THE NATIONAL CONFERENCE ON NEXT GENERATION ASSESSMENT SYSTEMS

Developing an Internationally Comparable Balanced Assessment System That Supports High-Quality Learning

Linda Darling-Hammond and Ray Pecheone

With Ann Jacquith, Susan Schultz,
Leah Walker, and Ruth Chung Wei

Center for
K–12 Assessment
& Performance Management

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Developing an Internationally Comparable Balanced Assessment System That Supports High-Quality Learning

Linda Darling-Hammond and Ray Pecheone

With Ann Jacquith, Susan Schultz,
Leah Walker, and Ruth Chung Wei

Stanford University

The ideas in Part 1 of this paper are organized around the principles reflected in the Memorandum of Understanding for a Balanced Assessment System, signed by representatives of 38 states. What follows are responses that are consistent with those principles and informed by conversations with design experts and members of the consortium, but they are not consortium decisions or policy and consequently are subject to evolution and change.

Contemporary efforts to create a set of Common Core Standards in the United States have been grounded in a desire to create more internationally competitive expectations by benchmarking learning objectives to those in high-performing nations abroad. Over the last two decades, all 50 states have developed standards for learning and tests to evaluate student progress. No Child Left Behind reinforced using test-based accountability to raise achievement, yet the United States has fallen further behind on international assessments of student learning since the law was passed in 2001.

On the Programme in International Student Assessment (PISA) tests in 2006, the United States ranked 35th among the top 40 countries in mathematics and 29th in science, a decline in both raw scores and rankings from three years earlier, which in turn was a decline since the first PISA results in 2000. Furthermore, in each disciplinary area tested, U.S. students scored lowest on the problem-solving items. The United States also had a much wider achievement gap than the most highly ranked jurisdictions, such as Finland, Canada, Hong Kong, Korea, the Netherlands, Australia, New Zealand, and Japan (Organisation for Economic Co-operation and Development, 2007). Even in the international PIRLS assessments of reading, where the U.S. scores relatively higher, scores have declined over the last several years (Mullis, Martin, Kennedy, & Foy, 2007).

Policy discussions in Washington often refer to these rankings when emphasizing the need to create more internationally competitive standards by benchmarking expectations in the U.S. to those in high-performing nations. The Common Core Standards initiative aims for standards that are fewer, higher, and deeper, based on analyses revealing that higher-achieving countries typically teach fewer topics more deeply each year; focus more on reasoning skills and applications of knowledge; and have a well-worked-out sequence of expectations grounded in developmental learning progressions within and across domains (see, for example, Mullis, Martin, Kennedy, & Foy [2007] and Valverde & Schmidt [2000]).

However, we must also examine *how* these topics are taught and assessed—so that we understand how other countries' education systems shape what students actually learn and can do. European and Asian nations that have steeply improved student learning have focused explicitly on creating curriculum guidance and assessments that focus on teaching central concepts in the disciplines in a thoughtfully organized way, as well explicitly higher-order cognitive skills: the abilities to find and organize information to solve problems, frame and conduct investigations, analyze and synthesize data, apply learning to new situations, self-monitor and improve one's own learning and performance, communicate well in multiple forms, work in teams, and learn independently.

Curriculum differences are reinforced by sharp divergence between the forms of testing used in the United States and those used in higher-achieving countries. Whereas U.S. tests rely primarily on multiple-choice items that evaluate recall and recognition of discrete facts, examinations in most high-achieving countries use primarily open-ended items that require students to analyze, apply knowledge, and write extensively. Furthermore, these nations' growing emphasis on project-based, inquiry-oriented learning has led to an increasing prominence for school-based tasks, which include research projects, science investigations, development of products, and reports or presentations about these efforts.

Because these assessments are embedded in the curriculum, they influence the day-to-day work of teaching and learning, focusing it on the use of knowledge to solve problems. Cognitively demanding performance tasks are incorporated into examination scores in systems as wide-ranging as the General Certificate of Secondary Education in Britain, the Singapore Examinations system, the Certification systems in Victoria and Queensland, Australia, the new examinations system in Hong Kong, and the International Baccalaureate Programme, which functions in more than 100 countries around the world.

As explained by the Hong Kong Education and Assessments Authority, which is increasing the use of school-based assessments in its examination system:

> The primary rationale for school-based assessments (SBA) is to enhance the validity of the assessment, by including the assessment of outcomes that cannot be readily assessed within the context of a one-off public examination, which may not always provide the most reliable indication of the actual abilities of candidates…. SBA typically involves students in activities such as making oral presentations, developing a portfolio of work, undertaking fieldwork, carrying out an investigation, doing practical laboratory work or completing a design project, help students to acquire important skills, knowledge and work habits that cannot readily be assessed or promoted through paper-and-pencil testing. Not only are they outcomes that are essential to learning within the disciplines, they are also outcomes that are valued by tertiary institutions and by employers. (Hong Kong Education and Assessment Authority, 2009, para 2)

A similar point was made by Achieve, a national organization of governors, business leaders, and education leaders, when it called for a broader view of assessment:

> States … will need to move beyond large-scale assessments because, as critical as they are, they cannot measure everything that matters in a young person's education. The ability to make effective oral arguments and conduct significant research projects are considered essential skills by both employers and postsecondary educators, but these skills are very difficult to assess on a paper-and-pencil test. (Achieve, 2004, p. 3)

The impetus to measure the skills that really matter for college- and career-readiness undergirds the political and educational press for new standards and assessments. As President Obama said in March of 2009:

> I am calling on our nation's Governors and state education chiefs to develop standards and assessments that don't simply measure whether students can fill in a bubble on a test, but whether they possess 21st century skills like problem-solving and critical thinking, entrepreneurship and creativity. (Obama, 2009, para 21)

Part I: An Assessment System that Promotes High-Quality Learning

Our priorities for what a new assessment system should accomplish are rooted in this concern for valid assessment of the deep disciplinary understanding and higher-order thinking skills that are increasingly demanded by a knowledge-based economy. They are also rooted in a belief that assessment must support ongoing improvements in instruction and learning, and must be educative for all members of the educational enterprise: students, parents, teachers, school administrators, members of the public, and policymakers.

Priorities for Assessment

Our assessment proposal is shaped by a set of principles that are shared by the systems of high-achieving nations, as well as a number of high-achieving states in the United States. These systems include the following:

- *Assessments are grounded in a thoughtful, standards-based curriculum and are managed as part of a tightly integrated system* of standards, curriculum, assessment, instruction, and teacher development. Large nations like Canada, China, and Australia manage curriculum and assessments at the state or provincial level, while small nations like Singapore and England—which have school populations about the size of Kentucky and California, respectively—have national systems managed by a Ministry of Education. Each of these jurisdictions has undertaken a careful process of developing standards (generally described as curriculum expectations) and curriculum guidance, often in the form of syllabi, to guide teachers' instruction in the classroom, as well as professional development that is organized around the curriculum.

Curriculum guidance is lean, but clear and focused on what students should know and be able to *do* as a result of their learning experiences. Assessment expectations are described in the curriculum.

Curriculum and assessments are organized around a well-defined set of learning progressions along multiple dimensions within subject areas. These guide teaching decisions, classroom-based assessment, and external assessment.

Teachers and other curriculum experts are involved in an extensively vetted curriculum development process and in the process of developing assessments grounded in the curriculum standards. These guide professional learning about curriculum, teaching, and assessment. Thus, everything that comes to schools is well-aligned and pulling in the same direction.

- *Assessments include* evidence of *<u>actual student performance</u> on challenging tasks that evaluate standards of 21st century learning.* Curriculum and assessments seek to teach and evaluate knowledge and skills in authentic ways that examine a broad array of skills and competencies and generalize to higher education and multiple work domains. They emphasize deep knowledge of core concepts within and across the disciplines, problem solving, collaboration, analysis, synthesis, and critical thinking.

  As a large and increasing part of their examination systems, high-achieving nations use open-ended performance tasks and school-based, curriculum-embedded assessments to give students opportunities to develop and demonstrate higher-order thinking skills: the abilities to find and organize information to solve problems, frame and conduct investigations, analyze and synthesize data, and apply learning to new situations. The curriculum and assessment systems evaluate students' abilities in projects, group work, open-ended tasks, and oral presentations, as well as examinations that include essays and open-ended tasks and problems, as well as selected response items, usually given at the end of a course or year.

- *<u>Teachers are integrally involved</u> in the development of curriculum and the development and scoring of assessments* for both the on-demand portion of state or national examinations and local tasks that feed into examination scores and course grades. States invest in extensive moderation of the scoring process to ensure consistency and to enable teachers to deeply understand the standards and to develop stronger curriculum and instruction. The moderated scoring process is a strong professional learning experience, and officials believe teacher involvement drives the instructional improvements that improve student learning, as teachers become more skilled at their own assessment practices and their development of curriculum to teach the standards. The assessment systems are designed to increase the capacity of teachers to prepare students for the demands of college and career in the 21st century.

- *Assessments are structured* to continuously improve teaching and learning. Assessment *as*, *of*, and *for* learning is designed to develop understanding of what learning standards are, what high-quality work looks like, and what is needed for student learning. It is also designed to foster instruction that supports transferable knowledge and skills. These outcomes are enabled by several features of assessment systems:

  The use of school-based, curriculum-embedded assessments provides teachers with models of good curriculum and assessment practice, enhances curriculum equity within and across schools, and allows teachers to see and evaluate student learning in ways that can feed back into instructional and curriculum decisions.

  Close examination of student work and moderated teacher scoring of both school-based components and externally developed open-ended examinations are sources of ongoing professional development that improve teaching.

  Developing both school-based and external assessments around learning progressions allows teachers to see where students are on multiple dimensions of learning and to strategically support their progress.

- *Assessment systems are designed to emphasize the validity and quality* of external assessment, with their primary aim to drive high-quality learning of ambitious intellectual skills. In order to maintain investments in high-quality, well-vetted expert processes of development and scoring, most countries implement external tests for students only once or twice prior to high school (generally around Grades 3 and/or 6), with continuous school-based assessment throughout these years. High-school examinations are generally selected from an array of subjects by students to demonstrate their areas of competence for colleges and employers.

- *Assessment and accountability systems use multiple measures to evaluate students and schools.* High-achieving countries use multiple measures to evaluate skills and knowledge needed for the demands of this dynamic, technological era. Students engage in a variety of tasks and tests that are both curriculum-embedded and on-demand, providing many ways to demonstrate and evaluate their learning. These are combined in reporting systems at the school and beyond the school level. School reporting and accountability is also based on multiple measures, including student achievement measures as one indicator among many. Other indicators often include student participation in challenging curriculum, progress through school, graduation rates, college-going, citizenship, safe and caring climate, and other indicators of school success and improvement.

- *Assessment and accountability systems are used primarily for information and improvement.* In most of these systems, student assessments are used to inform course grades, colleges, and employers, supports for individual student learning, and to shape curriculum improvement. The tests are typically not used to determine student graduation from high school; they set a higher standard linked to college and career expectations.

The fact that assessments are lower stakes allows them to be of higher quality, both in terms of the range of ways in which learning is to be measured and the performance standards that are set. The concept is that students should be striving for high standards, that performance should be measured along an extended continuum that provides substantial information about performance (not just against a single cut-point) and that this information should guide instructional decisions for individual children as well as curriculum and instructional improvement for schools and the system as whole.

This does not mean there are no consequences for poor performance. Outcomes are publicly reported, and the information is taken into account in a well-designed set of systems that focus on continual improvement for schools, including changes guided by school inspections and professional development supports organized by the Ministry or Department of Education.

A Conception of Powerful Learning and How to Get There

In the description above, we place considerable emphasis on the role of curriculum as a key lever for translating desired learning goals into assessments, instructional guidance, and professional development. We do that for two reasons.

First, the curriculum expresses the kinds of learning that is sought. The way new standards and assessments conceptualize and pursue knowledge, skills, and learning is critically important for the outcomes of education, and we believe that well-designed curriculum is central to enacting this conception.

Genuine readiness for college and 21st century careers requires students to find, evaluate, synthesize, and use knowledge in new contexts, to frame and solve non-routine problems, and produce new products. To meet these challenges, students need in-depth understanding of robust, *transferable* knowledge organized around big ideas and opportunities to explore the interconnections among key concepts. It also requires students to acquire well-developed thinking, problem solving, design, and communication skills. Research in the learning sciences has demonstrated that:

> "Usable knowledge" is not the same as a mere list of disconnected facts. Experts' knowledge is connected and organized around important concepts (e.g., Newton's second law of motion); it is "conditionalized" to specify the contexts to in which it is applicable and it supports understanding and transfer (to other contexts) rather than only the ability to remember. (Bransford, Brown, & Cocking, 1999, p. 9)

Learning that supports transfer involves organizing facts around general principles and understanding their reach, understanding why things happen as they do, drawing explicit connections among ideas, evaluating ideas in ways that draw distinctions as well as identifying commonalities, having multiple opportunities to apply learning in deliberate practice under increasingly complex conditions, and receiving feedback around both thinking and performance that helps students develop metacognitive abilities (self-regulated planning, learning and problem solving strategies, and reflection) that can drive further independent learning.

If assessments are to support transferable learning that allows students to later use knowledge and skills in generative ways, they need seek contextualized demonstrations of abilities and applications of knowledge to complex problems. To enable the development of these skills, both curriculum and assessment should emphasize a deep understanding of the central concepts and modes of inquiry within and across disciplines.

These disciplinary modes of inquiry are the means by which experts in the field construct and evaluate knowledge. For example, scientists use rigorous standards of evidence organized through scientific investigation; mathematicians engage in particular forms of mathematical reasoning and problem solving; historians assemble, weigh, and balance evidence in the course of historical analysis and inquiry; writers and literary analysts engage in forms of expression, interpretation, and critique that support communication and understanding. Applying knowledge in these ways supports more transferable learning.

Second, careful curriculum provides guidance that, like a roadmap, that can allow teachers to choose a number of paths to get to their destination. We argue that the curriculum and assessments should be organized around a well-defined set of learning progressions within a subject area, which can assist teachers in determining where students are along a learning continuum and guide their instruction, as well as formative and summative assessments, to support students to attain the college and career readiness standards.

Learning progressions are empirically-validated descriptions of how learning typically unfolds within a curricular domain or area of knowledge and skill. An understanding of learning progressions is important for teachers, so that they can scaffold instruction, identify gaps in students' understandings, target instruction, and follow up with appropriate assessments. The *progress map* in Figure 1 is an example of a learning progression from Australia's Developmental Assessment program. A student's progress in understanding number concepts can be charted on this continuum, which provides a picture of individual growth against a backdrop of normatively established expectations (Shepard, Hamerness, Darling-Hammond, & Rust, 2005). Similarly, the progression for developing early spelling shown in Table 1 provides an illustration of a normal sequence that can guide a teacher's planning, instruction, and assessment.

A sense of learning progressions is equally important for guiding both formative assessment, so that useful information is obtained about student learning progress and next steps, and summative assessment, so that growth can be more successfully evaluated. For this purpose, a more complete map leading to high levels of competence around central dimensions of learning in the subject matter domain is needed. This map is displayed at larger grain size than the classroom level examples offered here and may be supported by a set of finer grained progressions.

An example of a learning sequence shaping mathematics curriculum expectations and assessment in Victoria, Australia is included in the appendix. Organized around the six major dimensions of mathematics learning as defined in the curriculum—number; space; measurement, chance, and data; structure; and working mathematically—the standards show how students are expected to progress across six levels of growing competence during the first 10 grade levels of school.

Such learning progressions and sequences are not well-represented in most contemporary assessment programs in the United States. As assessment expert Lorrie Shepard has noted:

> One of the obstacles to the development of instructionally useful learning progressions has been the patchwork fashion in which large-scale assessment systems have been developed over time. State and national assessments, originally intended to monitor large-scale trends, focused on grade level expectations for milestone grades (e.g., 4, 8, 12). More recently, with increased requirements for individual testing, states have filled in the intervening grades and interpolated curricular expectations. However, these expectations, especially when "world class standards" are set on never-before-implemented curricula, do not necessarily reflect the developmental trajectory of real students. (Shepard et al., 2005, p. 284)

This task lies ahead of us in developing useful standards and assessments in the United States. While assessment development cannot wait until every aspect of learning sequences is worked out (there are years of research ahead to accomplish this), our proposal would draw on the work that has been done by researchers in the United States and abroad to create a map of learning progressions in the core areas of reading and mathematics where the evidence is strong. An ongoing aspect of refining the assessments over time would be based on refining our understanding of student learning processes and how those are represented in curriculum work.

Below is the lower portion of a counting and ordering progress map. The map shows examples of knowledge, skills, and understandings in the sequence in which they are generally expected to develop from grades one through five. This type of map is useful for tracking the progress of an individual child over time. An evaluation using tasks designed to tap specific performances on the map can provide a "snapshot" showing where a student is located on the map, and a series of such evaluations is useful for assessing a student's progress over the course of several years.

