A FRESH START
Creating the Next Generation of Assessment Systems

A report based on the National Conference on Next Generation Assessment Systems
March 8-9, 2010

by The Center for K-12 Assessment & Performance Management at ETS
INTRODUCTION

What Could the Next Generation Of Assessments Look Like?

Assessment policies and practices touch everyone in the education system, from students and parents to educators and government leaders, because they serve as the yardstick for our performance in education and a predictor of our nation’s economic and social future. For at least the last two decades, assessments and the accountability systems they support have created controversy and sometimes thwarted the intentions of education reforms. If nothing else, the national and state reform initiatives have shown the public and policymakers that assessment policy is critical to “get right” since it influences education practices at all levels of the system. The federal Race to the Top (RTTT) Assessment Program offers a unique opportunity to rethink and clarify the priority policy objectives, as well as how the assessment systems should integrate with and support curriculum, instruction and management systems in order to maximize student achievement. To frame these conversations, the new Center for K-12 Assessment and Performance Management (K-12 Center) at the Educational Testing Service (ETS) was pleased to plan the program for the National Conference on Next Generation Assessment Systems, held in Washington, DC, in March, 2010. Collaborating with the Conference host organizations, the Education Commission of the States and the Council of the Great City Schools, teams of state and urban district leaders as well as measurement experts and national policy leaders were brought together. The purpose was to stimulate thinking about new designs for assessment systems and to address the challenges and opportunities in the RTTT Assessment Program.

The conference centered on visions for the next generation of assessment systems, prepared by national leaders in the fields of curriculum, instruction and assessment. These four conceptual models are neither intended to cover the landscape of options nor to fuel a quest for “one best option.” Rather, they were brought forward in order to get the conversation started and prod deeper thinking about the options, their strengths, and their trade-offs.

Best-selling author and New York Times columnist Tom Friedman set the “why” for the conference with his views on how the platform for global collaboration and dissemination created by the World Wide Web has fundamentally changed the drivers of competitive advantage - for individuals as well as countries and companies – making creativity our most strategic resource. Andreas Schleicher of OECD concluded the conference with his overview of international developments in assessment policy, indicating that countries around the world are addressing similar issues and the United States has much to gain by studying those that have made dramatic improvements in student achievement. (See sidebar, page 14)

This special Education Week Supplement contains brief summaries of the papers delivered at the conference on the four assessment models as well as a table showing similarities and differences among the models. These papers have not been put through independent expert review, so the claims within each are those of the author(s). As consortia of states and districts strive to make the most of this opportunity to develop more useful and truly “next generation” assessment systems, they should take time to obtain expert review and critique of claims and assumptions within these designs. These design efforts will need to be informed about current technologies and methodologies as they push for a new generation of assessment systems. Also, even multi-component systems with significantly expanded features will not be able to serve all purposes with equal strength. The assessment consortia will need to clarify their priority purposes in order to design or adapt their system accordingly.

This conference on next generation assessment systems is part of the initial work of the K12 Center, established under the aegis of ETS last year, to independently stimulate and support improvements in the field of K-12 assessment. It is directed by Pascal (Pat) Forgione, former U.S. Commissioner of the National Center on Educational Statistics, former Delaware state superintendent and former superintendent of the Austin Independent School District in Texas. The Director of Programs for the Center is Nancy Doorey of Wilmington, Delaware.

The Conference papers, presentation slides and videos, as well as a set of research papers from a December 2009 Exploratory Seminar on two central measurement issues within the RTTT reform agenda, can be found at www.k12center.org.
The Uniqueness Of the Moment

There are few moments in history when our society is presented with an inflection point—a development that will create vastly expanded opportunities to connect with the ideas of others, collaborate, and create value. The printing press was one. August 9, 1995—the date on which Netscape held its first public offering for its Internet browser—was another. And, should the Common Core Standards be broadly adopted, American education may be on the verge of its own inflection point.

In his keynote speech at the National Conference on Next Generation Assessment Systems, New York Times columnist and Pulitzer Prize-winning author Thomas Friedman used this information to provide insight into the uniqueness and urgency of the moment that we face in American education, and its implications for our students and our schools. (See sidebar, page 14.)

