepared for postsecondary success in college or a career... Appendix A2-1 A high quality adaptive summative assessment system, including performance events, establishes high expectations and provides relevant information on achievement and growth to teachers, students, and others.
Teacher and student understanding of the CCSS and of the features of high-quality work that exemplify those standards.	p. 32 Teachers must also be involved in the formative and summative assessment systems so that they deeply understand and can teach in a manner that is consistent with the full intent of the standards, while becoming more skilled in their own assessment practices. p. 33 Assessment *as*, *of*, and *for* learning is designed to develop understanding of what learning standards are, what high-quality work looks like, what growth is occurring, and what is needed for student learning.
Improved teaching practice.	Appendix A4-2, Chart titled, "Overview of Development Process for Summative Assessment" The chart shows the follow set of causes and effect: The CCSS underlies two activity streams, Test Design and Development, and Professional Capacity Building. The former stream involves designing and developing an adaptive summative assessment and the latter stream involves the use of formative tools processes, and practices, and curriculum development. Together these activities lead to improved teaching (and, ultimately, increased learning). p. 32 Teachers must also be involved in the formative and summative assessment systems so that they deeply understand and can teach in a manner that is consistent with the full intent of the standards, while becoming more skilled in their own assessment practices.

Restated intended effect	Proposal statement and page
Teachers routinely diagnose learning needs and adjust instruction. Students routinely adjust their learning efforts meta-cognitively. School, district, and state officials understand better what students know and can do. Routine use of summative results to guide curriculum and professional-development decisions. Common policy across states and consortia based on internationally benchmarked standards. Support for education reform.	p. 33 Reporting of assessment results is timely and meaningful—offering specific information about areas of performance so that teachers can follow up with targeted instruction, students can better target their own efforts, and administrators and policymakers can more fully understand what students know and can do, in order to guide curriculum and professional development decisions. p. 33 Assessment *as*, *of*, and *for* learning is designed to develop understanding of what learning standards are, what high-quality work looks like, what growth is occurring, and what is needed for student learning. p. 37 The summative assessment will accomplish the following: 5. Promote policy alignment by establishing internationally benchmarked achievement standards that are common across Consortium States and that are comparable across multiple consortia. p. 115 Effective communication is critical in the short term to signal change, and over the longer term to build continuing support for reform. SBAC is committed to transparency and clarity in communicating to all stakeholders (e.g., legislators, policymakers, IHEs, the workplace, community members, parents, educators, and students) the principles of the Consortium, the purposes of each assessment component in a balanced system, and the practices and intended outcomes of this assessment system.
Ultimate effects	
Increased student learning. All students leave high school ready for college or careers.	p. 5 All students leave high school prepared for post-secondary success in college or a career through increased student learning and improved teaching. p. 102 (footnote) That is, the student is prepared for success, without remediation, in credit-bearing entry-level courses in an IHE, as demonstrated by an assessment score that meets or exceeds the Consortium's achievement standard for the final high school summative assessment in English language arts or mathematics. In addition, the Consortium expects students to be ready to enter and advance in a job or succeed in advanced training for a high-skill, high-wage job; able to read, comprehend, interpret, analyze and locate complex technical materials; use mathematics to plan, set priorities, and solve problems in the workplace; and pass a State-approved industry certification or licensure exam in the field.