1 Counts collections of objects to answer the question 'How many are there?'
Makes or draws collections of a given size
(responds correctly to Give me 6 bears)
Makes sensible estimates of the size of small collections up to 10
(for 7 buttons, 2 or 15 would not be a sensible estimate, but 5 would be)
Skip counts in 2s or 3s using a number line, hundred chart, or mental counting (2, 4, 6, …)
Uses numbers to decide which is bigger, smaller, same size
(If he has 7 mice at home and I have 5, then he has more)
Uses the terms first, second, third (I finished my lunch second)

2 Counts forward and backward from any whole number, including skip counting in 2s, 3s, and 10s
Uses place value to distinguish and order whole numbers
(writes four ten dollar notes and three one dollar coins as $43)
Estimates the size of a collection (up to about 20)
Uses fractional language (one-half, third, quarter, fifth, tenth) appropriately in describing and comparing things
Shows and compares unit fractions (finds a third of a cup of sugar)
Describes and records simple fractional equivalents
(The left over half pizza was as much as two quarters put together)

3 Counts in common fractional amounts
(two and one-third, two and two-thirds, three, three and one-third)
Uses decimal notation to two places
(uses 1.25 m for 1 m 25 cm; $3.05 for three $1 coins and one 5 cent coin; 1.75 kg for 1750 kg)
Regroups money to fewest possible notes and coins
(11 x $5 + 17 x $2 + 8 x $1 regrouped as 1 x $50 + 2 x $20 + $5 + $2)
Uses materials and diagrams to represent fractional amounts
(folds tape into five equal parts, shades 3 parts to show 3/5)
Expresses generalizations about fractional numbers symbolically
(1 quarter = 2 eighths and 1/4 = 2/8)

4 Counts in decimal fraction amounts (0.3, 0.6, 0.9, 1.2, …)
Compares and orders decimal fractions (orders given weight data for babies to two decimal places)
Uses place value to explain the order of decimal fractions
(which library book comes first-65.6 or 65.126?why?)
Reads scales calibrated in multiples of ten
(reads 3.97 on a tape measure marked in hundredths, labeled in tenths)
Uses the symbols =, <, and > to order numbers and make comparisons
(6.75 < 6.9; 5 x $6 > 5 x $5.95)
Compares and orders fractions (one-quarter is less than three- eighths)

5 Uses unitary ratios of the form 1 part to X parts
(the ratio of cordial to water was 1 to 4)
Understands that common fractions are used to describe ratios of parts to whole
(2 in 5 students ride to school. In school of 550, 220 ride bikes)
Uses percentages to make straightforward comparisons
(26 balls from 50 tries is 52%; 24 from 40 tries is 60%, so that is better)
Uses common equivalences between decimals, fractions, and percentages
(one-third off is better than 30% discount)
Uses whole number powers and square roots in describing things
(finds length of side of square of area 225 sq cm as a square root of 225)

SOURCE: Adapted from Masters and Forster (1996, p. 2). *Knowing What Students Know.* Reprinted with permission in Shepard (2005). Assessment. In L. Darling-Hammond and J. Bransford (Eds.), 2005, *Preparing Teachers for a Changing World*, San Francisco, CA: Jossey-Bass.

Figure 1. Progress Map for Counting and Ordering.

Table 1. Strategies in Children's Spelling

Prominent Strategy	Description	Example
Prephonemic	Letters are used to write words but the sound-symbol relationships are unrelated to target word.	"'C" for "hat"
Early phonemic	Some phonemes are represented by letters, typically most salient phoneme(s) in a word.	"DR" for "Dear"
Phonetic	Attempts are made to represent most sounds in words, often with a letter name that most closely resembles sound.	"wns" for "once"
Simple Associations	Simple vowels and consonants are represented correctly but complex patterns are not.	"bid" for "bird"
Strategic Extensions	With complex vowels and consonants, attempts reflect complex English patterns, although not the conventions of English.	"bote" for "boat"
Conventional		

Theory of Action

In sum, our theory of action is that an integrated system of curriculum and assessment (both formative and summative) that provides what Lauren Resnick has called "tests worth teaching to" (Resnick,1987) will support higher-quality, more coherent instruction. We believe that assessments that evaluate student work and reasoning will support more transferable learning and teaching, and provide more information to teachers and students. Furthermore, teacher involvement in developing, scoring, and using the results of assessments will support teacher understanding of curriculum and standards and will help improve instruction.

In line with this theory of action, our proposed system includes:

- *Summative tests* that assess student progress and mastery of core concepts and critical transferable skills using a range of formats: selected-response and constructed-response items, and performance tasks, designed together to assess the full range of standards.
- *Formative assessment tools* and supports, shaped around curriculum guidance that includes learning progressions.

- Focused *professional development* around curriculum and lesson development as well as scoring and examination of student work

- *Reporting systems* that provide first-hand evidence of student performance (beyond scores), as well as aggregated scores by dimensions of learning, types of students, schools, and districts.

Governmental Roles

In order to accomplish this, we imagine a systemic approach to transforming assessment of learning in the United States. In this system,

The federal government would:

- Maintain and continue to refine NAEP, using the new blueprints already established, to reflect the standards and more intellectually ambitious assessments of knowledge and skills

- Support research on the design, outcomes, and consequences of curriculum and assessments

- Allow, encourage, and fund the use of performance assessments for state assessment systems under ESEA, as well as the use of open-ended diagnostic assessments that can evaluate student performance over time.

- Support and fund initiatives to infuse knowledge of assessment and learning into pre- and in-service professional development.

States—working within consortia—would:

- Create Common Core Standards—mapped across the grade spans in a set of learning progressions around key dimensions of learning—to serve as the basis for state curriculum and assessment efforts.

- Adopt and augment the standards as appropriate to their context.

- Create and deploy a curriculum framework that addresses the standards—drawing on exemplars and tested curriculum models.

- Build and manage an assessment system that includes both on-demand and curriculum-embedded assessments that evaluate the full range of standards and allow evaluation of student progress. Consortia of states might create joint assessments and an assessment bank of performance tasks linked to the standards that can be used as part of both on-demand tests and curriculum-embedded assessments. These would be accompanied by rubrics that embody the standards, and clear examples of good work, benchmarked to performance standards.

- Create an oversight/moderation/audit system for ensuring the comparability of locally managed and scored assessment components.
- Ensure that teacher and leader education and development infuse knowledge of learning, curriculum, and assessment.
- Implement high-quality professional learning focused on examination of student work, curriculum and assessment development, and moderated scoring.

Districts and schools—perhaps also working in networks or consortia—would:

- Examine the standards and evaluate current curriculum, assessment, and instructional practice in light of the standards.
- Evaluate state curriculum guidance, and further develop and adapt curriculum to support local student learning, select and augment curriculum materials, and continually evaluate and revise curriculum in light of student learning outcomes.
- Design, select, and incorporate formative assessments into the curriculum, organized around the standards, curriculum, and learning progressions, to inform teaching and student learning.
- Participate in administering and scoring relevant portions of the on-demand and curriculum-embedded components of the assessment system, and examining student work and outcomes.
- Help design and engage in professional development around learning, teaching, curriculum, and assessment.
- Engage in review and moderation processes to examine assessments and student work, within and beyond the school.

How the Assessment System Would Operate

Drawing from successful practices in the U.S. and abroad, a consortium of states creating the proposed assessment system would:

- *Develop curriculum frameworks.* When the Common Core Standards have been released, vetted, and adopted, consortia of states would work with curriculum and assessment experts to develop (or adapt from previously successful work) curriculum frameworks mapped to the standards and learning progressions. There has been enormous investment in the United States in high-quality curriculum, for example through NSF and other organizations at the national level, and in many states and districts. Other English-speaking nations have also developed high-quality curriculum materials linked to standards and learning progressions that should be evaluated in this process. This effort would inventory and cull from efforts with a strong evidence base of success in building out curriculum frameworks around which states can organize deeper curriculum development and assessment development at the state and local level, along with instructional supports and professional development.

- *Create a digital curriculum and assessment library.* The results of this effort should ultimately be made available online in a digital platform that offers materials for curriculum building and, eventually, model syllabi for specific courses linked to the standards, formative and summative assessment tasks and instruments, and materials for training teachers and school leaders in both strategies for teaching specific curriculum concepts/units and assessment development and scoring. Formative assessment tasks linked to specific standards could be accessed from an Assessment Task Bank, like that recently developed in Hong Kong, so that they are available both for large-scale and classroom use. In addition, as described below, an electronic scoring platform would be developed and made available across the states.

- *Develop state and local assessments.* Initially, the state consortium would work to create a common reference examination, which includes selected-response, constructed-response and performance components aimed at higher-order skills, linked to the Common Core Standards for Grades 3-8, like the NECAP assessment recently developed by a set of New England states. This assessment would be designed to incorporate more analytic selected-response and open-ended items than many tests currently include and would include a small number of strategically selected curriculum-embedded performance assessments at the classroom level that are part of the summative assessment, while also providing formative information.

These curriculum-embedded components would be developed around core concepts or major skills that are particularly salient in evaluating students' progress in English language arts and mathematics. Exemplars to evaluate and build upon are already available in many states and in nations like England that have developed a set of "tests and tasks" for use in classrooms that help teachers evaluate students' learning in relation to well-described learning progressions in reading, writing, mathematics, and other subjects.

Curriculum-embedded components would link to the skills evaluated in the "on-demand" test, allowing for more ambitious tasks that take more time and require more student effort than can be allocated in a 2- or 3-hour test on a single day; these components would evaluate skills in ways that expect more student-initiated planning, management of information and ideas, interaction with other materials and people, and production of more extended responses that reveal additional abilities of students. (These might include oral presentations, exhibitions, or product development, as well as written responses.)

In the context of summative assessments, curriculum-embedded tasks would be standardized, scored in moderated fashion, and scores would be aggregated up to count as part of the external assessment. Curriculum-embedded assessments would also include marker tasks that are designed to be used formatively to check for essential understandings and to give teachers useful information and feedback as part of ongoing instruction. Thoughtful curriculum guidance would outline the scaffolding and formative assessment needed to prepare students to succeed on the summative assessments.

A design much like this one was developed by the New Standards project in the 1990s, and has been implemented in states like Vermont, Kentucky, and Maine that have tied a set of performance tasks to a reference examination in English language arts and mathematics.

All components of the system would incorporate *principles of universal design* that seek to remove construct-irrelevant aspects of tasks that could increase barriers for non-native English speakers and students with other specific learning needs. In addition, designers who are skilled at developing linguistically supportive assessments and tests for students with learning disabilities would be engaged from the beginning in considering how to develop the assessments for maximum access, as well as how to design appropriate accommodations and modifications to enable as many students as possible to be validly assessed within the system.

The emphasis on evaluating *student growth over time* and on tying standards to a conception of learning progressions should encourage a growth oriented frame for both the on-demand examination and the more extended classroom assessments. Ideally, the reference exam would incorporate computer-based adaptive testing that creates vertically scaled assessments based on the full range of learning progressions in ELA and math. This would allow students to be evaluated in ways that give more accurate information about their abilities and their growth over time. This approach should not preclude evaluation of grade-level standards, which could be part of any students' assessment, nor should it preclude a significant number of constructed-response, open-ended items, as the technology for machine-scoring structured open-ended items is now fairly well-developed. As described later, strategic use of partial teacher scoring for these items would also be a desirable element of the system to support teachers' understanding of the standards and assessments, and their planning for instruction.

The emphasis on evaluating student growth should also inform the development of the curriculum-embedded elements of the system, which should be selected or developed to strategically evaluate students' progress along the learning continuum. Centrally developed tasks administered and scored by teachers with moderation (see below), using common rubrics, would be part of the set of reported examination scores. (Some of these tasks may ultimately be scored using artificial intelligence technologies as well as teacher scoring.) Existing tools such as the Developmental Reading Assessment and the Primary Learning Record, which evaluate student progress along a learning continuum in ways that can inform both instruction and reporting, should be examined as well for their contribution to the classroom-embedded component of the assessment system.

In sophisticated state systems, it may be possible to begin to incorporate information about student learning that teachers develop from their own classroom evidence, linked to the standards and learning progressions and guided by the curriculum frameworks. This is the primary approach to assessment before high school in countries like Finland, England, New Zealand, and Australia. This approach is likely to be most productive of more sophisticated and adaptive teaching and well-supported student learning. This could be an optional aspect of the Consortium's work for states and communities with interest and capacity.

At the *high-school level*, the Consortium would explore several options for assessment:

- *Course- or syllabus-based systems* like those in England, Australia, Singapore, Hong Kong, Alberta (Canada), as well as the International Baccalaureate. Generally conceptualized as end-of-course-exams in this country, this approach should become a more comprehensive course assessment approach like that pursued in these other countries. Such an approach would include within-course performance assessments that count toward the examination score, as well as high-quality assessment end-of-course components that feature constructed-response as well as selected response items. Within-course performance assessments would tap central modes of inquiry in the disciplines, ensuring that students have the opportunity to engage in scientific investigations; literary analyses and other genres of writing, speaking, and listening; mathematical modeling and applications; and social scientific research. Such an approach might require an ELA and math assessment at a key juncture that evaluates an appropriate benchmark level for high-school standards, and then, as in high-achieving nations, allow for pursuit of other courses/assessments that are selected by students according to their interests and expertise. These could serve as additional information on the diploma for colleges and employers.

- *Standards-driven systems* that might include a more comprehensive benchmark assessment in ELA and mathematics, complemented by collections of evidence that demonstrate students' abilities to meet certain standards within and across the disciplines. This set of assessments would allow more curriculum flexibility in how to meet the standards. Systems like these are used in some provinces in Canada and Australia, in states like Rhode Island, Wyoming, Nebraska, and New Hampshire, and in school organizations like Envision Public Schools, New Tech High, Asia Society schools, and the New York Performance Standards Consortium. Sometimes these sets of evidence are organized into structured portfolios, such as the Technology portfolio in New Hampshire and the broader Graduation portfolios in these sets of schools that require specific tasks in each content area, scored with common rubrics and moderation.

- *A mixed model* could combine elements of both course- and standards-driven models, allowing some demonstrations of proficiency to occur in any one of a range of courses (rather than a single, predetermined course) or even outside the bounds of a course, like the efforts by some states to allow students to pass courses via demonstrations of competence rather than seat time (e.g., NH, OH). Such a system could also include specific components intended to develop and display research and inquiry skills that might also be interdisciplinary, such as the Project Work requirements in England, Singapore, and the International Baccalaurate, and the Senior Project requirements in Pennsylvania and Ohio.

- *Develop moderation and auditing systems for teacher-scored work.* State consortia would develop protocols for managing moderation and auditing systems and training scorers so as to enable comparable, consistent scoring of performance assessments. In other nations' and states' systems that include these features routinely, procedures have been developed to ensure both widespread teacher involvement—often as part of professional development time—and to create common standards and high levels of reliability in evaluating student work. A range of models are possible, and the consortium would serve as a resource to individual states in developing and implementing strong, efficient approaches.

- *Provide time and training for teachers and school leaders.* To implement an integrated system of curriculum, assessment, and instruction, time must be set aside for teacher development and participation in the system. Creative use of existing professional development days and incentives provided by recertification requirements (e.g., continuing education units) can be part of this commitment. In order to secure benefits for the quality of teaching and learning, states will need to designate concrete commitments to support teacher engagement in curriculum and assessment development, scoring, and analysis.

- *Develop technology to support the system.* Technology should be used to enhance these assessments in a number of ways: by delivering the assessments; in online tasks of higher-order abilities, allowing students to search for information or manipulate variables and tracking information about the students' problem-solving processes; in some cases, scoring the results or delivering the responses to trained scorers / teachers to assess from an electronic platform. Such a platform may also support training and calibration of scorers and moderation of scores, as well as efficient aggregation of results in ways that support reporting and research about the responses. This use of technology is already being used in the International Baccalaureate Programme assessment system, which includes both on-demand and classroom-based components.

- In order to gain the efficiency and cost benefits of machine scoring and the teaching and learning benefits of teachers' moderated scoring, a mixed system would be developed where computer-based scoring is incorporated on constructed-response tasks where useful—though teachers would score some of these tasks for anchoring and learning purposes—while other tasks that require human scoring engage most teachers in scoring to support improvements in instruction.

In Sum

We believe it is critical that the system be constructed with the entire teaching and learning system in mind, rather than simply shoving another set of tests into systems that are already fragmented by efforts to force-fit parts that were never designed to create an integrated system. Decisions should be made as much on the basis of what schools and teachers need in order to support instruction as on the basis of how to derive scores that can be published and compared. Both are important, of course. But at the end of the day, high levels of equitable learning are the real goal, and this goal will only be achieved if educators are able to teach more knowledgeably and effectively.

Part 2: Answers to Guiding Questions

Rigorous Standards and Good Instructional Practices

How Does Your Model Ensure That the Assessments Measure Achievement of Standards That Are Based on College- And Career-Readiness?

The Common Core standards are intended to outline the foundational knowledge and skills that will prepare students for college- and career-readiness in a thoughtful sequence. Our model will further illuminate the learning progressions that lead to the college- and career-ready standards within central domains. Assessment items and tasks will be designed to convey what's important to learn as well as providing an opportunity to examine students' understanding. As Lorrie Shepard has suggested, the assessments should *embody t*he range and depth of what we want students to understand and be able to do, not just serve as remote proxies.

In our proposal, the nature of the items and tasks is a key aspect of benchmarking to college- and career-readiness. Curriculum guidance and assessments should seek to teach and evaluate knowledge and skills that generalize to higher education and work settings. Assessments that evaluate student performance, analysis, reasoning, and critical thinking skills will support more transferable learning and teaching. Feedback from these types of assessments will provide students, parents, and teachers with evidence-based input to determine where they are in terms of ultimately being prepared to achieve college and career readiness standards. Analysis of student performance on the assessments will help teachers determine students' gaps in knowledge and skills enable them to specifically adjust instruction to strategically support students' learning.

What Types or Items, Tasks, and Tools Does Your Model Include and How Will They Be Scored?

A coherent, comprehensive assessment system consists of a set of strategically selected measures to assess students' knowledge and mastery of college and career readiness standards and their progress en route to these goalposts. These measures should be selected based on their appropriateness for the construct to be assessed, and they should fit together as a set to evaluate the full range of standards. Our assessment system will include analytic selected-response items, short and extended constructed-response items, and standardized performance tasks in each grade level tested. Depending on the grade level, embedded performance tasks per grade or course would range from 1 to 4 tasks of varying length. Student performance on the on-demand examination is intended to be combined with the embedded performance measures to contribute to a total score on the grade specific accountability measure.