In Friedman’s estimation, the impact of the printing press, although transformational, was exceeded by that of Netscape, which he referred to as “the mother of all inflection points.” This new tool made it possible for users of differing computers and software to interact seamlessly, to share their ideas and products globally and to access those of others, virtually for free. In doing so, Friedman argued, we moved from a world in which “value was created in vertical silos of command and control” to a world in which “value is created horizontally by those who connect and collaborate with.” The power of one’s ideas became much more important than one’s location.

The adoption of Common Core Standards and common assessments by consortia of states would change the education world just as dramatically, by replacing “silos” of individual state standards with a national platform for connection and collaboration that will enable American educators and the education sector to tap their collective genius. A break-through strategy developed in Miami for helping 3rd-grade English language learners master a core standard would be immediately useful in Memphis, Middletown, and Minneapolis. The educational marketplace, which currently must fragment its work to develop and adapt material to 50 sets of state standards, will be able to focus on the Common Core Standards and associated learning expectations, resulting in significantly increased competition for high-impact tools and support for this set of core skills and concepts. Small subgroups of students for whom there are often too few specialized tools—those with low-incidence learning disabilities, for example—would become large enough across a number of states to spur greater R&D investments by educators and companies alike.

The release of Netscape did not change the way we work and interact overnight. Friedman explained, but was the initial, critical inflection point. Over the five years following the 1995 public offering, more than $1 trillion dollars were invested in fiber optic cable, and new industry standards for data transmission were adopted. Similarly, in order to realize the full potential presented by the Common Core Standards for American education, additional fundamental components of the platform are required. The Race to the Top (RTTT) grant program will provide some needed seed funding for the development of shared assessments, but without sufficient agreement on performance standards and data systems, fragmentation could again create silos. It will take significant effort and strong political will and leadership to realize the potential of this “inflection point.”

Educational representatives from the state, district and national levels gathered at the National Conference on Next Generation Assessment Systems were quick to see the connection between Friedman’s description of “the Netscape moment” and the opportunity currently before them. A panel discussion, which can now be viewed at www.k12center.org, featured three individuals with distinctly differing political views and many years of experience in education reform efforts.

Roy Romer has served as the Governor of Colorado, the Superintendent of Los Angeles Unified School District, and the first Chair of the National Education Goals Panel in 1991. He characterized this moment as “quite extraordinary … like one of those great waves off of Hawaii that comes along every 40 years.” Romer offered six points of advice to state and district leaders, cautioning that it will take time and require both strong commitment and flexibility as state leaders figure out the type of shared governance structure needed to support the common standards and assessments over time and update them as needed.

Chester Finn has served as the Executive Director of Fordham’s Governance Center, a Senior Advisor to President Clinton, and a Senior Fellow at the New America Foundation. Finn explained, “This is a once-in-a-generation opportunity for our kids,” one that will require strong leadership from states as they seek to find the right balance between respect for local control and the need to create maximum leverage for a vital national priority—improved educational achievement. In terms of crafting designs for next generation assessment systems, he urged leaders to obtain the advice of measurement experts, perhaps through a Request for Information process, but to remain focused on and use that information to press for solutions.

Chester E. Finn, Jr., a former Assistant Secretary of Education in the George H.W. Bush administration, current President of the Fordham Foundation and a well-known critic of reform efforts to date, noted, “It isn’t often that a society gets a chance to start afresh, and I think that moment is here.” He commended the quality of the newly released Common Core Standards, and urged states and districts to evaluate carefully options for new assessment systems and to clarify the purposes they want to optimize and the trade-offs and costs they are willing to accept.
Clarifying the Purposes Of the New Assessments

The fundamental purposes of state assessment systems have both evolved and expanded over the past two decades. At the onset of standards-based reform in the early 1990’s, many states set about developing standards and accompanying statewide assessments to monitor improvement toward those academic goals. Assessments served as the yardstick for the performance of the system, and for districts as well. Public reporting of results were the only “stakes” in most states. By the late 1990’s, state legislatures began to require that the assessment results be used for school and, in some cases, individual student and teacher accountability. The No Child Left Behind Act of 2001 mandated a national policy of using state assessment result for annual school and district accountability, and required the disaggregation of results by subgroups. With the increasing importance of the test results, additional demands developed — for more solid measurement of individual growth in addition to status, and for more timely and instructionally actionable feedback. The set of purposes these test were to serve quickly exceeded the measurement capacity of any individual assessment.

The RTTT Assessment Program incorporates the vision of an integrated system of assessments adopted by a consortium of states. Summative, interim and formative assessments will be carefully and thoughtfully aligned to college- and career-readiness standards. Also, the reporting systems will be designed and implemented to provide students, parents and educators with a more complete picture of the progress to date, current needs and future trajectory of each student. However, even such systems of assessments cannot be optimized for all purposes. As nearly every Conference presenter pointed out, states and districts will need to make clear, thoughtful choices with their constituencies about the priority purposes of their new assessment systems and the trade-offs they are willing to accept in order to join consortia with compatible views. Failure to gain clarity and consensus on these priority policy objectives of the new systems will undermine their longer term political viability, cautioned a number of presenters.

The Colorado Department of Education has spent a great deal of time over the past twelve months building consensus on the priority purposes of their assessment and accountability systems. They have asked constituents to clarify the most important questions the state should be able to answer at each level of the system (i.e., student, classroom, school, district and state).

Building on the work in Colorado, the K12 Conference attendees also wanted to ensure that the new reporting systems provide results that are useful to students and parents in understanding where the student currently is in relationship to the academic standards and future college- and career-readiness. The four assessment system models discussed over the following pages take a variety of approaches to providing instructional guidance at the student level. They do this primarily through the use of aligned formative or interim assessments that are closely tied to recent instruction. The degree to which one can obtain instructionally actionable feedback —
One of the fundamental tensions and choices states and districts will need to grapple with as they design their new assessment systems: the instructional value of the common assessment systems – both summative and formative – will be directly related to the degree to which states are willing to accept curricular consistency.

from either summative or formative assessments — will be determined in large part by the degree to which the sequence and content of instruction is defined and consistent across those using a particular assessment. This reveals one of the fundamental tensions and choices states and districts will need to grapple with as they design their new assessment systems: the instructional value of the common assessment systems – both summative and formative – will be directly related to the degree to which states are willing to accept curricular consistency.

At one end of the spectrum, consortia may choose to build common end-of-year summative assessments based on the learning progression milestones defined within the Common Core Standards, leaving the development of formative and/or interim assessments to individual states or districts. At the other end of the spectrum, as we see in the Tucker model for example, consortia could choose to come to agreement at the more detailed level of course syllabi or units of instruction, making it possible to co-invest in shared instructionally sensitive formative and summative assessments, as well as instructional materials. The Resnick-Berger model offers an interesting but as yet untested alternative model of breaking the summative assessment into a series of end-of-unit exams, augmented by an item bank of formative items and an engine to help teachers compose individualized diagnostic tests.

The point here is that consortia will need to think through the components of the system from which they want to acquire “what’s next?” feedback on the individual student level and the trade-offs, in terms of curricular uniformity, cost, and personnel time, they are willing to accept.

2. Determining “On Track to College- and Career-Readiness”

A second theme found across the responses of state, district and policy leaders is that the primacy of NCLB’s focus on the percent of students meeting annual proficiency targets has been superseded by a desire to know, based growth patterns, whether or not the students, at the individual, school, and district levels, are “on track” to meeting college- and career-readiness standards by the time they finish high school.

The Administration’s newly released blueprint for ESEA reauthorization calls for new assessments that more accurately measure individual student growth and to use that information within new accountability systems to recognize teachers, schools and districts that are putting students on track to college-and career-readiness. It has been widely recognized that current NCLB-compliant assessments were not designed to create, and have serious limitations in creating, such longitudinal projections due in part to the restriction that all items be on grade level. In order to better measure growth, three of the four assessment system models described on the following pages call for the use of computer adaptive delivery and the removal of the current federal restriction to grade-level content.

Conference participants also revealed a strong degree of consensus on the need for assessments that require the application of skills and knowledge, a priority strongly advocated by Andreas Schleicher of OECD/ PISA during his keynote address on the second day. (You will find his presentation at www.k12center.org.) On the summative assessments in particular, survey respondents called for greater use of technology for delivery, scoring, and measurement of the expanded range of skills that must be assessed.