This approach to assessment that combines on-demand (sit-down test) measures with curriculum-embedded performance tasks is common in high-achieving countries, as indicated by Table 2, below. The curriculum-embedded components, which generally involve standardized tasks and scoring processes, count for between 20% and 100% of the total examination score. In our proposed system, we expect the consortium of states to evaluate the relative weights of on-demand items and curriculum-embedded performance tasks. We anticipate that these tasks could comprise from 20-50% of the total score, depending on the subject and grade level, and the judgments of the states.

Analytic selected-response items that are judiciously designed to evaluate students' knowledge in a manner that goes beyond simple recall of facts to an evaluation of evidence, patterns, or conditions. Students may be asked to recognize patterns or discrepancies and infer potential interpretations or causes; identify underlying problems or conflicts and select a solution strategy to resolve the conflict; or examine a data set and interpret the results presented and the significance of the pattern of results. Students must do more than apply factual knowledge; they must analyze the situation, consider multiple options or evidence, and make a judgment.

Table 2. Assessment Components in Selected Countries

	End-of-Course Exam	School-Based Assessment
Alberta, Canada (4 Core Academic Subjects)	25% MC 25% Open-Ended	50% Coursework
England (Choice of Subjects)	40-75% Open-Ended (Written and Extended Problems)	25-60% Performance Tasks
Singapore (Choice of Subjects)	70-80% Open-Ended (Written, Oral, Problems)	20-30% Performance Tasks + Project Work
Victoria, Australia (Choice of Subjects)	Up to 50%—Combination of MC and Open-Ended	50% or more Coursework & Performance Tasks
Queensland, AU (Choice of Subjects)	No end-of-course component	100% Performance Tasks
Hong Kong (Choice of Subjects)	70-80%—Combination of mostly open-ended and some MC	20-30% Performance tasks with plans to increase
International Baccalaureate (Choice within Core Curriculum)	Generally 50%—Combination of MC and Open-Ended	50% Coursework and Performance Tasks

Below we contrast a traditional item measuring basic recall (from a U.S. History test) with an analytic item developed by Alberta, Canada history teachers as part of Alberta's diploma examination—both evaluating knowledge of the same period of history. The latter requires deeper historical content knowledge as well as the ability to compare and contrast situations across historical periods and contexts.

> Who was president of the United States at the beginning of the Korean War?
> a) John F. Kennedy
> b) Franklin D. Roosevelt
> c) Dwight Eisenhower
> d) Harry Truman
> e) Don't know

> A feature common to the Korean War and the Vietnam War was that in both conflicts:
> a) Soviet soldiers and equipment were tested against American soldiers and equipment.
> b) The United States became militarily involved because of a foreign policy of containment.
> c) The final result was a stalemate; neither side gained or lost significant territory.
> d) Communist forces successfully unified a divided nation.

These kinds of questions require different kinds of curriculum and teaching, with one focusing on memorizing names and dates without context, and the other focusing on deeper analysis of historical events and situations, and their genesis and outcomes.

Constructed-response items that demonstrate thinking and allow students to defend their ideas. Short and extended constructed-response items can measure applied skills as well as content knowledge, and can help reveal students' reasoning to teachers engaged in scoring the responses. The questions can be designed to reveal students' misconceptions as well as solution strategies. These can be flagged for identification in the criterion-based scoring rubric.

Constructed-response items can provide students with various types of stimuli (e.g., timelines, maps, graphs, pictures or other visuals , charts, short readings, problem statements, etc), Students' responses demonstrate use of complex thinking skills such as formulating comparisons or contrasts; proposing cause and effects; identifying patterns or conflicting points of view; categorizing, summarizing, or interpreting information; and developing generalizations, explanation, or evidence based conclusions.

Items that evaluate students' reasoning and depth of understanding may often have several components applying various problem solving skills that probe the depth of understanding. For example, a constructed-response task from an Alberta, Canada applied mathematics assessment requires students to think about the concepts of normal distribution, mean, and standard deviation in increasingly complex ways, and to explain their thinking as they proceed through the several parts of the problem.

Alberta, Canada High-School Applied Mathematics Problem

Six students from a particular class are travelling by airplane to Europe. They are each allowed to take two suitcases, and no suitcase should have a mass over 32 kg. Any suitcase with a larger mass is classified as overweight, and an extra charge is applied.

At check-in, the mass of each suitcase is determined. The table below lists the mass of each students' suitcase. [Table follows.]

2. a. • Calculate, to the nearest tenth of a kilogram, the mean and standard deviation of the masses of all 12 suitcases.

Mean = ________ kg Standard deviation = ________ kg

The airline knows that suitcase masses model a normal distribution. Using your answers from the previous bullet, **determine** the percentage, to the nearest whole number, of all suitcases that the airline can estimate will be over 32 kg. ________

Use the following additional information to answer the next part of the question.

The students do not want to leave anything behind or pay extra charges for overweight suitcases, so they decide to repack some of the suitcases. They decide that those students with overweight suitcases will first move items from one of their suitcases to the other. If a student still has an overweight suitcase, she must then move some items into another student's suitcase.

b. Complete the following table to show how the students could rearrange their items in the suitcases so that there are no overweight suitcases.

c. Explain how repacking the suitcases may affect the mean and standard deviation of the masses of all 12 suitcases. **Justify** your answer.

Items may draw specifically on texts read and work done throughout the year, as for example, this item from the International Baccalaureate Programme examination.

International Baccalaureate Programme Constructed-Response English Item

On the English A1 exam in the International Baccalaureate Programme, students may choose from essay questions like the following, which treat different genres of literature based on texts they have studied during the year:

1. Using two or three of the works you have studied, discuss how and to what effect writers have used exaggeration as a literary device.
2. Acquiring material wealth or rejecting its attractions has often been the base upon which writers have developed interesting plots. Compare the ways the writers of two or three works you have studied have developed such motivations.
3. Discuss and compare the role of the speaker or persona in poems you have studied. You must refer closely to the work of two or three poets in your study and base your answer on a total of three or four poems. (International Baccalaureate Organization, 2005).

Curriculum-embedded performance tasks that require problem framing, as well as problem solving and inquiry. Curriculum-embedded performance tasks focus on the application of knowledge, skills, and analytical thinking specific to the discipline.

In English language arts, quite often these tasks are organized by genre of writing and literary analysis. For example, the Kentucky writing portfolio, much like the collection of performance tasks in the British GCSE system, requires particular kinds of writing tasks written to specifications and scored with a task-based rubric that reflects the demands of the task. In 2009, the set of performance tasks in Kentucky included one in each of the following categories, written to particular specifications:

- Reflective Writing
- Personal Expressive Writing/Literary Writing
- Transactive Writing
- Transactive Writing with an Analytical or Technical Focus (Grade 12 only).

Evidence shows that, unlike the problems in scoring non-standardized portfolios, these collections of specific tasks can be scored consistently with high levels of reliability (Commonwealth of Kentucky, 2009).

Performance tasks often probe an understanding of particular content or concepts, as well as the ability to analyze and organize information; produce a product, analysis, or model; and explain and defend ideas. For example in a physics course, students studying Newton's Laws may be asked to research automobiles on the market and to discuss what it means to be energy efficient. Using what they have learned about automobiles and the concepts of acceleration, motion, force, and friction, students could design, build, and test the efficiency of a model vehicle.

In the example below, from the Ohio Performance Assessment Project, students apply their understanding of a central theme in American literature to a task that requires selecting, analyzing, interpreting, and explaining texts.

Ohio Performance Assessment Project English Language Arts Performance Task

Imagine that you are editing an online digital anthology for 11-12th graders entitled, "Perspectives on the American Dream." Your job is to prepare the introduction to this anthology. In your introduction, please do the following things:

a) Decide which texts you want to include and in which order (you must include at least six texts). Texts can include books, poems, songs, short stories, essays, photographs, articles, films, television shows, or Internet media. The six texts must represent at least two different perspectives and must include at least two different types of text (e.g., print text, visual media, audio media, multi-media, digital media). At least two of the six texts must be print (written) texts.
b) Identify and discuss different perspectives on the American dream represented in the six texts you selected.
c) Write a short paragraph about each text, in which you make clear why you have included it and how it relates to the other texts in your anthology.
d) Propose a set of questions to focus readers as they consider the perspectives represented in these texts.

In a mathematics example from the Ohio Performance Assessment Project, students are asked to evaluate how heating costs may change as a simultaneous function of temperature, fuel costs, and savings due to insulation. The task requires students to apply their knowledge of ratio, proportion, and algebraic functions to a complex, real-world problem. They must engage in analysis and modeling of multiple variables. The response requires a display, explanation, and defense of their ideas.

Gas Bills, Heating Degree Days, and Energy Efficiency

Here is a typical story about an Ohio family concerned with saving money and energy by better insulating their house.

Kevin and Shana Johnson's mother was surprised by some very high gas heating bills during the winter months of 2007. To improve the energy efficiency of her house, Ms. Johnson found a contractor who installed new insulation and sealed some of her windows. He charged her $600 for this work and told her he was pretty sure that her gas bills would go down by "at least 10 percent each year." Since she had spent nearly $1,500 to keep her house warm the previous winter, she expected her investment would conserve enough energy to save at least $150 each winter (10% of $1,500) on her gas bills.

Ms. Johnson's gas bill in January 2007 was $240. When she got the bill for January 2008, she was stunned that the new bill was $235. If the new insulation was going to save only $5 each month, it was going to take a very long time to earn back the $600 she had spent. So she called the insulation contractor to see if he had an explanation for what might have gone wrong. The contractor pointed out that the month of January had been very cold this year *and* that the rates had gone up from last year. He said her bill was probably at least 10% less than it would have been without the new insulation and window sealing.

Argon Energy Co.	**Customer** Ms. Arlene Johnson 42 Bluebonnet Avenue Columbus OH 43205	**Bill Date** Jan 31, 2008 **Account #** 56-73342B Residential
Current Itemized Bill		
December 30 reading actual		8300
January 31 reading actual		8488
Total units used January 2008		188
January 2008:	1108 heating degree days 0 cooling degree days	
Price per unit @ $1.25		$235
Energy Use History		
Total units used January 2007		200
January 2007:	1000 heating degree days 0 cooling degree days	
TOTAL CURRENT CHARGES		**$235**

Ms. Johnson compared her January bill from 2008 to her January bill from 2007. She found out that she had used 200 units of heat in January of 2007 and was charged $1.20 per unit (total = $240). In 2008, she had used 188 units of heat but was charged $1.25 per unit (total = $235) because gas prices were higher in 2008. She found out the average temperature in Ohio in January 2007 had been 32.9 degrees, and in January of 2008, the average temperature was more than 4 degrees colder, 28.7 degrees. Ms. Johnson realized she was doing well to have used less energy (188 units versus 200 units), especially in a month when it had been colder than the previous year.

Since she used gas for heating only, Ms. Johnson wanted a better estimate of the savings due to the additional insulation and window sealing. She asked Kevin and Shana to look into whether the "heating degree days" listed on the bill might provide some insight.

Ohio Performance Assessment Project "Heating Degrees" Task

On the basis of the situation painted above and some initial information to help them begin to research "heating degree days" on the internet, students are given two tasks:

1. Assess the cost-effectiveness of Kevin and Shana's mom's new insulation and window sealing. In their assessment, you must do the following:
   - Compare Ms. Johnson's gas bills from January 2007 and January 2008.
   - Explain Ms. Johnson's savings after the insulation and sealing.
   - Identify circumstances under which Ms. Johnson's January 2008 gas bill would have been at least 10% less than her January 2007 bill.
   - Decide if the insulation and sealing work on Ms Johnson's house was cost-effective and provide evidence for this decision.
2. Create a short pamphlet for gas company customers to guide them in making decisions about increasing the energy efficiency of their homes. The pamphlet must do the following:
   - List the quantities that customers need to consider in assessing the cost-effectiveness of energy efficiency measures.
   - Generalize the method of comparison used for Ms. Johnson's gas bills with a set of formulas, and provide an explanation of the formulas.
   - Explain to gas customers how to weigh the cost of energy efficiency measures with savings on their gas bills.

Note that the examples of items and tasks offered above would build on a common curriculum framework as outlined in the Common Core Standards, but would not require that every state or district have adopted the same curriculum approach or materials.

Student-designed project work requires the student to design and implement a project plan, engage in research and problem solving strategies as appropriate, write up results, and, often, present and defend them orally to evaluators. Students develop and use meta-cognitive abilities as they manage a process of idea development, construction, evaluation, self-reflection, and revision. This kind of work is increasingly included in other nation's assessment systems. For example, in Queensland (Australia), Great Britain, and Singapore, all students complete science investigations as a component of the examination score in science classes (usually counting from 20 to 30% of the total score). This requires them to plan an investigation to answer a particular question; make and record observations, measurements, methods and techniques; interpret and evaluate observations and experimental data; and write up and critique their results and methods.

In addition, college-bound students in England, Singapore, and in International Baccalaureate Programme courses across 125 countries must complete a major Project Work assignment. For similar reasons, Ohio, Pennsylvania, and Rhode Island are among states in the United States that have instituted a senior project or related requirement as part of their graduation requirements. These jurisdictions believe that this kind of activity is essential to become truly college and career-ready, as it emulates the kinds of work students will need to do in these settings.

We believe that the experience of defining and pursuing a problem, collecting and evaluating evidence and information, and developing a product or solution with written explanation is important and should be part of a collection of evidence students can offer their potential higher education institutions and employers. In our proposed system, states will be supported to develop project work elements of their assessment systems, especially, although not exclusively, at the secondary level. Because of their newness in the context of U.S. assessment systems, these tasks would be used initially to inform local and state-level decisions, rather than factoring into comparative summative judgments across jurisdictions. Over time, as the capacity to implement and evaluate this kind of work increases, such work might be integrated into the system as a whole.

Project work in places like Singapore is directly linked to the idea of career readiness:

> In carrying out the Project Work (PW) assessment task, students are intended to acquire self-directed inquiry skills as they propose their own topic, plan their timelines, allocate individual areas of work, interact with teammates of different abilities and personalities, gather and evaluate primary and secondary research material. These PW processes reflect life skills and competencies such as knowledge application, collaboration, communication and independent learning, which prepare students for the future workplace. (Singapore Examinations and Assessment Board, 2009)

Project Work in Singapore

Project Work (PW) is an interdisciplinary subject that is compulsory for all pre-university students. As an interdisciplinary subject, it breaks away from the compartmentalization of knowledge and skills to focus on interdisciplinary outcomes by requiring students to draw knowledge and apply skills from across different subject domains. The goals for this experience are embedded in the requirements for the task and its assessment, which are centrally set by the Singapore Examinations and Assessment Board. The tasks are designed to be sufficiently broad to allow students to carry out a project that they are interested in while meeting the task requirements:

- It must foster collaborative learning through group work. Together as a group which is randomly formed by the teacher, students brainstorm and evaluate each others' ideas, agree on the project that the group will undertake and decide on how the work should be allocated amongst themselves.
- Every student must make an oral presentation: Individually and together as a group, each student makes an oral presentation of his / her group project in the presence of an audience
- Both product and process are assessed: There are 3 components for assessment:
  —the Written Report which shows evidence of the group's ability to generate, analyze and evaluate ideas for the project;
  —the Oral Presentation in which each individual group member is assessed on his/her fluency and clarity of speech, awareness of audience as well as response to questions. The group as a whole is also assessed in terms of the effectiveness of the overall presentation;
  —the Group Project File in which each individual group member submits three documents related to 'snaphsots' of the processes involved in carrying out the project. These documents show the individual student's ability to generate, analyze and evaluate
    (i) preliminary ideas for a project
    (ii) a piece of research material gathered for the chosen project and
    (iii) insights and reflections on the project.

Assessment is school-based and criterion-referenced. While task setting, conditions, assessment criteria, achievement standards and marking processes are externally specified by SEAB, the assessment of all three components of PW is carried out by classroom teachers, using a set of assessment criteria provided by the board. All schools are given exemplar material that illustrates the expected marking standards. The Board provides training for assessors and internal moderators. Like all other assessments, the grading is both internally and externally moderated.

Which Components of Your System Are Used For Summative, High-Stakes Purposes?

In our comprehensive assessment system, multiple measures of learning and performance will be used to evaluate students' skills and knowledge for summative purposes. The results of on-demand tests using selected-response and constructed-response items will be combined with the weighted results of reliably-scored curriculum-embedded performance tasks to create consistent measures for gauging student attainment and for comparing results across schools, districts, and states.

The consortium will construct a relatively lean summative assessment that is comparable across states (including a small number of performance tasks at each tested grade), and usable for deriving individual student scores and aggregating scores to the classroom, school, district, and state levels. In addition, individual states will build out more extensive and flexibly used performance components in formative and interim assessments, as well as instructional supports, for their own state and local evaluation and instructional purposes. States will work together to create and share many of these tools, and to link them to the standards and to curriculum materials, but will use them to differing extents and in different ways depending on their needs and systems.

The system will be designed, like that in Massachusetts, to release items and tasks so that they can serve as learning tools for teachers, parents, students, and the public. To the extent that students' actual responses are available to teachers and students, they can also serve formative purposes. Other formative assessments will also be developed and shared by members of the consortium and mapped to the same standards, core concepts, and modes of inquiry, using similar assessment strategies and made available to teachers across states through the online curriculum and assessment library. These will be connected to curriculum frameworks and standards to support their use.

On the issue of stakes, we would argue, as economist Richard Murnane suggested in his study of Vermont's assessment system (Mumane & Levy, 1996), that medium stakes can be preferable to high stakes of the kind that often lead to unintended negative consequences for student participation in school and teachers' instructional practices (for a review of research of the effects of testing on teaching and student participation in school, see Darling-Hammond & Rustique-Forrester, 2005). That is, the use of rich assessments to inform stakeholders about educational performance (both because what students know and can do is made visible and because it produces useful, interpretable scores) can produce significant attention to educational improvement and support, as well as needed information for teachers, parents, policymakers, colleges, and employers.

We agree with the cautions of the *Standards for Psychological Testing* (American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education, 1999) that test scores should always be combined with other evidence about student learning and performance when high-stakes decisions are to be made about students. And we note that high-achieving nations with high-school examination systems typically do not use them directly to determine whether students will graduate from high school. Instead, they use scores to inform course grades, to provide information on the transcript that informs colleges and vocational / technical pathways about students' proficiencies in different areas, and to determine which areas of achievement warrant certification. This allows the standards to be higher and the process of combining evidence for decision making to be well-grounded. The same principle should pertain for using assessments to inform judgments about educators and schools. Evidence from tests should always be combined with other evidence about performance as the basis for decision making.

Technology Use and Reporting

How Would Your Design Utilize Technology in Delivery, in Scoring, and to Extend the Range and Complexity of Skills and Processes That Can Be Assessed?

A technology infrastructure will be a key instrument in the administration, scoring, data collection, and score reporting for both on-demand exams and the curriculum embedded performance task components. This technology platform will significantly reduce the financial and human resource burdens of implementing and maintaining the system as required by psychometric standards. In our proposed system, technology will be used to:

- Deliver both on-demand and curriculum-embedded assessments to students and teachers;
- Use adaptive computer technology not only to deliver tests electronically but also to create assessments that are responsive to the test-taker's performance and allow better measurement of growth.
- Deliver online tasks of higher-order abilities, allowing students to engage in online simulations; analyze and draw conclusions from an "inbox" of artifacts that can be explored using online tools; search for information; build and test models; manipulate variables, and create a wide range of responses;
- Score selected items (both selected-response and constructed-response)
- Deliver the responses on other items or tasks to trained scorers/teachers to assess from an electronic platform.
- Support training and calibration of scorers and moderation of scores;
- Enable efficient aggregation of results in ways that support reporting and research about the responses.

We combine our discussion of technology use for assessment, scoring, and reporting because we will build an interactive web-based platform that permits much more efficient and robust collection, sharing, evaluation, and aggregation of evidence about student learning. Digitized student responses will allow states to manage all the relevant balanced assessment data on a common platform which will enable teachers and students to both manage their learning and *significantly accelerate* the speed in which assessment information will be returned to the students and schools to support improvement. Moving to an electronic interface can enable tasks to be tailored to individual students, allow students and teachers to manage and monitor their own learning, train and certify scorers, assign responses to multiple scorers, support calibration and auditing of scorers, generate reports, and create interfaces with other systems.

Moving to a more balanced, multiple-measure reporting system depends, in part, on the development of intelligent technologies that can capture and transform information that goes beyond simple test scores to include both formative and summative student performance data, ranging from simple and complex text to digital media (including exhibitions of student work). Many data management systems that are currently in use provide accessible and relevant demographic and standardized test score data. These systems, however, are not generally structured to provide actionable, just–in-time evidence around academic factors and early warning indicators (formative information) that schools, districts, and states can use to guide curriculum, instruction and assessment. More importantly, these systems don't always put the power of classroom-based data in the hands of teachers and educators to support struggling students and to monitor student progress over time—so called *on-track measures*. Examining patterns of early warning indicators can uncover systemic weaknesses and enable schools and districts to address them in real time. Therefore, ready access to actionable data that is embedded in the school's culture and norms can guide the development of preventive and proactive strategies used to strategically target resources to high-leverage areas of need that will lead to improved student outcomes and school improvement.

Perhaps the most critical factor in the development of a reporting system to support 21st century learning is the development of a system that provides a web- based interface for teachers, students, state department staff, and parents. Below we briefly outline the access, use, and reporting functions of a system already under development.

Reporting in this context is not conceptualized simply as communicating assessment results but as part of a comprehensive learning system that provides actionable information to teachers, students and states that is used both to build capacity as well as to support district or state accountability.

Student Web-Interface

Collection of assessments from students. A web-based system will enable students to retrieve, upload, and store their own work products, including work assigned by the teacher, performance tasks that are part of the formative or summative assessment system, and traditional tests. Additionally, the system will enable students to monitor their own progress and provide access to peers/faculty/administrators to review the work, to guide revisions, where appropriate, and to ensure that the work is complete and ready for teacher review or grading—or upload to a system where the test or task will be externally scored. Given that student work products can vary from test input to word documents to multimedia components (e.g., spreadsheets, computer-generated graphs and other graphics, etc.), the web-based system will need to be able to convert all work products into a common format for use in a web-based scoring environment. Once the student work is completed for each assessment, teachers will need to verify all of their work and to facilitate the uploading of tests or tasks to an inter-operable electronic platform designed to support the scoring of student work. When formally submitted to the electronic platform, teachers and/or scorers will be notified that student work is available to be scored. Note, that the electronic interface to support the management of student work products should be customized and adapted developmentally to provide all students ready access to the system regardless of grade level (k-12). One explicit purpose of providing this capability is to support students to manage their own learning.

Teacher Web-Interface

Management of teacher training. The system will enable states or a testing company to train scorers electronically, to calibrate scorers to pre-scored benchmarks and to judge whether teachers were able to score reliably. Based on a standardized training module and evidence that teachers meet reliability standards, scorers will be certified to score student work as one key component of a balanced assessment system.

Management of teacher scoring. Once student work is uploaded and converted, certified teacher scorers will have access to the online rubric for scoring work products on the specific standards being assessed. Once all scorers register on the electronic platform, scorers will be assigned to specific tasks (including constructed-response items or performance tasks). Scoring can be at the item, task, or test level. The electronic system will have the capability to assign scorers to particular candidates for scoring. For some purposes (especially where formative diagnostic information is a primary goal), teachers might score their own student's work, while for others, they would be assigned to score the work of other students in the same or a different school, in the same or a different state. In some cases, human scorers may score a sample of tasks that are otherwise computer scored, for purposes of both double scoring where needed and providing teachers with the opportunity to understand the standards and assessments as they inform instruction.

To inform scorer judgment, the rubrics will include references to the descriptors for each level of mastery being assessed. As part of the scoring process, scorers will be able to annotate the work product to justify and identify the evidence they used to support their rubric ratings for a specific task(s) as well as to provide relevant feedback to students.

To ensure reliability of scoring, the electronic system will allow candidate work to be double-scored. Failing or near-failing work samples can be automatically assigned for double-scoring. To ensure that scoring is calibrated across a state or consortium of states, the electronic platform will allow scorers to participate in a central audit of assessments through a randomly selected stratified sample. Audited assessments that have large score discrepancies from original scores will be rescored by other trainers as part of a moderation process to ensure consistency. The system will monitor the scoring and can inform individuals or institutions where there is evidence that they have unreliable scores.

Management of classroom practice. Eventually, the system should also allow teachers to manage and monitor classroom work for formative and summative purposes, as well as storing demographic and other test related student data. With assessment development software, it can help teachers develop formative performance tasks that are aligned to college ready standards and score student work at the school-level using common rubrics. The system will enable teachers to share their teacher/student work products with peers and administrators for review and comment within or beyond the district. It will also enable teachers to align student work to content and performance standards and to benchmark performance against learning progressions. The goal is to both support and provide tools to enable teachers to manage their own teaching and learning.

System Management and Reporting

The development of an electronic interface should become a self correcting real time system that provides access to information that can be used both as an early warning data system to monitor curriculum and instruction and learning and to manage, at scale, the state assessment and accountability system. Reports can be generated to monitor student progress on tests and performance tasks, as well as classroom assignments. Reports can:

- Identify teachers who have completed workshops and online training
- Track input and scoring of student work
- Evaluate and produce reports of inter-rater agreement
- Produce reports of agreement with validity sets (e.g., comparisons between assigned scores and expert scored assessments)
- Provide student, class, building and district summaries
- Return scored student work (released items and tasks) to teachers and students to allow further analysis and feedback, and to support instruction

Reporting student results. Through the use of an inter-operable electronic platform scores can be turned around quickly (within days or weeks) for both performance measures and for on demand standardized assessments (including open ended and constructed-responses). Reporting functions can include:

1. *Student summaries:* Reports that summarize student performance on both standardized components and performance measures. These reports can include both a quantitative summary of scores and comments and feedback from scorers based on rubrics and student performance to guide teacher planning and student improvement.

2. *Class/school/ district/ state summaries:* Once scores have been assigned and verified, the data can be used to provide quantitative and descriptive data to all stakeholders. Scores can be broken down and reported by item type or subscales; broken down by standards and/or aligned to learning progressions as well as disaggregated and reported across all relevant NCLB sub-groups. Aggregated or disaggregated comparisons can also be generated across teachers, classrooms, schools, districts and states. These data can not only provide status information but be disaggregated to identify relative strengths and weaknesses of the individual, school, or classroom.

3. *Parent and community summaries:* Because assessment tasks and responses will be digitized, teachers can regularly report student progress to parents regarding what their child knows and is able to do as well as to provide in the report to parents a narrative of student progress and resources that can be used to support their child's learning. Student performance can be explained, as it is in NAEP reports, with examples of benchmarked released items, so that stakeholders understand the kind of learning and performance being discussed. Finally digitized teacher assignments and completed formative assessments can provide transparency to parents about teacher expectations for learning and ready access to their child's actual work samples.

Measuring Growth

This approach builds on two key approaches to measuring growth:

- Evaluating students' movement along a (long) vertically scaled continuum that sets standards for performance characteristics at different points along the scale (the NAEP approach). Use of computer adaptive testing can more efficiently measure students' location and progress along such a continuum by identifying where students are in their development on particular strands and testing more completely their understanding and skill, while also evaluating a subset of items / tasks that are grade-specific. It is key in this conception that assessment not be constructed to a limited conception of *grade level standards*, which fails to capture student growth in areas or on dimensions where their development falls outside a narrowly-conceived grade level band.

- Evaluating movement along well-defined learning progressions assessed through a collection of evidence about students' developing abilities (e.g., the approach used by the *Assessing Pupils' Progress* system in the United Kingdom, the Developmental Reading Assessment (DRA) used to evaluate literacy development in many states and districts, or the Balanced Assessment System in mathematics, developed by the Shell Centre in London, with colleagues in the United States).

The proposed assessment system is designed to combine scores from on-demand (sit-down) tests and more extended performance tasks in a continuous scale. This scale will evaluate student's development along multiple learning dimensions that are keyed to the core domains of knowledge and skill within a discipline (e.g., within reading, dimensions such as comprehension, analysis, interpretation; within writing, dimensions such as organization, clarity, development of ideas, conventions, and so on). These scores can be examined and aggregated by dimension and combined to produce domain-level scores.

As described above, the use of adaptive computer technologies to administer the curriculum embedded performance tasks and the end-of-course on-demand examinations, as well as to score some open-ended items and tasks, is a critical factor that will facilitate the use of student scores for formative purposes and for tracking student progress over time on a common set of learning progressions within a discipline, as well as their achievement on grade-level standards.

It is important to note that, while researchers have made substantial headway on examining these progressions—and some nations, like Australia, New Zealand, and the United Kingdom, have integrated this notion into curriculum and assessment—these progressions should never be interpreted as lock-step expectation or an absolute sequence of prerequisites, as research understandings of these unfolding sequences is not perfect.

Furthermore, children's knowledge acquisition is influenced by individual proclivities and as well teaching and out-of-school experiences, and learners will always differ somewhat in their individual trajectories. Thus, assessments must be designed to evaluate a wide range of abilities along the continuum, understanding that the development of understanding will be uneven both within and across students.

Within this system, individual performance tasks in the same discipline, offered at different points in time, will be designed based on templates or task shells to reflect specific dimensions of learning. The technology for building such items and tasks has evolved significantly over the last two decades (Lane, 2010). This constancy in the set of constructs measured by some tasks will also allow for better measurement of growth over time along key learning progressions for that discipline. The common scoring rubrics used to score these tasks will be designed to be indexed to the learning progressions for that particular discipline.

This approach has been used successfully by the Collegiate Learning Assessment, given at the beginning and end of college, which evaluates the development of a stable set of cognitive skills using common templates and rubrics for task construction and scoring. Furthermore, this complex open-ended performance task is computer-scored (with double scoring by humans and a reliability of about .9 between machine and human scorers), suggesting the prospect that evaluations of growth can be conducted in a reliable and timely manner.

What Types of Questions Related to Student Growth Will Your Design Be Able tTo Address?

The use of multiple assessment formats (high-quality multiple-choice, constructed-response, performance tasks) that sample from each of the key domains of knowledge and skill within a discipline can provide rich information about a student's progress along multiple dimensions. These multiple assessment formats and events provide a means to triangulate evidence about a student's performance on the same dimension of learning across item formats and occasions. When students' scores across the multiple assessment formats are combined, and when these scores are indexed to learning progressions within a discipline, this assessment design could allow for a more comprehensive and balanced picture of student learning at discrete points in time, as well as the ability to track progress along the same learning dimensions over time (year to year, as well as during a single year). As a result, this assessment design could provide more useful formative assessment information that can provide guidance around curriculum and instructional decisions, and sufficient data from across occasions to make summative decisions. A useful example of how this approach can work is available in the *Assessing Pupil Progress* system in use in England which provides assembles evidence around aggregated evidence at key stages.

Because this design can evaluate growth and progress in multiple ways, and because it incorporates means for evaluating growth along a full achievement continuum, it may also be useful in providing multiple indicators of student learning gains for informing judgments about the contributions of teachers and schools to achievement. Based on the findings of researchers about the challenges with value-added measures at the classroom level (Braun, Chudowsky, & Koenig, 2010) this information also needs to be accompanied by, and interpreted with, information the characteristics of students, resources, and school policies.

Accessibility

New systems of assessment must be designed to allow all students to fully participate and demonstrate what they know and are able to do in an environment that embraces students with different instructional, linguistic, and academic needs. Our proposed system will start with the full range of students in mind, using, to the greatest extent possible, principles of universal design in the development of the assessments, maintaining consciousness about the clarity and accessibility of items and tasks throughout the process. In addition, in our design and field testing of both on-demand and curriculum-embedded components, we will aim to identify and test variations in assessment design that are meant to provide access to and more validly measure the content knowledge and skills of special populations.

The goal of the Consortium will be to build upon the on-going work in states that has already been done to provide greater access to learning and assessment opportunities for English learners, students with disabilities, and students from economically disadvantaged communities. New approaches will also be sought to ensure more comparable and productive accommodations for special populations of students. In addition to the need for more effective accommodations for a wide range of special education students, significant headway is needed with respect to accommodations for English language learners (ELL). An expert panel convened in 2007-2008 found that the majority of accommodations used in state testing programs actually came from special education, and do not address the linguistic needs of ELLs (Rivera, 2008).

To help guide the development of new and more effective strategies to provide all students opportunities to learn and perform on complex performance items/tasks that are aligned to college and workplace standards—we will put in place a blue ribbon panel, chaired by Kenji Hakuta and Guadalupe Valdez at Stanford University, that will help develop and oversee new directions for the development of tools, strategies and specific accommodations/modifications that will provide greater access to English learners, students with disabilities, and others from culturally and economically diverse families.

To develop accessible assessments, prototype items and tasks will be field tested through small pilots and evaluated through a method of micro-experimentation (within-subjects design) and revision, with the most promising variations further tested through a larger-scale field trial across the Consortium states. The effectiveness of these design variations will be evaluated by examining the contribution of specific task features and accommodations to the performance of students from special populations.

The goal of a *micro-experimentation* methodology, coupled with construct and predictive validity studies, is to improve the quality of the assessment scales and tools used to measure the academic achievement of underserved student populations, whose academic progress has proven difficult to measure validly and fairly, and to heighten their opportunities for success.

Technical Quality: Valid, Fair, and Reliable Assessments

Validity refers to the soundness of interpretations, decisions, or action related to the test scores. The theory of action undergirding this proposed system places special emphasis on construct validity—the extent to which the assessments measure students' actual performance abilities—and consequential validity—the ways in which the assessments influence teaching and learning. The study design for the proposed assessment system will includes multiple rounds of pilot testing accompanied by a full set of reliability and validity studies. These will assess the quality, relevance and rigor of the items or tasks, as well as possible sources of bias in relationship to subgroups defined by race/ethnicity, as well as special education, language, and economic status.

Construct and consequential validity. During the pilot years, we will collect qualitative data about the items/tasks and their impact on curriculum, instruction and learning. We will sponsor studies in classrooms where tasks are being administered and observe students engaged in tasks to understand the knowledge and skills they use to approach them. Developers will also interview students and teachers about their perceptions of the knowledge and skills they believe are tapped by the tasks. During the task completion, we may ask students to describe for us the content and goal(s) of the question, to evaluate student comprehension of the item, and to provide a running commentary on their own work and progress. This will help us to monitor any troublesome features of the task that may keep it from working in the manner intended. Later, we will collect evidence about how the use of the tasks influences curriculum and instruction in classrooms, and we will collect evidence about the eventual achievement on a range of measures of students in the pilot classrooms and similar students in non-pilot classrooms. The qualitative data will help us to form and evaluate our arguments around validity and reliability.

Reliability of measurement. In the field test stage, lessons learned from the multiple pilots will inform and guide changes and modifications to our items and performance tasks. The results from the large scale field tests (years 3 and 4) should provide traditional metrics of reliability and validity based on quantitative evidence. Among these will be analyses of test-retest reliability, based on administrations of tasks to the same students on different occasions during the pilot; measures of internal consistency (e.g., Cronbach's alpha and Kreppendorf's alpha); measures of split-half reliability; and measures of inter-rater agreement.