3. Identifying Areas for Improvement

All three stakeholder group samples – state, district, and national policy leaders - placed high priority on the use of student gains data in the next generation systems for the identification of students who are making inadequate growth and require focused attention/intervention. In addition, these new systems should identify those teachers whose students (all, by specific subgroups, by subject area) are making inadequate gains in order to provide targeted professional development or other intervention. Moreover, participants would like these systems to identify an individualized professional development plan for every teacher, based on patterns in the progress of their students.

Several of the models described herein include the use of a small number of extended tasks or projects, perhaps given over the course of the school year, within the summative/ accountability system. Herein, then, lies another tension that the consortia will need to address. While such extended tasks require the authentic demonstration of students’ ability to apply skills and knowledge, scores on these tasks have proven difficult to equate sufficiently for purposes of longitudinal trends. Several of the authors call for piloting and evaluating such components prior to use through parallel activities. A point of differentiation among the models, however, is in the use of teacher scoring of summative assessments as a professional development activity. The Tucker model calls for professional scoring of summative assessments, whereas the Darling-Hammond/ Pocheone model includes moderated teacher scoring as a core attribute of the design. The Lazer and Resnick/Berger models discuss the benefits of teacher involvement in scoring, but leave it as an option within the summative components. This, then, is another decision that requires clarity in the design priorities of consortia and the trade-offs they are willing to accept.

4. Reporting Improvement Over Time and Closing of Gaps

At the school, district and state level, all groups placed priority on longitudinal assessment data that will reveal whether there is improvement over time and whether long-standing achievement gaps are being reduced. Accurate longitudinal trends require consistency in both the content standards and the performance standards over time. Several of the models described herein include the use of a small number of extended tasks or projects, perhaps given over the course of the school year, within the summative/ accountability system. Herein, then, lies another tension that the consortia will need to address. While such extended tasks require the authentic demonstration of students’ ability to apply skills and knowledge, scores on these tasks have proven difficult to equate sufficiently for purposes of longitudinal trends. Several of the authors call for piloting and evaluating such components prior to use through parallel activities. A point of differentiation among the models, however, is in the use of teacher scoring of summative assessments as a professional development activity. The Tucker model calls for professional scoring of summative assessments, whereas the Darling-Hammond/ Pocheone model includes moderated teacher scoring as a core attribute of the design. The Lazer and Resnick/Berger models discuss the benefits of teacher involvement in scoring, but leave it as an option within the summative components. This, then, is another decision that requires clarity in the design priorities of consortia and the trade-offs they are willing to accept.

MITCHELL CHESTER
Massachusetts Commissioner of Education

it’s been really helpful to get explicit about the aspirations for this next generation of assessments but it’s also going to require, from this point forward, some very serious conversations about the trade-offs, priorities, etc.”
within the summative system. Additional educational measurement work needs to be initiated to provide the science for these next generation assessment systems.

5. Ensuring Our Students are Internationally Competitive

The state, district and national policy respondents were in strong agreement that the most important policy objective at the state and national levels is to place each state’s educational progress into a national and international context, identifying the subject areas or major skills that require greater focus in order to ensure that our students are prepared to be internationally competitive. This will require comparability of scores across states on that portion of summative assessments that assess the Common Core Standards (states may augment the Common Core with up to an additional 15% of State-specific standards).

In order to compare our students’ performance to their peers internationally, consortia will need to either benchmark the performance levels of the new assessments to those of international assessments or to create statistical links between the reporting systems of the new assessments and international studies. Only the Tucker model embeds this attribute of international comparability by utilizing examination systems that are already internationally benchmarked.

Building Next Generation Assessment Systems

As consortia of states put together their wish lists for their new assessments systems, it will be critical to clearly identify the key policy objectives to be achieved. No system – even next generation assessment systems with multiple integrated components – can serve all potential purposes well. It will be equally important to share openly with the public the strengths of the new design as well as its limitations, and to be clear with the public and education constituents about the assumptions inherent in the proposed assessment system.

In addition, as Mitchell Chester, the Commissioner of the Massachusetts Department of Education, pointed out, leaders will need to take time to think through the trade-offs involved in any given design. Many of the desired features involve significant initial and/or ongoing costs, such as the technology infrastructure required for computer-based testing. Teacher involvement in scoring builds understanding and buy-in, but is costly and may stand in tension with the desire for rapid results. The desire for shared and highly instructionally useful assessments requires a willingness to utilize a common curriculum. Comparability of scores across multiple states and/or consortia stands in tension, at least in the near term, with the desire to utilize extended performance tasks or projects.