To develop performance tasks that are generalizable and reliable in their measurement and scoring, tasks must be carefully designed and equated. To create performance tasks, developers outline the performance outcomes that are tightly linked to the standards and provide the foundation for the development of the task components. Use of carefully designed and tested task shells or templates can support the creation of tasks that are comparable across versions of the assessment.

Using the performance outcomes as the criteria for the students' performance, a rubric is constructed to clearly define and accurately distinguish key dimensions of a students' performance. Rubrics consist of a continuum of student performance on each dimension or a point system indicating the degree to which the outcomes were met. Using rubrics aligned with the performance outcomes enables students to understand the criteria for assessing their work products at the beginning of the project and provides the teacher with a valuable method for providing specific feedback about the strengths and areas for improvement of a students' performance. It also allows reliable scoring of the tasks for purposes of summative judgment.

Scoring reliability. A key issue in the use of open-ended items and tasks is ensuring consistency in scoring. Experience in the United States and in nations that include performance tasks in their assessment systems suggests several steps needed to ensure strong inter-rater reliability. Careful scorer training and calibration against benchmarks is an important first step. Scorers will be certified based on their ability to score to benchmarks accurately. Scores will also be moderated. Moderation is a process to ensure that the same assessment standards are applied to every student. Social moderation, which is practiced in many countries, engages teachers with expert scorers in reviews within or beyond the school to evaluate how well scores reflect consistent standards, and work with teachers to calibrate scores based on further discussion and training. In some cases, individuals or panels of experts re-score a representative sample of student performance tasks and identify the degree to which student scores are consistent with or discrepant from the independent objective scoring of calibrated teachers. A related approach is the audit, in which a proportion of tasks (e.g., 10%) is sampled and re-scored. Scorers or schools with high levels of disagreement can be flagged for further investigation which can lead to retraining or rescoring of tasks.

Another approach that we will explore is statistical moderation of scores, if there are instances when teachers score their own students' work within schools or districts. Statistical moderation is used to evaluate, and in some cases, adjust the level and spread of each school's (or other unit's) assessment of its students in comparison to the same students' scores on another common examination evaluating the same constructs. It can also be used to support equating of tasks.

Item/task comparability and quality. To examine task comparability, our data collection design will also include links between items and tasks using common students. This will allow us to calibrate all of the tasks in what can be thought of as a pre-equating phase. During this pre-equating phase we may choose to implement equating processes based on the distribution of responses, like equipercentile equating, or equating processes based on more advanced technique like kernal equating. This decision will be made based on technical requirements of our sample and to ensure that our interpretations of scores align with the specific goals of the assessment. Future task administrations will rely on the pre-equating phase for score comparison. In addition, this process may be repeated over time to control for the possibility of task drift.

Finally, we will analyze data using latent trait analysis procedures. After data are collected in the *field test phase*, we will employ an item response model to provide the basis for inference about student ability captured by the assessments. This analysis will provide estimates of the task difficulty and discrimination as well as estimates of the student ability. It will also provide us with additional quantitative evidence on the quality of the tasks, overall reliability of the group of tasks, and validity of the assessment. The results of this study will also provide evidence for task revision, scoring rubric revision, or assessment design revision to improve the quality of the performance assessment for future implementations.

Bias and fairness. We will conduct bias and fairness analyses by adding parameters for differential task function (DTF) to the item response model. DTF can be thought of as the analogous concept to Differential item functioning (DIF), which is employed for item types such as multiple-choice or constructed-response items. We would estimate a DTF parameter for each task by group combination. The size of these parameters will help us to determine whether the task is fair for all groups of students. The technical interpretation of the parameter estimate is the amount by which the task would have a relative increase or decrease in difficulty for students of equal ability but from different groups. We will examine DTF according to gender, race, ethnicity, age, and also important learning groups like English language learners and studies with disabilities. This will help us to determine whether students of a specific demographic group are disadvantaged by the inclusion of a certain task in their performance assessment.

Design team and technical advisors. These responses outline the initial thinking of some participants in the Consortium, but planning for the development of a valid and reliable system will be ongoing. Fortunately, a number of the top curriculum and assessment experts in the nation have agreed to consult on design and technical issues. They include:

Jamal Abedi, CRESST, UC Davis

Eva Baker, CRESST, UCLA

David Conley, University of Oregon

Brian Gong, National Center for the Improvement of Assessment

Ed Haertel, Stanford

Joan Herman, CRESST

Rich Hill, National Center for the Improvement of Assessment

Bob Linn, National Center for the Improvement of Assessment

Scott Marion, National Center for the Improvement of Assessment

Ray Pecheone, Stanford

Jim Pellegrino, University of Illinois—Chicago

Ed Roeber, Michigan

Lorrie Shepard, CRESST, University of Colorado at Boulder

Informing Instruction and Leadership

The most important aspect of the proposed system is that it will conceptualize and support the development of formative and summative assessment elements simultaneously, and its items and tasks will *embody* the standards, modeling the kinds of work students should be encouraged and enabled to produce. The performance components of the system, especially, will both illustrate and provide the kinds of instruction students should be engaged in. Enough items and tasks will be released, so that students, teachers, administrators, and parents have a strong concrete understanding of how the standards are reflected in assessments.

High-quality instruction requires continuous learning. The ongoing information loop that is built into the comprehensive assessment system will provide the data for continuous learning. The technology platform will provide both teachers and students access to analogous formative tasks, linked to standards (with rubrics linked to learning progressions and standards), as well as scored student work on both formative and summative assessments, embedded in a professional development system. Teachers' involvement in scoring will further reinforce their understanding of the standards. As Marc Tucker notes, this internalization of the standards "has a powerful effect on instruction … and is the most powerful form of formative assessment available" (Tucker, 2010).

The proposed system will provide various users (students, parents, teachers, principals, district leaders, and state policymakers) with the particular information they need to make informed decisions. The organization of the assessment system will be coherent, meaning that all components of the system, at all levels, will align to the big conceptual ideas in the common standards. Leaders at various levels of the system will be able to use these assessment data to guide policymaking both to support and stimulate educational programs that will prepare all students for the demands of college and career.

At the classroom level, curriculum-embedded performance tasks (both formative and summative) that teachers administer will provide teachers with an array of data about students' depth of content knowledge, disciplinary skills, problem-solving and analytic reasoning capabilities. These data are generated as students work through rich, curriculum embedded tasks. The immediacy of the data enables teachers to adjust their instructional practices (e.g., to provide additional direct instruction regarding a particular skill) during instruction as teachers become aware of misconceptions and/or the extent of students' disciplinary skills (e.g., note-taking, observation, or disciplinary reasoning skills). Teachers can make these instructional adjustments for individual students and for groups of students as the data warrants. When teachers are part of the scoring of performance-based assessments, they recognize that these assessments are instructionally sensitive and learn to recognize evidence of a high-quality performance. With the use of rubrics, exemplars and other assessment tools, students will also be provided with rich data—feedback—on their performances that they can use to adjust and improve their approach to learning.

Results on performance tasks coupled with the results of on-demand test components can provide schools and districts with a more complete picture of the content knowledge and skills that all students have in a particular subject area as well evidence of what students have demonstrated that they know how to do in that discipline—such as develop a persuasive essay, pose a worthy and reasonable research question, design an inquiry, collect data, and draw evidence-based conclusions. Various types of data, which provide information about students' basic knowledge and skills, as well as higher levels of cognitive complexity, can provide schools and districts with a richer and more accurate picture of students' college and career readiness.

In addition, a comprehensive assessment system will have the capacity to show evidence of students' learning progress over time both during a school year and across years. Such information, disaggregated by school, district, and/or student demographic data, can inform the development of curricula and materials, course syllabi, pedagogical approaches, formative, classroom-based assessment and professional development foci.

A comprehensive assessment system that is accurate, credible and fair can inform policy decisions as well as leadership at all levels of the system. An assessment system that is aligned to a set of common standards, in which results are comparable across states, will support cross-state and within-group analyses of differences in student content knowledge and reasoning abilities. Combined with other evidence about school resources and teaching practices, it can inform how educator preparation and professional development, technology, instructional interventions, and the use of school time and resources are designed are allocated at the national, state, and local levels.

Leveraging Common Standards and Assessments

The proposed system will leverage the adoption of common standards and assessments by creating a substantial consortium of states that uses to Common Core Standards to create common assessments, which will enable cost savings in assessment development and administration (see "Costs" below). In addition, the system will build on leading-edge work states and assessment developers have already conducted (for example the development of computer adaptive technology for assessment in Oregon; the development, testing, and reliable scoring of high-quality performance tasks in Connecticut, Kentucky, Maine, Ohio, and Vermont; the development of new technology platforms for assessment sharing and scoring by Teachscape and others). The leveraging of these investments will accelerate improvements in assessment quality, as well as the richness and timeliness of reporting, and help support instruction, as described above.

Implementation Timeline

To reframe and rethink current state practices regarding standards based reform and accountability will require a theory of action to understand how the development and adoption of common standards might be expected to work within and across states. Expectations for systemic change at the state and federal level call for significant changes in curriculum, instruction and assessment to enable more equitable and higher levels of learning. The press for systemic change is rooted in the belief that nationally we need clear and challenging common core standards and a coherent structure of state and district leadership to support innovation that enables all students the opportunity to successfully compete in this digital, knowledge-based world. Clearly common standards will need to become embedded in practice by supporting and resourcing new directions in curricula, teacher training, instruction (including instructional materials and digital tools) and assessment practices to meet the higher expectations they convey.

Moreover, states that are joining together in consortia have considerable experience in standards based reforms and accountability, which provides the foundation for examining the impact of assessment policies on learning within and across states. Our assessment of the opportunity costs and benefits of learning from past experience is grounded in our belief that we can accelerate the pace and effectiveness of reforms in the development of the next generation of accountability through a systematic and thoughtful review of existing tools and practices. The footprints of reform will be framed by our ability to identify promising practices and challenges to curriculum, instruction and assessment in the current system and use this information to shape the development of a more balanced and equitable accountability system.

If this reform is to be successful, it will require a systematic and sustained involvement from all levels of the education system—simultaneously respecting community norms, culture and context. Therefore, the Balanced Assessment Consortium will seek new relationships among districts, states, assessment experts, and testing and technology support providers to forge dynamic partnerships around the design and development of a new assessment and accountability system that actively and intentionally involves all levels of the education system.

Overview of Goals and Timeline

- Conduct a comprehensive review of Consortium assets, best practices and challenges that will guide the development of a Balanced Assessment System.
- Recruit and convene a Technical Advisory Committee to guide and shape the development and design of a comprehensive Balanced Assessment System.
- Create development teams to convene design/development committees; develop items and tasks, in collaboration with selected test vendor(s); conduct small scale pilots, large scale field trials, and validity/reliability research; and manage implementation to support cross-state scale up.
- Design and release RFP, evaluate bids, and contract with technology vendor to develop a technology platform to support and manage the information and data/assessment needs that support the implementation of a Balanced Assessment system at all levels of the system (i.e., student, teacher, classroom, district, and state).

- Design and release RFP, evaluate bids, and contract with independent evaluator to conduct ongoing evaluation of design/development, pilots, and field trials (Years 1-4).

Year 1: Design and Development of Assessment Instruments

- Convene an expert team to review currently available curriculum and assessment tools and formulate recommendations to guide the design of a Balanced Assessment system that can be implemented within and across Consortia states.

- Develop skeletal curriculum frameworks / learning progression outlines based on the Common Core Standards, as initial guidance for assessment development and more extensive curriculum development at the state and local level.

- Convene design/development committees to develop: (a) performance outcomes, (b) common scoring rubrics, (c) on-demand standardized assessments across all grade levels, and (d) curriculum-embedded performance tasks.

- Organize cross-state vehicles to develop and share formative and benchmarked assessments aligned to common core standards

- Select pilot districts/schools, composition of design/development committees for each content area and across all targeted grade levels across Consortia States.

- Secure external review and comment on assessment instruments from stakeholders.

- Revise assessment instruments in response to external review.

- Prepare school districts and school administrators for piloting assessment instruments.

- Prepare pilot teachers and site-based or regional coaches to support the piloting of assessment instruments

- Begin design and development of technology platform [Contractor, Consortia TAC].

- Work with state boards, chiefs, legislators to begin to define policies for use of assessment instruments in state's assessment and accountability program [States with the assistance of CCSSO].

Year 2: Small Scale Pilots and Redesign

- Pre-Pilot and Pilot items and new item formats (including performance components) to inform the development of grade-level reference examinations.

- Pre-pilot and Pilot draft embedded performance tasks across pilot sites (2 rounds—fall and spring) in designated content areas

- Provide professional development and coaching to participating teachers and school administrators implementing new reference examinations and performance tasks
- Conduct scorer training and scoring sessions (fall and/or spring)
- Collect score data and student demographic/achievement data on other measures to conduct validity and reliability studies
- Collect quantitative and qualitative data to evaluate the pilot including preliminary evidence of reliability and validity (including bias and sensitivity reviews)
- Make revisions to assessment instruments (on-demand reference tests and embedded assessments) in response to feedback from pilots; redesign on-demand test items including constructed-response and embedded performance tasks in preparation for second year field trial
- Refine learning progressions based on score data and work samples from pilot
- Conduct pilots and evaluation of technology platform
- Continue to work with state leaders on defining policies for use of assessment instruments in state's assessment and accountability program [States, CCSSO]

Year 3: Field Trial and Capacity Building for Scale-Up

- Build infrastructure (human, structural and technical) within states to participate in ongoing design of assessment instruments with contractor, to conduct professional development (including teachers, students and administrators), coaching, scorer training, scoring and reporting.
- Investigate and re-purpose existing and new resources within states to support coherent practice and avoid duplication and/or competing agendas and programs.
- Provide training (face to face and electronic) and professional development for teacher leaders, coaches, and school administrators, moving toward a regional/local/State/Consortia model of implementation.
- Conduct large-scale field trial of all assessment instruments; build and pilot multiple and equitable forms for the grade-level reference examinations and a pre-equated task bank for the embedded curricula tasks (formative and summative).
- Conduct regional training of chief trainers, scorer training and conduct scoring sessions, moving toward a regional/local or Consortia-wide model for scoring (including a mix of human and artificial intelligence scoring models).

- Revise and refine assessment instruments based on feedback from field trial [Contractor, Consortia TAC with full states involvement].
- Refine learning progressions based on score data and work samples from field trial.
- Conduct validity/reliability studies (including bias and sensitivity review) for both the on-demand test and embedded performance tasks, including experimenting with strategies for combining and integrating data from both the reference tests and the embedded performance measures. [Contractor, Consortia TAC, and states)].
- Conduct preliminary standard setting process to establish proficiency levels for both the reference tests and performance measures [Contractor, Consortia TAC, states].
- Conduct qualitative and quantitative evaluation studies.
- Conduct large-scale field trial and evaluation of technology platform.
- Continue to work with State leaders on defining policies for use of assessment instruments in state’s assessment and accountability program [with CCSSO].

Year 4-5: Scale-Up

- Support all schools in administering the new assessments, with low or moderate stakes (completion of both reference assessments across grade levels and embedded performance tasks) [States, Districts].
- Provide training of trainers professional development for teacher leaders, coaches, and school administrators, moving toward a regional/local/State/Consortia model of scoring and implementation.
- Conduct regional/local professional development sessions—reference test and performance task orientation, school-based coaching, scorer training, studying score data and student item level performance and embedded student work samples.
- Continue to build and pilot new forms/items/tasks.
- Conduct validity/reliability studies (including Bias and sensitivity reviews) as well as evaluation studies.
- Design and Conduct standard setting processes for assessments.
- Refine assessment tools, including learning progressions, for full implementation.

- Support state leaders and legislators in defining policies for use of assessment instruments in state's assessment and accountability program for full implementation in year 5 [States and CCSSO].

Costs

We have argued that a balanced assessment system that includes performance components can better measure student abilities important for life in the 21st century and secure a number of benefits for the system, which are important to consider when thinking about costs.

When thinking about the cost of assessments, people too often confuse expenditures with cost. Expenditures refer to the dollar amount a state spends on assessment. Costs include both expenditures and the opportunity costs of a particular decision in terms of other valued outcomes. Furthermore, a full analysis of costs should include an estimate of benefits associated with investments. The single dollar figure associated with spending on tests does not capture these trade-offs in the overall education funding system.

Studies have documented the opportunity costs of high-stakes tests that are narrow in their focus and format, in terms of reduced classroom emphasis on the kinds of learning that promote transfer: deep treatment of topics through research, inquiry, and applications of learning to a variety of contexts; extended writing and defense of ideas; development of higher order thinking skills (Darling-Hammond & Rustique-Forrester, 2005; Picus, Adamson, Montague, & Owens, 2010). Studies have also documented the instructional benefits of the use of rich performance tasks that model good teaching and learning (for a review, see Darling-Hammond & Rustique-Forrester, 2005).

In addition, current testing systems provide very little textured information to help teachers improve learning: The tests deliver scores, rather than evidence of student work that can be closely examined and understood in terms of a learning continuum for specific skills. They reveal little of students' thinking, reasoning, and misconceptions, and almost nothing about their actual performance beyond the bounds of recognizing or guessing answers in items where they are already supplied. Because the time for testing and test preparation often does little to help students acquire transferable knowledge and skills, teachers often feel it is "lost" to instruction, rather than that it reflects, supports, and reinforces instruction. Data in the form of scores is supplied months after students have left school for the summer. Thus, the opportunity costs of current tests are fairly high and they produce relatively few benefits in terms of expanded knowledge about important student learning for students and teachers. The flip side of these opportunity costs illustrates some of the potential benefits accrued when using a performance assessment system that is information-rich in the ways that we have described.