The four assessment system models described in this Supplement are intended to help states and districts further and deepen their discussions about what matters most to them. They do not cover the entire range of options available to consortia, but do represent four different and comprehensive approaches advocated by nationally respected experts. The full-length papers by these authors, as well as additional conference materials, can be found at www.k12center.org/publications.html. Note that these papers and the approaches they describe have not been subjected to public and formal critique by independent experts. Also note that the approaches are described at a high level, with important details yet to be described and, in many cases, yet to be worked out as they represent next-generation designs. Again, we emphasize that it will be important for consortia to determine how to obtain, within the parameters of local laws and regulations, the advice of independent experts to fully vet the feasibility and cost of such designs.

As the conference panelists noted, it is very rare to have the opportunity to start afresh and to design significantly improved systems to support student achievement. The Common Core Standards and the RTTT Assessment Program provide just such an opportunity for American public education to move ahead forcefully in creating new systems to enhance student learning across our nation.

We now direct you to the accompanying four next generation assessment system models for your consideration.

“...The state, district and national policy respondents were in strong agreement that the most important policy objective at the state and national levels is to place each state’s educational progress into a national and international context, identifying the subject areas or major skills that require greater focus in order to ensure that our students are prepared to be internationally competitive.”
Developing an Internationally Comparable Balanced Assessment System

LINDA DARLING-HAMMOND & RAY PECHEONE
Stanford University
(with Ann Jacquith, Susan Schultz, Leah Walker and Ruth Chung Wei)

This model proposes a balanced assessment system that integrates curriculum and assessments, both formative and summative, and is designed to support higher quality, more coherent instruction. It would be implemented by consortia of states committed to developing curriculum frameworks based on the Common Core Standards and learning progressions. The components of the assessment system— all designed to go beyond recall of facts and show students’ abilities to evaluate evidence, problem solve and understand contexts - include analytic selected-response items, short and extended constructed-response items and standardized performance tasks in each grade level tested. The results of on-demand tests (which include the first two types of items) combined with weighted results of reliably scored curriculum-embedded performance tasks would provide student achievement data that could be compared across schools, districts and states (summative assessments). The system also would be able to show student growth along multiple dimensions. Technology would be used to deliver tests; collect student work for teacher/scorer use and for training scorers; manage the scoring/reporting processes and eventually help teachers to manage classroom practice. Technology would also support computer adaptive testing and computer scoring of some open-ended items. This system, properly implemented, would provide a more performance-based and useful assessment system at potentially no more cost than the present less rich and less helpful systems in use.

Despite standards-based reform efforts in the United States over the past two decades, student achievement in this country continues to slide backwards compared to that of other students on international tests. The Common Core Standards under development by the states could help align expectations for our students with those of high-achieving countries, but how the topics are taught and assessed also must be addressed. Our reliance on multiple-choice items that emphasize recall of discrete facts contrasts with the exams in higher performing nations and states in this country that includes:

- 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.
- Assessments include evidence of actual student performance on challenging tasks that evaluate standards of 21st century learning.
- Teachers are integrally involved in the development of curriculum and the development and scoring of assessments which include both the on-demand portion of state or national examinations and local tasks that feed into examination scores and course grades.
- Assessments are structured to continuously improve teaching and learning (e.g., school-based curriculum embedded assessments and a close examination of student work).
- Assessment systems are designed to emphasize the validity and quality of external assessment aimed at driving high-quality learning of ambitious intellectual skills.
- Assessment and accountability systems use multiple measures to evaluate students and schools.
- Assessment and accountability systems are used primarily for information and improvement.

These principles put considerable emphasis on curriculum as a key lever for two reasons. First, the curriculum expresses the kinds of learning that is sought, such as transferable learning and deep understanding of the central concepts in each discipline studied. Second, carefully constructed curriculum can help teachers understand how learning unfolds, or the typical learning progression within specific domains in a discipline.

This proposed assessment system would include: summative tests that measure student progress and mastery of core concepts with a range of assessment formats (multiple-choice, constructed response, and performance tasks); formative assessments shaped around curriculum and learning progressions; focused professional development; and reporting systems that provide first-hand evidence of student performance along with aggregated scores.

Principles of the Model

The priorities of this assessment system model focus on assessing deep disciplinary understanding and higher-order thinking skills, as well as on using assessment to continually improve instruction and learning. It shares a set of principles with high-performing nations and states in this country that includes:

- 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.
- Assessments include evidence of actual student performance on challenging tasks that evaluate standards of 21st century learning.
- Teachers are integrally involved in the development of curriculum and the development and scoring of assessments which include both the on-demand portion of state or national examinations and local tasks that feed into examination scores and course grades.
- Assessments are structured to continuously improve teaching and learning (e.g., school-based curriculum embedded assessments and a close examination of student work).
- Assessment systems are designed to emphasize the validity and quality of external assessment aimed at driving high-quality learning of ambitious intellectual skills.
- Assessment and accountability systems use multiple measures to evaluate students and schools.
- Assessment and accountability systems are used primarily for information and improvement.

These principles put considerable emphasis on curriculum as a key lever for two reasons. First, the curriculum expresses the kinds of learning that is sought, such as transferable learning and deep understanding of the central concepts in each discipline studied. Second, carefully constructed curriculum can help teachers understand how learning unfolds, or the typical learning progression within specific domains in a discipline.

This proposed assessment system would include: summative tests that measure student progress and mastery of core concepts with a range of assessment formats (multiple-choice, constructed response, and performance tasks); formative assessments shaped around curriculum and learning progressions; focused professional development; and reporting systems that provide first-hand evidence of student performance along with aggregated scores.

Roles for Each Level of Government

The federal government would continue strong investments in education, including: refining the National Assessment of Educational Progress; supporting research on curriculum and assessment design; allowing and encouraging the use of performance assessments under the Elementary and Secondary Education Act; and funding initiatives that infuse knowledge of assessment and learning into pre- and in-service professional development.

Working through consortia, states would create common core standards mapped across grade spans in a set of learning progressions; adopt and
augment the standards as appropriate for a state’s context; create and deploy a curriculum framework that incorporates the standards; build and manage an assessment system of both on-demand and curriculum-embedded assessment components that evaluate the full range of standards and allow evaluation of student progress; create an audit system for ensuring the comparability of teacher managed or scored assessment components; and ensure that the education and development of teachers and leaders infuses knowledge of learning, curriculum and assessment.

Districts and schools would use state curriculum guidance to further refine/revise and continually improve their curriculum; design, select and incorporate formative assessments based on the standards, curriculum and learning progressions; participate in administering and scoring relevant portions of the assessment system and examining student work and outcomes; and engage in the review of assessments and student work within and beyond the school.

How the Assessment System Would Work

Drawing from successful practices already documented in this country and abroad, a consortium of states adopting this blended system would:

- Develop curriculum frameworks based on the Common Core Standards and knowledge of learning progressions, using and expanding upon evidence-based curriculum and assessment efforts.
- Create a digital curriculum and assessment library that would ultimately include materials for curriculum building, model syllabi for specific courses, formative and summative assessment tasks and materials to train both teachers and school leaders on all aspects of the curriculum and assessments.
- Develop state and local assessments. Initially, the consortium would create a common reference examination that includes selected-response, constructed-response and performance components for grades 3-8 and at least one high school grade. The curriculum-embedded components would be developed around core concepts and central skills in English/language arts (E/LA) and math. They would link to the concepts and skills evaluated in the “on-demand” test and would allow the measurement of skills that require more time and student effort than current snapshot tests. Curriculum-embedded tasks used for summative purposes would be standardized, scored in moderated fashion, and the scores aggregated up to count as part of the external assessment. (Over time some of the open-ended items and tasks might also be scored, in part, by computer.) Curriculum-embedded assessments could be used in formative assessments and give teachers useful information for instruction.

All components of the system would incorporate principles of universal design intended to remove construct-related barriers for non-native English speakers and students with other specific learning needs.

The model addresses measurement of student growth along a continuum of performance standards and related learning progressions, not limited by grade level and placed on a vertical scale, in both “on-demand” exams and the more extended classroom assessments.