At the same time, it is important to acknowledge that there are greater expenditures associated with the development and human scoring of open-ended items and tasks, especially when they need to be scored in ways that assure high levels of consistency.

Previous studies and cost estimates from current programs provide relatively similar estimates of these costs from multiple sources. For example, adjusted to current dollars, costs estimates from several studies for development and scoring of assessments that include substantial performance components have ranged from about $45 to $55 per pupil, based on the practices used in the early efforts undertaken in the United States (for a review, see Picus et al., 2010). This compares to about $20 per pupil for a largely multiple-choice test battery. The ratio of about 2 or 3 to 1 in terms of costs between performance-based and selected-response tests is also fairly constant across studies. (Performance assessments in European and Asian countries tend to cost considerably less—generally in the vicinity of $7-$12 per test per pupil—because of the more highly-developed routines and systems and the engagement of teachers in scoring.)

We commissioned a study from the Assessment Solutions Group (ASG) to estimate the costs and potential cost savings of the new assessments if developed by a consortium of states with technology supports and several approaches to scoring (Topol, Olson, & Roeber, 2010). ASG developed cost models providing an apples-to-apples comparison for two types of tests: a typical summative multiple-choice test with a few constructed-response items and high-quality assessment that includes more constructed-response items and new item types, such as performance events (relatively short curriculum-embedded tasks) and more ambitious performance tasks. In this study, ASG used empirically-based cost data and their cost modeling software to determine the costs of each type of assessment.

Table 3 shows the number of multiple-choice and extended response items for each grade in a typical state system, followed by a reduced number of multiple-choice items and the addition of performance tasks and events in the new high-quality assessment (HQA). The models are based on an NCLB-type assessment system (English language arts and mathematics tests in Grades 3-8 and Grade 10).

This model estimates that a current multiple-choice test in a typical state costs around $20 per pupil. In the same typical state, the high-quality assessment would cost around $55 per pupil before cost reduction strategies are applied. However, Figure 2 shows the results of estimating costs for both the multiple-choice test and HQA and the incremental cost-savings achieved through four different approaches:

- The cost savings of *participating in a consortium*. The model includes state consortia sizes of 10, 20, and 30 states. The use of a state consortium reduces costs by an average of $15 per pupil. The consortium approach represents a significant decrease in assessment cost.

- *Uses of technology* for online test delivery, distributed human scoring of some of the open-ended items, and automated scoring for certain constructed-response items. Together these innovations account for cost savings of about $3 to $4 per pupil, and are likely to account for more as efficiencies are developed in programming and using technology for these purposes.

- Two approaches to the use of *teacher-moderated scoring*. The final cost-reduction strategy, teacher-moderated scoring, can net both substantial cost reductions as well as the potential professional development benefits we have discussed earlier. ASG estimates two different models for teacher-moderated scoring, one a professional development model with no additional teacher compensation beyond that supported by the state or district for normal professional development days and the other assuming a $125/day stipend to teachers. These strategies for using teachers as scorers, reducing costs by an additional $10 to $20 per pupil in a 30-state consortium.

Table 3. Summative Assessment Design

Summative Assessment	**Item Counts**				
Mathematics	Multiple-Choice	Short Constructed-Response	Extended Constructed-Response	Performance Event	Performance Task
Current Typical Assessment	50	0	2	0	0
High-Quality Assessment	25	2 (1 in Grade 3)	2 (0 in Grade 3, 1 in Grade 4)	2	2 (0 in Grade 3, 1 in Grade 4)
Summative Assessment	**Item Counts**				
English Language Arts	Multiple-Choice	Short Constructed-Response	Extended Constructed-Response	Performance Event	Performance Task
Current Typical Assessment (Reading)	50	0	2	0	0
Current Typical Assessment (Writing)*	10	0	1	0	0
High-Quality Assessment (Reading)	25	2 (1 in Grade 3 & 4)	2 (1 in Grade 3 & 4)	2	1
High-Quality Assessment (Writing)*	10	2 (1 in Grade 3 & 4)	2 (1 in Grade 3 & 4)	2	0

*Administered in Grades 4, 7, and 10

Combining every possible cost-saving strategy provides a per-pupil cost for the high-quality assessment of just under $10 per pupil, or around half of the estimated cost of the typical summative state test. This figure matches the costs found in many international performance assessments, which tend to run between about $7 and $12 per test per pupil—often for quite extensive performance assessments. Given that not all efficiencies will fall into place immediately, and that scoring costs will start out higher as the system is instituted and teachers are newly trained, we would feel more comfortable suggesting that it is possible to develop and administer a much more performance-based and educationally informative assessment system at no more than the same cost we are currently paying for tests with fewer benefits and larger opportunity costs.

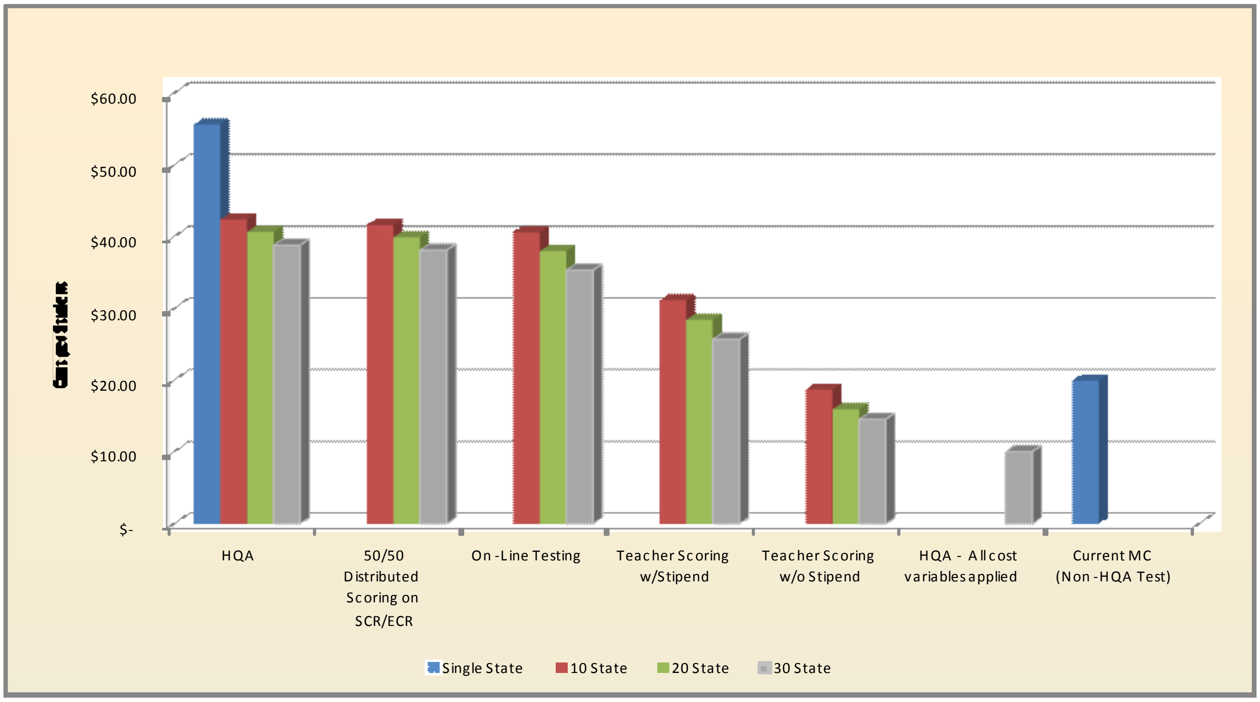

Figure 2. Diminishing Expenditures per Capita for High-Quality Assessments.

While some of the cost-saving approaches, such as online delivery, can also apply to multiple-choice tests, most cost-saving measures are achieved through different modes of scoring more extensive performance events and tasks. When combined with the benefits of professional development, teacher-moderated scoring has the potential to increase the efficiency of assessment systems and encourage instruction of higher-order thinking skills that would be tested using performance assessments.

Limitations and Trade-Offs

Assessment systems in the United States today encounter a number of tensions and trade-offs because of all the demands that are placed upon them. Three important tensions are;

- *Growth vs. Grade-Level Standards:* Evaluating growth well (and better assessing English learners and students with special needs, as well as others who are achieving well above or below the median) means measuring students along a broad achievement continuum rather than focusing only on evaluating grade-level standards or focusing on cut scores around a proficiency mark. Our system, while it will accommodate evaluation of students against grade-level standards, emphasizes the measurement of learning and growth over time.

- *Learning Value vs. Speed of Reporting:* High-quality assessments that engage educators in evaluating student work will use item types and scoring processes that cannot be universally scored by machine within a time-frame like 72 hours. While some results in a computer supported system may be turned around almost immediately, some elements will take longer to evaluate. At the same time, teachers involvement can be structured to give them insights into student performance in real-time and strategic use of distributed human scoring of open-ended items and tasks should allow results within weeks rather than the current 5 to 6 months for today's paper-and-pencil testing programs. We do privilege the quality of items / tasks and the involvement of teachers in scoring performances over the value of rapid-fire return of all scores.

- *Local Management vs. Ease of Moderation:* While teacher involvement in local scoring can provide them with valuable, nearly immediate information about their own students' thinking and learning, this approach requires extensive ongoing moderation to ensure comparability. More centralized approaches to teacher scoring that involve moderation, calibration, and auditing offer greater comparability but delay scoring. For summative purposes, we recognize the importance of creating well-functioning moderation and auditing systems to maintain credibility of the system.

Value Versus Burden

A balanced and comprehensive assessment system aligned to a set of common standards will place burdens on the educational system: it calls for developing a broadly shared vision of what college and career readiness means; it requires the development of high-quality teaching in every classroom; it asks teachers to become knowledgeable about standards and corresponding bigdisciplinary ideas; it asks teachers and leaders to be assessment literate, capable of participating in the assessment system, and well-informed about common benchmarks and their representation in curriculum; it calls for an easy-to-use, reliable online technology platform; and it necessitates a comprehensive and coherent approach to professional development that aligns with the goals of the assessment system while meeting the needs of local educators.

In short, from one vantage point, a balanced and comprehensive assessment system could tax the capacity of our educational system. From another perspective, the development of a comprehensive approach to assessment will reshape the system in ways that will bring considerable instructional benefit that cannot be achieved in any other way.

Building a truly comprehensive assessment system provides an opportunity for states to invest in a coherent system of education that will actually build the capacity of the system to provide a higher quality of education for all students, while also creating a system that is both dynamic and resource generating. A comprehensive assessment system will generate resources such as knowledge, local leadership, assessment tools, curricula, organizational routines, and structures to deepen and support ongoing learning.

Among the values of a balanced and comprehensive assessment system is that it would embody the range and depth of what we want students to know and do. Such a system can measure and help stimulate a much greater range of competencies students are expected to develop and demonstrate. It can provide assessment data in usable formats to all the various constituents who have a stake in these results and who need this data to make informed decisions to improve student learning: students, parents, teachers, principals, district leaders, state policymakers, higher education and employers. It can develop coherence by intentionally connecting assessment, instruction and curriculum. It can create assessments that are "instructionally sensitive," and thereby make quality teaching and the qualities of excellent teaching more explicit and visible. We need an assessment system with these capabilities.

System Level

The value of a balanced assessment system, developed in a consortium of states with strategic attention to design and implementation decisions, will outweigh the burden of its implementation. While developing a comprehensive assessment system has costs associated with it in terms of the time and money required to build the range of assessments and the accompanying reporting mechanisms, this investment will yield recurring dividends in the form of ongoing learning about high-quality instructional practices that are fitted to the particular needs of real students. Having multiple methods for measuring what students know and can do provides a richer picture of student learning and related instructional practices that can better inform policymakers as well as teachers. Of significant value is that a comprehensive assessment system strengthens the educational system at the classroom level and at the statehouse by providing the type of data to each level of the system that will enable better decision-making to occur.

As research has documented, "It is exceedingly difficult for policy to change practice…. Change ultimately is a problem of the smallest unit" (McLaughlin, 2005, p. 60). Investing in an approach to assessment that includes teacher-scored, curriculum-embedded tasks will reach into classrooms to stimulate change for teachers and students. By engaging teachers in the development, use and scoring of these assessments, through coherent and meaningful professional development, teachers will develop a shared conception of high-quality instruction over time and through practice. They will internalize what counts as evidence of high-quality student work. Teachers and administrators will develop knowledge of high-quality assessment design principles and of the inter-relationships among assessment, instruction, and curriculum. The engagement of educators in assessment development also enables assessment designers to create more valid, reliable, and fair assessments, because they gain fine-grained information about the contexts in which the assessments are used. In order to achieve these effects, a comprehensive assessment system, rather than a single state test administered at the end of the year, is necessary.

A comprehensive assessment system will also require a significant investment of resources for professional development (PD). However, as a nation, we already invest billions of dollars in teacher quality and countless hours in professional development that leverages less knowledge and change in practice than engagement in assessment has been found to yield. A coherent assessment system could re-direct professional development dollars to more meaningful use and create a paradigm shift about how professional development is conceptualized and organized around the work of learning and instruction.

Finally, the development of a technology platform is critical for managing costs and building a system that can go to scale. A digitized web-based platform can help manage affordable scoring processes, including scorer training modules, facilitate the moderation processes necessary for developing valid and reliable performance-based assessment components and streamline reporting functions. A technology platform can also facilitate learning by creating living archives of practice for students, teachers, principals, and district leaders. Investing in such a technology platform becomes a mechanism for de-privitizing practice, a mechanism that can stimulate learning in, from and across remote and geographically disparate regions. The potential for learning and the opportunities for capacity building associated with such a technology investment are great and extend beyond the management of assessment data and the reporting of results.

Component Level

The set of components—annual summative tests, performance tasks, formative assessment tools, professional development and reporting systems—included in the overall comprehensive assessment system creates costs and implementation burdens, but these are the necessary cost of achieving a system that is truly coherent. Furthermore, these components are inter-related; one strengthens and supports the other. In combination, they can create a powerful system. Any other approach to assessment will fail to create the coherent, fair, and reliable educational system we need in order to educate all students for college and career readiness in this country.

Costs associated with a comprehensive assessment system are reflected in money and time. Of particular concern in this regard may be the cost of using performance tasks that meet the requirements of valid and reliable assessment. While the development, use and scoring of performance tasks does require time and expertise, the value of performance tasks and formative assessment tools far outweighs their cost. Many international examples within high-achieving nations (e.g., Alberta (Canada), Victoria (Australia), Finland, Singapore, Hong Kong) illustrate the various benefits of using performance tasks as summative measures of learning: deeply engaging teachers and students in learning; making rigorous and cognitively-demanding instruction commonplace and increasing students' achievement levels, as measured by international assessments.

Why are performance tasks such a valuable component of a balanced assessment system? Performance assessments become a systemic mechanism for *instructional resourcing*, that is, the creation in practice of instructional resources, such as the knowledge, technology, organizational and relational assets that support teaching and learning (Jacquith, 2009). Thus, *instructional resourcing* occurs in the development, use and scoring of performance tasks. As teachers participate in these activities, they gain knowledge, develop tools, and build professional relationships. As administrators and leaders begin to develop the school and district-based conditions to support the effective use of performance tasks, other sorts of organizational, knowledge and technology resources are generated, for instance, in the form of particular structures, routines or roles to support meaningful use of performance tasks for learning.

Involving teachers in the development of performance tasks engages teachers in the construction of tasks that are well suited to their curricular needs, deepens teachers' knowledge of high-quality assessment and increases the likelihood that performance assessments will enable and support effective instruction. Performance tasks, whether they are conceived of as on-demand or performance-based, are opportunities for students to problem-solve, think with disciplinary content in novel circumstances, and they are intended to be embedded in curricula. In this way, performance tasks closely connect assessment, instruction and curriculum. When teachers score performance tasks, scoring can become a rich learning experience too. Teachers develop an understanding of the features of "common" benchmark performances. This, in turn, can deepen teachers' knowledge of instructional moves to support students' learning. When teachers score student work samples, an occasion is created that de-privitizes practice and makes high-quality (and poor quality) instruction public and visible to colleagues. It is well documented that by creating windows into instructional practice, educators can learn from one another and inspire each other. Opportunities for school and district leaders are created to see first-hand which instructional patterns lead to particular characteristics of a performance and to design professional development experiences that are embedded in teachers' daily practice and specifically fitted to their needs. In these ways, the implementation of performance tasks can continuously generate a vast array of needed instructional resources to strengthen instructional practice and improve student learning.

References

Achieve. (2004). *Do graduation tests measure up? A closer look at state high school exit exams: Executive summary.* Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for psychological testing.* Washington, DC: American Educational Research Association.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academies Press.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop.* Washington, DC: National Academies Press.

Commonwealth of Kentucky. (2009). *Commonwealth accountability testing system 2007–08 technical report.* Kentucky Department of Education and Measured Progress.

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. In J. Herman & E. Haertel (Eds.), *The uses and misuses of data in accountability testing. The 104th yearbook of the national society for the study of Education: Part II* (pp. 289–319). Malden, MA: Blackwell Publishing.

Gentry, J. R. (1982). An analysis of developmental spelling in GNYS AT WRK. The Reading Teacher, 36, 192-199.

Hong Kong Education and Assessment Authority. (2009). School-based assessment: Changing the assessment culture. Retrieved October 1, 2009, from Retrieved from http://www.hkeaa.edu.hk/en/sba/.

Jaquith, A. (2009). *The creation and use of instructional resources: The puzzle of professional development* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3364072)

Lane, S. (2010). *Performance assessment: The state of the art.* Stanford, CA: Stanford Center for Opportunity Policy in Education.

McLaughlin, M. (2005). Listening and learning from the field: Tales of policy implementation and situated practice. In A. Lieberman (Ed.), *The roots of educational change* (pp. 58-72). New York, NY: Springer.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's Progress in international reading literacy study in primary schools in 40 countries.* Boston, MA: Boston College, Lynch School of Education,,TIMSS & PIRLS International Study Center.

Murnane, R., & Levy, F. (1996). *Teaching the new basic skills.* New York, NY: The Free Press.

Obama, B. (2009, March 10). *Remarks by the President to the Hispanic Chamber of Commerce on a complete and competitive American education.* Retrieved from http://www.whitehouse.gov/the_press_office/Remarks-of-the-President-to-the-United-States-Hispanic-Chamber-of-Commerce.

Organisation for Economic Co-operation and Development. (2007). *Programme for international student assessment 2006.* Paris, France: Author.

Pennock-Roman, M., & Rivera, C. (2007). The differential effects of time on accommodated vs. unaccommodated content assessments for English language learners. Houston, TX: Center on Instruction.

Picus, L, Adamson, F., Montague, W., & Owens, M. (2010). *A new conceptual framework for analyzing the costs of performance assessment.* Stanford, CA: Stanford Center for Opportunity Policy in Education.

Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academies Press.

Schmidt, W. H., Wang, H. C., & McKnight, C. (2005). Curriculum coherence: An examination of U.S. mathematics and science content standards from an international perspective. *Journal of Curriculum Studies, 37*(5), 525-559.

Shepard, L., Hammerness, K., Darling-Hammond, L., & Rust, F., with Baratz Snowden, J., Gordon, E., Gutierrez, & Pacheco, A. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275-326). San Francisco, CA: Jossey-Bass.

Singapore Examinations and Assessment Board. (2009). *Project work assessment document.* Singapore: Author.

Topol, B., Olson, J., & Roeber, E. (2010). *The cost of new higher-quality assessments: A comprehensive analysis of the potential costs for future state assessments.* Stanford, CA: Stanford Center for Opportunity Policy in Education.

Tucker, M. (2010). *An assessment system for the United States: Why not build on the best?* Washington, DC: National Center for Education and the Economy.

Valverde, G. A., & Schmidt, W. H. (2000). Greater expectations: Learning from other nations in the quest for 'world-class standards' in US school mathematics and science. *Journal of Curriculum Studies, 32*(5), 651-687.

Appendix

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)					
Dimension Description	The *Number* dimension focuses on developing students' understanding of counting, magnitude and order. The natural (counting) numbers with zero extend to positive and negative signed whole numbers (integers) and through part-whole relations and proportions of whole numbers to the rational numbers (fractions and finite decimals or infinite recurring decimals). Proportions of lengths involving sides and/or diagonals of right-angled triangles and rectangles and arcs of a circle lead to the introduction of certain irrational real numbers such as the square root of 2, the golden ratio *phi* and fractions or multiples of *pi*. Principal operations for computation with number include various algorithms for addition (aggregation), subtraction (disaggregation) and the related operations of multiplication, division and exponentiation carried out mentally, by hand using written algorithms, and using calculators, spreadsheets or other numeric processors for calculation.	The *Space* dimension focuses on developing students' understanding of shape and location. These are connected through forms of representation of two- and three-dimensional objects and the ways in which the shapes of these objects and their ideal representations can be moved or combined through transformations. Students learn about key spatial concepts including continuity, edge, surface, region, boundary, connectedness, symmetry, invariance, congruence and similarity. Principal operations for computation with space include identification and representation, construction and transformation by hand using drawing instruments, and also by using dynamic geometry technology.	The *Measurement, chance and data* dimension focuses on developing students' understanding of unit, measure and error, chance and likelihood and inference. Measure is based on the notion of unit (*informal*, *formal* and *standard*) and relates number and natural language to measuring characteristics or attributes of objects and/or events. Various technologies are used to measure, and all measurement involves error. Students learn important common measures relating to money, length, mass, time and temperature, and probability – the measure of the chance or likelihood of an event. Other measures include area, volume and capacity, weight, angle, and derived rates such as density, concentration and speed. Principal operations for computation with measurement include the use of formulas for evaluating measures, the use of technology such as dataloggers for direct and indirect measurement and related technologies for the subsequent analysis of data, and estimation of measures using comparison with prior knowledge and experience, and spatial and numerical manipulations.	The *Structure* dimension focuses on developing students' understanding of set, logic, function and algebra. It is fundamental to the concise and precise nature of mathematics and the generality of its results. Key elements of mathematical structure found in each of the dimensions of Mathematics are membership, operation, closure, identity, inverse, and the commutative, associative and distributive properties as well as other notions such as recursion and periodic behaviour. While each of these can be considered in its own right, it is in their natural combination as applied to elements of number, space, function, algebra and logic with their characteristic operations that they give rise to the mathematical systems and structures that are embodied in each of these dimensions. Principal operations for computation with structure include mental, by hand and technology-assisted calculation and symbolic manipulation by calculators, spreadsheets or computer algebra systems, with sets, logic, functions and algebra.	*Working mathematically* focuses on developing students' sense of mathematical inquiry: problem posing and problem solving, modelling and investigation. It involves students in the application of principled reasoning in mathematics, in natural and symbolic language, through the mathematical processes of conjecture, formulation, solution and communication; and also engages them in the aesthetic aspects of mathematics. In this dimension the nature, purpose and scope of individual work is connected to that of the broader mathematical community, and the historical heritage of mathematics through the discourse of working mathematically. Mental, by hand and technology-assisted methods provide complementary approaches to working mathematically.

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)					
Level 1	At Level 1, students form small sets of objects from simple descriptions and make simple correspondences between those sets. They count the size of small sets using the numbers 0 to 20. They use one-to-one correspondence to identify when two sets are equal in size and when one set is larger than another. They form collections of sets of equal size. They use ordinal numbers to describe the position of elements in a set from first to tenth. They use materials to model addition and subtraction by the aggregation (grouping together) and disaggregation (moving apart) of objects. They add and subtract by counting forward and backward using the numbers from 0 to 20.	At Level 1, students recognise copy and draw points, lines and simple free-hand curves. They identify basic two-dimensional shapes such as triangles, circles and squares and three-dimensional solids and objects such as boxes and balls. They recognise the interior and exterior of shapes and objects. They sort geometric objects according to simple descriptions. They place and orientate shapes according to simple descriptions such as *next to*, *beside*, *in front of*, *behind*, *over* and *under*. They develop and follow simple instructions to move and place shapes and objects in familiar situations in relation to what they can see, and to move themselves from one place to another.	At Level 1, students compare length, area, capacity and mass of familiar objects using descriptive terms such as *longer*, *taller*, *larger*, *holds more*, and *heavier*. They make measurements using informal units such as paces for length, handprints for area, glasses for capacity, and bricks for weight. They recognise the continuity of time and the natural cycles such as day/night and the seasons. They correctly sequence days of the week. They use informal units such as heartbeats and hand claps at regular intervals to measure and describe the passage of time. They recognise and respond to unpredictability and variability in events, such as getting or not getting a certain number on the roll of a die in a game or the outcome of a coin toss. They collect and display data related to their own activities using simple pictographs.	Standards do not apply for this dimension at this level. See learning focus at http://vels.vcaa.vic.edu.au	At Level 1, students use diagrams and materials to investigate mathematical and real life situations. They explore patterns in number and space by manipulating objects according to simple rules (for example, turning letters to make patterns like *bqbqbq*, or flipping to make *bdbdbdbd*). They test simple conjectures such as 'nine is four more than five'. They make rough estimates and check their work with respect to computations and constructions in *Number*, *Space*, and *Measurement, chance and data*. They devise and follow ways of recording computations using the digit keys and +, − and = keys on a four function calculator. They use drawing tools such as simple shape templates and geometry software to draw points, lines, shapes and simple patterns. They copy a picture of a simple composite shape such as a child's sketch of a house.
Level 2	At Level 2, students model the place value of the natural numbers from 0 to 1000. They order numbers and count to 1000 by 1s, 10s and 100s. Students skip count	At Level 2, students recognise lines, surfaces and planes, corners and boundaries; familiar two-dimensional shapes	At Level 2, students make, describe and compare measurements of length, area, volume, mass and time using informal units. They recognise the	Standards do not apply for this dimension at this level. See learning focus at http://vels.vcaa.vic.edu.au	At Level 2, students make and test simple conjectures by finding examples, counter-examples and special

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)					
Level 2 cont…	by 2s, 4s and 5s from 0 to 100 starting from any natural number. They form patterns and sets of numbers based on simple criteria such as odd and even numbers. They order money amounts in dollars and cents and carry out simple money calculations. They describe simple fractions such as one half, one third and one quarter in terms of equal sized parts of a whole object, such as a quarter of a pizza, and subsets such as half of a set of 20 coloured pencils. They add and subtract one- and two-digit numbers by counting on and counting back. They mentally compute simple addition and subtraction calculations involving one- or two-digit natural numbers, using number facts such as complement to 10, doubles and near doubles. They describe and calculate simple multiplication as repeated addition, such as $3 \times 5 = 5 + 5 + 5$; and division as sharing, such as 8 shared between 4. They use commutative and associative properties of addition and multiplication in mental computation (for example, $3 + 4 = 4 + 3$ and $3 + 4 + 5$ can be done as $7 + 5$ or $3 + 9$).	including rectangles, rhombuses and hexagons, and three-dimensional shapes and objects including pyramids, cones, and cylinders. They arrange a collection of geometric shapes, such as a set of attribute blocks, into subsets according to simple criteria, and recognise when one set of shapes is a subset of another set of shapes. They recognise and describe symmetry, asymmetry, and congruence in these shapes and objects. They accurately draw simple two-dimensional shapes by hand and construct, copy and combine these shapes using drawing tools and geometry software. They apply simple transformations to shapes (*flips*, turns, slides and enlargements) and depict both the original and transformed shape together. They specify location as a relative position, including left and right, and interpret simple networks, diagrams and maps involving a small number of points, objects or locations.	differences between non-uniform measures, such as hand-spans, to measure length, and uniform measures, such as icy-pole sticks. They judge relative capacity of familiar objects and containers by eye and make informal comparisons of weight by hefting. They describe temperature using qualitative terms (for example, cold, warm, hot). Students use formal units such as hour and minute for time, litre for capacity and the standard units of metres, kilograms and seconds. Students recognise the key elements of the calendar and place in sequence days, weeks and months. They describe common and familiar time patterns and such as the time, duration and day of regular sport training and tell the time at hours and half-hours using an analogue clock, and to hours and minutes using a digital clock. Students predict the outcome of chance events, such as the rolling of a die, using qualitative terms such as certain, likely, unlikely and impossible. They collect simple categorical and numerical data (count of frequency) and present this data using pictographs and simple bar graphs.		cases and informally decide whether a conjecture is likely to be true. They use place value to enter and read displayed numbers on a calculator. They use a four-function calculator, including use of the constant addition function and *x* key, to check the accuracy of mental and written estimations and approximations and solutions to simple number sentences and equations.

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)					
Level 3 Level 3 cont…	At Level 3, students use place value (as the idea that ‘ten of these is one of those’) to determine the size and order of whole numbers to tens of thousands, and decimals to hundredths. They round numbers up and down to the nearest unit, ten, hundred, or thousand. They develop fraction notation and compare simple common fractions such as ${}^{3}/_{4} > {}^{2}/_{3}$ using physical models. They skip count forwards and backwards, from various starting points using multiples of 2, 3, 4, 5, 10 and 100. They estimate the results of computations and recognise whether these are likely to be over-estimates or under-estimates. They compute with numbers up to 30 using all four operations. They provide automatic recall of multiplication facts up to 10 × 10. They devise and use written methods for: whole number problems of addition and subtraction involving numbers up to 999 multiplication by single digits (using recall of multiplication tables) and multiples and powers of ten (for example, 5 × 100, 5 × 70) division by a single-digit divisor (based on inverse relations in multiplication tables). They devise and use algorithms for the addition and subtraction of numbers to two decimal places, including situations involving	At Level 3, students recognise and describe the directions of lines as vertical, horizontal or diagonal. They recognise angles are the result of rotation of lines with a common end-point. They recognise and describe polygons. They recognise and name common three-dimensional shapes such as spheres, prisms and pyramids. They identify edges, vertices and faces. They use two-dimensional nets, cross-sections and simple projections to represent simple three-dimensional shapes. They follow instructions to produce simple tessellations (for example, with triangles, rectangles, hexagons) and puzzles such as tangrams**.** They locate and identify places on maps and diagrams. They give travel directions and describe positions using simple compass directions (for example, N for North) and grid references on a street directory**.**	At Level 3, students estimate and measure length, area, volume, capacity, mass and time using appropriate instruments. They recognise and use different units of measurement including informal (for example, paces), formal (for example, centimetres) and standard metric measures (for example, metre) in appropriate contexts. They read linear scales (for example, tape measures) and circular scales (for example, bathroom scales) in measurement contexts. They read digital time displays and analogue clock times at five-minute intervals. They interpret timetables and calendars in relation to familiar events. They compare the likelihood of everyday events (for example, the chances of rain and snow). They describe the fairness of events in qualitative terms. They plan and conduct chance experiments (for example, using colours on a spinner) and display the results of these experiments. They recognise different types of data: non-numerical (categories), separate numbers (discrete), or points on an unbroken number line (continuous).They use a column or bar graph to display the results of an experiment (for example, the frequencies of possible categories).	At Level 3, students recognise that the sharing of a collection into equal-sized parts (division) frequently leaves a remainder. They investigate sequences of decimal numbers generated using multiplication or division by 10. They understand the meaning of the ‘=’ in mathematical statements and technology displays (for example, to indicate either the result of a computation or equivalence). They use number properties in combination to facilitate computations (for example, 7 + 10 + 13 = 10 + 7 + 13 = 10 + 20). They multiply using the distributive property of multiplication over addition (for example, 13 × 5 = (10 + 3) × 5 = 10 × 5 + 3 × 5). They list all possible outcomes of a simple chance event. They use lists, Venn diagrams and grids to show the possible combinations of two attributes. They recognise samples as subsets of the population under consideration (for example, pets owned by class members as a subset of pets owned by all children). They construct number sentences with missing numbers and solve them.	At Level 3, students apply number skills to everyday contexts such as shopping, with appropriate rounding to the nearest five cents. They recognise the mathematical structure of problems and use appropriate strategies (for example, recognition of sameness, difference and repetition) to find solutions. Students test the truth of mathematical statements and generalisations. For example, in: number (which shapes can be easily used to show fractions) computations (whether products will be odd or even, the patterns of remainders from division) number patterns (the patterns of ones digits of multiples, terminating or repeating decimals resulting from division) shape properties (which shapes have symmetry, which solids can be stacked) transformations (the effects of slides, reflections and turns on a shape) measurement (the relationship between size and capacity of a container).

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)					
	money. They add and subtract simple common fractions with the assistance of physical models.				Students use calculators to explore number patterns and check the accuracy of estimations. They use a variety of computer software to create diagrams, shapes, tessellations and to organise and present data.
Level 4 Level 4 cont…	At Level 4, students comprehend the size and order of small numbers (to thousandths) and large numbers (to millions). They model integers (positive and negative whole numbers and zero), common fractions and decimals. They place integers, decimals and common fractions on a number line. They create sets of number multiples to find the lowest common multiple of the numbers. They interpret numbers and their factors in terms of the area and dimensions of rectangular arrays (for example, the factors of 12 can be found by making rectangles of dimensions 1 × 12, 2 × 6, and 3 × 4). Students identify square, prime and composite numbers. They create factor sets (for example, using factor trees) and identify the highest common factor of two or more numbers. They recognise and calculate simple powers of whole numbers (for example, $2^4 = 16$). Students use decimals, ratios and percentages to find equivalent representations of common fractions (for example, ${}^{3}/_{4} = {}^{9}/_{12} =$	At Level 4, students classify and sort shapes and solids (for example, prisms, pyramids, cylinders and cones) using the properties of lines (orientation and size), angles (less than, equal to, or greater than 90°), and surfaces. They create two-dimensional representations of three dimensional shapes and objects found in the surrounding environment. They develop and follow instructions to draw shapes and nets of solids using simple scale. They describe the features of shapes and solids that remain the same (for example, angles) or change (for example, surface area) when a shape is enlarged or reduced. They apply a range of transformations to shapes and create tessellations using tools (for example, computer	At Level 4, students use metric units to estimate and measure length, perimeter, area, surface area, mass, volume, capacity time and temperature. They measure angles in degrees. They measure as accurately as needed for the purpose of the activity. They convert between metric units of length, capacity and time (for example, L–mL, sec–min). Students describe and calculate probabilities using words, and fractions and decimals between 0 and 1. They calculate probabilities for chance outcomes (for example, using spinners) and use the symmetry properties of equally likely outcomes. They simulate chance events (for example, the chance that a family has three girls in a row) and understand that experimental estimates of probabilities converge to the theoretical probability in the long run. Students recognise and give consideration to different data types in forming questionnaires and sampling. They distinguish between categorical and	At Level 4, students form and specify sets of numbers, shapes and objects according to given criteria and conditions (for example, 6, 12, 18, 24 are the even numbers less than 30 that are also multiples of three). They use Venn diagrams and Karnaugh maps to test the validity of statements using the words *none*, *some* or *all* (for example, test the statement '*all* the multiples of 3, less than 30, are even numbers'). Students construct and use rules for sequences based on the previous term, recursion (for example, the next term is three times the last term plus two), and by formula (for example, a term is three times its position in the sequence plus two). Students establish equivalence relationships between mathematical expressions using properties such as the distributive property for multiplication over addition (for example, 3 × 26 = 3 × (20 + 6)). Students identify relationships between variables and describe them with language and words (for example, how hunger depends varies with time of the day). Students recognise that addition and subtraction, and multiplication and division are inverse operations. They use words and symbols to form simple equations. They solve equations by trial and error.	At Level 4, students recognise and investigate the use of mathematics in real (for example, determination of test results as a percentage) and historical situations (for example, the emergence of negative numbers). Students develop and test conjectures. They understand that a few successful examples are not sufficient proof and recognise that a single counter-example is sufficient to invalidate a conjecture. For example, in: number (all numbers can be shown as a rectangular array) computations (multiplication leads to a larger number) number patterns (the next number in the sequence 2,4,6 … must be 8) shape properties (all parallelograms are