At the High School Level

The Consortium would explore several options for assessment at the high school level, including:

- Course- or syllabus-based systems like those used in many countries and by the International Baccalaureate (IB) program and the New York Regents system. These would include within-course performance assessments and high-level
end-of-course exams.
- Standards-driven systems with a more comprehensive benchmark assessment in E/LA and math, complemented by collections of evidence that demonstrate students’ abilities to meet certain standards in and across the disciplines. These assessments allow more curricular flexibility in meeting the standards.
- A mixed model that includes both course- and standards-driven models, allowing some demonstration of proficiency in any given course or even outside the boundaries of a course.
- Develop moderation and auditing systems for teacher-scored work. Training would ensure consistent scoring.
- Teachers and school leaders will also need time and training to implement an integrated system of curriculum, instruction and assessment. States and districts must be committed to support teacher engagement in curriculum and assessment development, scoring and analysis. Moreover, states will need to develop technology to support the system.

**Technology Use and Reporting**

A technology platform is key to this assessment model and would be used to:
- Deliver both on-demand and curriculum-embedded assessments to students and teachers.
- Make use of adaptive technologies to deliver tests electronically and create assessments that are responsive to the test taker’s performance.
- Deliver online tasks of higher order abilities.
- Score multiple-choice and some open-ended items and tasks.
- Deliver responses to other tasks that are assessed by trained scorers/teachers using 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.

This model would build an interactive web-based platform that allows robust and efficient 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 and give students and teachers information on assessment performance rapidly. Tasks can be tailored to individual students, training on scoring can be delivered efficiently, and it can be used to audit scoring and generate reports.

**Accessibility and Technical Quality**

With the guidance of a blue-ribbon panel, this assessment system model will help develop the tools, strategies and specific modifications/ accommodations that will allow a wide range of students with specific language or learning needs to be assessed with accessible items and performance tasks/items aligned to college-ready standards.

**Informing Instruction and Leadership**

The ongoing information loop that is built into this blended assessment system will provide the data for continuous learning by students, teachers, school leaders and policymakers. Aligning the assessment system to common standards will enable cross-state and within-group analyses of student content knowledge and foster the capacity to make decisions about professional development, teacher preparation, technology, instructional interventions and the use of school time and resources.

**Goals and Timelines**

The development work will begin with a comprehensive review of Consortium assets, best practices and challenges. A Technical Advisory Committee will be convened to guide the design and development of the balanced assessment system, which will be the focus of the first year of implementation, followed by small scale pilots and any needed modifications in the second year while further design and development work is being completed. Field trials and building capacity for scale up of the assessment model follows in the third year as areas needing further pilot testing are fine-tuned. Readiness for full scale up across the Consortium will occur by year five.

**Costs**

A single dollar figure of test expenditures does not capture the cost-benefit trade-offs such as the benefits of instructionally useful assessments, or, conversely, the narrowing of the curriculum because assessment systems choose what may appear to be the least costly item types.

This paper includes a detailed cost model which used empirical estimates of costs from current testing programs to estimate expected savings from certain kinds of efficiencies. The analysis estimates that a contemporary largescale multiple-choice test battery in a typical state costs about $20 per student. In the same typical state, a high-quality assessment that includes more constructed-response items and an average of two to four additional performance tasks would cost about $55 per student. However, three strategies can yield cost savings: participating in a consortium; using online technology to deliver and score tests; using teacher-scoring of open-ended items (rather than vendor scoring), and using some computer-based scoring of open-ended items. Combining all of these strategies would reduce the cost of high-quality assignments to about $10 to $20 per student, depending on the teacher scoring model used. The authors contend that, with careful planning and attention to the costs of various decisions, it is possible to develop and administer the system they propose at no more than the same cost we are paying for current tests.

**LINDA DARLING-HAMMOND**

is Charles E. Ducommun Professor of Education at Stanford University, where she has launched the Stanford Center for Opportunity Policy in Education and the School Redesign Network and served as faculty sponsor for the Stanford Teacher Education Program. She is a former president of the American Educational Research Association and member of the National Academy of Education. She has been involved in a number of initiatives to develop standards and assessments for both students and teachers, as well as licensing standards and assessments for teachers. She is author of A License to Teach, Authentic Assessment in Action, and Powerful Learning, among other books and articles on these topics. From 1994-2001, Darling-Hammond served as executive director of the National Commission on Teaching and America’s Future, a blue-ribbon panel whose 1996 report, What Matters Most: Teaching for America’s Future, led to sweeping policy changes affecting teaching and teacher education. In 2006, this report was named one of the most influential affecting U.S. education and Darling-Hammond was named one of the nation’s ten most influential people affecting educational policy over the last decade. She recently served as the leader of President Barack Obama’s education policy transition team.