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)						
	0.75 = 75% = 3 : 4 = 6 : 8). They explain and use mental and written algorithms for the addition, subtraction, multiplication and division of natural numbers (positive whole numbers). They add, subtract, and multiply fractions and decimals (to two decimal places) and apply these operations in practical contexts, including the use of money. They use estimates for computations and apply criteria to determine if estimates are reasonable or not.	software). Students use the ideas of size, scale, and direction to describe relative location and objects in maps. They use compass directions, coordinates, scale and distance, and conventional symbols to describe routes between places shown on maps. Students use network diagrams to show relationships and connectedness such as a family tree and the shortest path between towns on a map.	numerical data and classify numerical data as discrete (from counting) or continuous (from measurement). They present data in appropriate displays (for example, a pie chart for eye colour data and a histogram for grouped data of student heights). They calculate and interpret measures of centrality (mean, median, and mode) and data spread (range).		rectangles) chance (a six is harder to roll on die than a one). Students use the mathematical structure of problems to choose strategies for solutions. They explain their reasoning and procedures and interpret solutions. They create new problems based on familiar problem structures. Students engage in investigations involving mathematical modelling. They use calculators and computers to investigate and implement algorithms (for example, for finding the lowest common multiple of two numbers), explore number facts and puzzles, generate simulations (for example, the gender of children in a family of four children), and transform shapes and solids.
Level 5	At Level 5, students identify complete factor sets for natural numbers and express these natural numbers as products of powers of primes (for example, 36 000 = $2^5 \times 3^2 \times 5^3$). They write equivalent fractions for a fraction given in simplest form (for example, $^2/_3 = {}^4/_6 = {}^6/_9 = \ldots$). They know the decimal equivalents for the unit fractions $^1/_2$, $^1/_3$, $^1/_4$, $^1/_5$, $^1/_8$, $^1/_9$ and find equivalent	At Level 5, students construct two-dimensional and simple three-dimensional shapes according to specifications of length, angle and adjacency. They use the properties of parallel lines and transversals of these lines to calculate angles that are supplementary, corresponding, allied (co-	At Level 5, students measure length, perimeter, area, surface area, mass, volume, capacity, angle, time and temperature using suitable units for these measurements in context. They interpret and use measurement formulas for the area and perimeter of circles, triangles and parallelograms and simple composite shapes. They calculate the surface area and volume of	At Level 5, students identify collections of numbers as subsets of natural numbers, integers, rational numbers and real numbers. They use Venn diagrams and tree diagrams to show the relationships of intersection, union, inclusion (subset) and complement between the sets. They list the elements of the set of all subsets (power set) of a given finite set and comprehend the partial-order relationship between these subsets with respect to inclusion (for example, given the set {*a*, *b*, *c*} the	At Level 5, students formulate conjectures and follow simple mathematical deductions (for example, if the side length of a cube is doubled, then the surface area increases by a factor of four, and the volume increases by a factor of eight). Students use variables in

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)					
Level 5 cont… Level 5 cont…	representations of fractions as decimals, ratios and percentages (for example, a subset: set ratio of 4:9 can be expressed equivalently as $^4/_9 = 0.4 \approx 44.44\%$). They write the reciprocal of any fraction and calculate the decimal equivalent to a given degree of accuracy. Students use knowledge of perfect squares when calculating and estimating squares and square roots of numbers (for example, $20^2 = 400$ and $30^2 = 900$ so $\sqrt{700}$ is between 20 and 30). They evaluate natural numbers and simple fractions given in base-exponent form (for example, $5^4 = 625$ and $(^2/_3)^2 = {^4/_9}$). They know simple powers of 2, 3, and 5 (for example, $2^6 = 64$, $3^4 = 81$, $5^3 = 125$). They calculate squares and square roots of rational numbers that are perfect squares (for example, $\sqrt{0.81} = 0.9$ and $\sqrt{^9/_{16}} = {^3/_4}$). They calculate cubes and cube roots of perfect cubes (for example, $\sqrt[3]{64} = 4$). Using technology they find square and cube roots of rational numbers to a specified degree of accuracy (for example, $\sqrt[3]{200} = 5.848$ to three decimal places). Students express natural numbers base 10 in binary form, (for example, $42_{10} = 101010_2$), and add and multiply natural numbers in binary form (for example, $101_2 + 11_2 = 1000_2$ and $101_2 \times 11_2 = 1111_2$). Students understand ratio as both set: set comparison (for example,	interior) and alternate. They describe and apply the angle properties of regular and irregular polygons, in particular, triangles and quadrilaterals. They use two-dimensional nets to construct a simple three-dimensional object such as a prism or a platonic solid. They recognise congruence of shapes and solids. They relate similarity to enlargement from a common fixed point. They use single-point perspective to make a two-dimensional representation of a simple three-dimensional object. They make tessellations from simple shapes. Students use coordinates to identify position in the plane. They use lines, grids, contours, isobars, scales and bearings to specify location and direction on plans and maps. They use network diagrams to specify relationships. They consider the connectedness of a network, such as the ability to travel through a set of roads between towns.	prisms and cylinders. Students estimate the accuracy of measurements and give suitable lower and upper bounds for measurement values. They calculate absolute percentage error of estimated values. Students use appropriate technology to generate random numbers in the conduct of simple simulations. Students identify empirical probability as long-run relative frequency. They calculate theoretical probabilities by dividing the number of possible successful outcomes by the total number of possible outcomes. They use tree diagrams to investigate the probability of outcomes in simple multiple event trials. Students organise, tabulate and display discrete and continuous data (grouped and ungrouped) using technology for larger data sets. They represent uni-variate data in appropriate graphical forms including dot plots, stem and leaf plots, column graphs, bar charts and histograms. They calculate summary statistics for measures of centre (mean, median, mode) and spread (range, and mean absolute difference), and make simple inferences based on this data.	corresponding power set is $\{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{a, c\}, \{a, b, c\}\}$.) They test the validity of statements formed by the use of the connectives *and, or, not,* and the quantifiers *none*, *some* and *all*, (for example, 'some natural numbers can be expressed as the sum of two squares'). They apply these to the specification of sets defined in terms of one or two attributes, and to searches in data-bases. Students apply the commutative, associative, and distributive properties in mental and written computation (for example, 24×60 can be calculated as $20 \times 60 + 4 \times 60$ or as $12 \times 12 \times 10$). They use exponent laws for multiplication and division of power terms (for example $2^3 \times 2^5 = 2^8$, $2^0 = 1$, $2^3 \div 2^5 = 2^{-2}$, $(5^2)^3 = 5^6$ and $(3 \times 4)^2 = 3^2 \times 4^2$). Students generalise from perfect square and difference of two square number patterns (for example, $25^2 = (20 + 5)^2 = 400 + 2 \times (100) + 25 = 625$. And $35 \times 25 = (30 + 5)(30 - 5) = 900 - 25 = 875$). Students recognise and apply simple geometric transformations of the plane such as translation, reflection, rotation and dilation and combinations of the above, including their inverses. They identify the identity element and inverse of rational numbers for the operations of addition and multiplication (for example, $^1/_2 + {^{-1}/_2} = 0$ and $^2/_3 \times {^3/_2} = 1$). Students use inverses to rearrange simple mensuration formulas, and to find equivalent algebraic expressions (for example, if $P = 2L + 2W$, then $W = {^P/_2} -$ L. If $A = \pi r^2$ then $r = \sqrt{^A/_\pi}$). They solve simple equations (for example, $5x + 7 = 23$, $1.4x - 1.6 = 8.3$, and $4x^2 - 3 =$	general mathematical statements. They substitute numbers for variables (for example, in equations, inequalities, identities and formulas). Students explain geometric propositions (for example, by varying the location of key points and/or lines in a construction). Students develop simple mathematical models for real situations (for example, using constant rates of change for linear models). They develop generalisations by abstracting the features from situations and expressing these in words and symbols. They predict using interpolation (working with what is already known) and extrapolation (working beyond what is already known). They analyse the reasonableness of points of view, procedures and results, according to given criteria, and identify limitations and/or constraints in context. Students use technology such as graphic calculators, spreadsheets, dynamic geometry software and computer algebra systems for a

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)					
	number of boys : number of girls) and subset: set comparison (for example, number of girls : number of students), and find integer proportions of these, including percentages (for example, the ratio number of girls: the number of boys is 2 : 3 = 4 : 6 = 40% : 60%). Students use a range of strategies for approximating the results of computations, such as front-end estimation and rounding (for example, $925 \div 34 \approx 900 \div 30 = 30$). Students use efficient mental and/or written methods for arithmetic computation involving rational numbers, including division of integers by two-digit divisors. They use approximations to π in related measurement calculations (for example, $\pi \times 5^2 = 25\pi = 78.54$ correct to two decimal places). They use technology for arithmetic computations involving several operations on rational numbers of any size.			13) using tables, graphs and inverse operations. They recognise and use inequality symbols. They solve simple inequalities such as $y \le 2x + 4$ and decide whether inequalities such as $x^2 > 2y$ are satisfied or not for specific values of x and y. Students identify a function as a one-to-one correspondence or a many-to-one correspondence between two sets. They represent a function by a table of values, a graph, and by a rule. They describe and specify the independent variable of a function and its domain, and the dependent variable and its range. They construct tables of values and graphs for linear functions. They use linear and other functions such as $f(x) = 2x - 4$, $xy = 24$, $y = 2^x$ and $y = x^2 - 3$ to model various situations.	range of mathematical purposes including numerical computation, graphing, investigation of patterns and relations for algebraic expressions, and the production of geometric drawings.

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)					
Level 6 Level 6 cont… Level 6 cont…	At Level 6, students comprehend the set of real numbers containing natural, integer, rational and irrational numbers. They represent rational numbers in both fractional and decimal (terminating and infinite recurring) forms (for example, $1^4/_{25} = 1.16$, $0.\overline{47} = {}^{47}/_{99}$). They comprehend that irrational numbers have an infinite non-terminating decimal form. They specify decimal rational approximations for square roots of primes, rational numbers that are not perfect squares, the golden ratio φ, and simple fractions of π correct to a required decimal place accuracy. Students use the Euclidean division algorithm to find the greatest common divisor (highest common factor) of two natural numbers 9 (for example, the greatest common divisor of 1071 and 1029 is 21 since $1071 = 1029 \times 1 + 42$, $1029 = 42 \times 24 + 21$ and $42 = 21 \times 2 + 0$). Students carry out arithmetic computations involving natural numbers, integers and finite decimals using mental and/or written algorithms (one- or two-digit divisors in the case of division). They perform computations involving very large or very small numbers in scientific notation (for example, $0.0045 \times 0.000028 = 4.5 \times 10^{-3} \times 2.8 \times 10^{-5} = 1.26 \times 10^{-7}$).	At Level 6, students represent two- and three-dimensional shapes using lines, curves, polygons and circles. They make representations using perspective, isometric drawings, nets and computer-generated images. They recognise and describe boundaries, surfaces and interiors of common plane and three-dimensional shapes, including cylinders, spheres, cones, prisms and polyhedra. They recognise the features of circles (centre, radius, diameter, chord, arc, semi-circle, circumference, segment, sector and tangent) and use associated angle properties. Students explore the properties of spheres. Students use the conditions for shapes to be congruent or similar. They apply isometric and similarity transformations of geometric shapes in the plane. They identify points that are invariant under a given transformation (for example, the point (2, 0) is invariant under reflection in the *x*-axis, so the *x* axis intercept of the graph of $y = 2x - 4$ is also invariant	At Level 6, students estimate and measure length, area, surface area, mass, volume, capacity and angle. They select and use appropriate units, converting between units as required. They calculate constant rates such as the density of substances (that is, mass in relation to volume), concentration of fluids, average speed and pollution levels in the atmosphere. Students decide on acceptable or tolerable levels of error in a given situation. They interpret and use mensuration formulas for calculating the perimeter, surface area and volume of familiar two- and three-dimensional shapes and simple composites of these shapes. Students use pythagoras' theorem and trigonometric ratios (sine, cosine and tangent) to obtain lengths of sides, angles and the area of right-angled triangles. They use degrees and radians as units of measurement for angles and convert between units of measurement as appropriate. Students estimate probabilities based on data (experiments, surveys, samples, simulations) and assign and justify subjective probabilities in familiar situations. They list event spaces (for combinations of up to three events) by lists, grids, tree diagrams, venn diagrams and karnaugh maps (two-way tables). They calculate probabilities for	At Level 6, students classify and describe the properties of the real number system and the subsets of rational and irrational numbers. They identify subsets of these as discrete or continuous, finite or infinite and provide examples of their elements and apply these to functions and relations and the solution of related equations. Student express relations between sets using membership, $\in$, complement, $'$, intersection, $\cap$, union, $\cup$, and subset, $\subseteq$, for up to three sets. They represent a universal set as the disjoint union of intersections of up to three sets and their complements, and illustrate this using a tree diagram, venn diagram or karnaugh map. Students form and test mathematical conjectures; for example, 'What relationship holds between the lengths of the three sides of a triangle?' They use irrational numbers such as, π, φ and common surds in calculations in both exact and approximate form. Students apply the algebraic properties (closure, associative, commutative, identity, inverse and distributive) to computation with number, to rearrange formulas, rearrange and simplify algebraic expressions involving real variables. They verify the equivalence or otherwise of algebraic expressions (linear, square, cube, exponent, and reciprocal, (for example, $4x - 8 = 2(2x - 4) = 4(x - 2)$; $(2a - 3)^2 = 4a^2 - 12a + 9$; $(3w)^3 = 27w^3$; $(x^3y) / xy^2 = xy^2y^{-1}$; ${}^4/_{xy} = {}^2/_x \times {}^2/_y$). Students identify and represent linear, quadratic and exponential functions by table, rule and graph (all four quadrants of	At Level 6, students formulate and test conjectures, generalisations and arguments in natural language and symbolic form (for example, 'if m^2 is even then m is even, and if m^2 is odd then m is odd'). They follow formal mathematical arguments for the truth of propositions (for example, 'the sum of three consecutive natural numbers is divisible by 3'). Students choose, use and develop mathematical models and procedures to investigate and solve problems set in a wide range of practical, theoretical and historical contexts (for example, exact and approximate measurement formulas for the volumes of various three dimensional objects such as truncated pyramids). They generalise from one situation to another, and investigate it further by changing the initial constraints or other boundary conditions. They judge the reasonableness of their results based on the context under consideration. They select and use

Victoria Essential Learning Standards: Mathematics, Levels 1-6 (Grades 1-10)						
	They carry out exact arithmetic computations involving fractions and irrational numbers such as square roots (for example, $\sqrt{18} = 3\sqrt{2}$, $\sqrt{(^3/_2)} = {}^{(\sqrt{6})}/_2$) and multiples and fractions of π (for example $\pi + {}^{\pi}/_4 = {}^{5\pi}/_4$). They use appropriate estimates to evaluate the reasonableness of the results of calculations involving rational and irrational numbers, and the decimal approximations for them. They carry out computations to a required accuracy in terms of decimal places and/or significant figures.	under this transformation). They determine the effect of changing the scale of one characteristic of two- and three-dimensional shapes (for example, side length, area, volume and angle measure) on related characteristics. They use latitude and longitude to locate places on the Earth's surface and measure distances between places using great circles. Students describe and use the connections between objects/location/events according to defined relationships (networks).	complementary, mutually exclusive, and compound events (defined using *and*, *or* and *not*). They classify events as dependent or independent. Students comprehend the difference between a population and a sample. They generate data using surveys, experiments and sampling procedures. They calculate summary statistics for centrality (mode, median and mean), spread (box plot, inter-quartile range, outliers) and association (by-eye estimation of the line of best fit from a scatter plot). They distinguish informally between association and causal relationship in bi-variate data, and make predictions based on an estimated line of best fit for scatter-plot data with strong association between two variables.	the Cartesian coordinate system) with consideration of independent and dependent variables, domain and range. They distinguish between these types of functions by testing for constant first difference, constant second difference or constant ratio between consecutive terms (for example to distinguish between the functions described by the sets of ordered pairs {(1, 2), (2, 4), (3, 6), (4, 8) …}; {(1, 2), (2, 4), (3, 8), (4, 14) …}; and {(1, 2), (2, 4), (3, 8), (4, 16) …}). They use and interpret the functions in modelling a range of contexts. They recognise and explain the roles of the relevant constants in the relationships $f(x) = ax + c$, with reference to gradient and y axis intercept, $f(x) = a(x + b)\ 2 + c$ and $f(x) = ca^x$. They solve equations of the form $f(x) = k$, where k is a real constant (for example, $x(x + 5) = 100$) and simultaneous linear equations in two variables (for example, $\{2x - 3y = -4$ and $5x + 6y = 27\}$ using algebraic, numerical (systematic guess, check and refine or bisection) and graphical methods.	technology in various combinations to assist in mathematical inquiry, to manipulate and represent data, to analyse functions and carry out symbolic manipulation. They use geometry software or graphics calculators to create geometric objects and transform them, taking into account invariance under transformation.