**DR. RAYMOND PCEHONE**

is currently launching a new center, the Stanford Center for Assessment Learning and Equity (SCALe) at Stanford University that focuses on both the development of pre-service licensure portfolio assessments and a performance-based system for student assessment to support the development of the next generation of assessment in the U.S. Over the past decade, Ray was Co-Executive Director of the Stanford School Redesign and Leadership Network (SRN_LEADS). Previously, while serving as Bureau Chief with the Connecticut State Department of Education, Dr. Pcehone oversaw the development and implementation of curriculum frameworks and evaluation programs for teachers and administrators, the Beginning Educator Support and Training Program (BEST). He was a leader in the development of the innovative assessments for the National Board for Professional Teacher Standards (NBPTS) and the Interstate Teacher Assessment and Support Consortium (INTASC). Dr. Pcehone has made numerous conference presentations and published extensively on topics including teacher quality and assessment, as well as student performance assessment and data use.

**FOR MORE INFORMATION**

on this assessment system model, please see the paper by Linda Darling-Hammond and Ray Pcehone at www.k12center.org/publications.html
High-Level Model for an Assessment Of Common Standards

This model proposes an integrated assessment system, not a single test, and focuses on the technical details of a summative system for elementary and middle schools. Using the Common Core Standards, it calls for end-of-year tests for grades 3-8, which could be used to measure student growth if the standards cohere across grade levels. The summative/accountability components of the integrated assessment system might also include periodic classroom tests and collections of student work, which will be made easier if the common standards lead to a common sequence of learning objectives. Summative tests also could be used to inform formative assessments. The end-of-year and periodic components of the assessment should be computer-based. An assessment system for high schools would contain some of the same elements as the K-8 system. However, rather than choosing a specific approach, the paper offers policymakers two models: end-of-domain assessment or end-of-course assessment.

Educators generally agree on the need for improved assessment systems, but there is far less consensus on the priorities for uses of the new assessments. It is impossible for one assessment or even one assessment system to fulfill everyone’s goals, but there are some goals that are shared by various stakeholders. Other goals will require choices; an assessment system cannot do everything equally well.

The author of this paper made certain assumptions in creating this model. Among these are:

- Common core standards will be adopted by a large number of states within 1-2 years.
- The common core standards will cohere across grades so that assessments of the standards will support meaningful estimates of student growth.
- States and/or consortia may want to measure some attributes beyond those covered in the common core standards.
- Universal computer-based testing will be possible in 3-4 years. However, the technology may not be adequate to test all students in mass administrations.
- Major elements of the new assessment system must be made operational within 3-4 years, but the system can and will continue to evolve.

Efficiencies from pooled test development and psychometric work will make modest increases possible in the per pupil operational costs of assessment, allowing some use of human scoring in the system.

The goals of the common core assessment can only be met by an assessment system, not by a single test.

Another assumption is that any assessment system must meet two overarching goals: new instructionally relevant measurement based on common standards; and sound measurement that meets professional technical standards for high-stakes use.

Elements of a Comprehensive K-8 Assessment System

The design calls for an integrated system with formative components and summative/accountability end-of-year assessments, which optimally may also include other components, such as periodic classroom tests and assessments of collected student work. Formative and summative components will work better if they measure the same standards. Moreover, an integrated system that includes formative assessments relieves pressure on summative assessments to provide varied information, especially for classroom-based decisions. An integrated system is versatile and could include interim tests, diagnostic tests and item banks.

Structure of the Summative System

The summative/accountability assessment system will include (but may not be limited to) end-of-year assessments at grades 3-8 in English language arts (ELA) and mathematics. This will produce individual student scores, as well as aggregate scores. Annual testing is recommended to support student growth modeling and to take a snapshot of system progress at a fixed point in time. This would aid in comparability, a major goal of the system. For these reasons, end-of-year testing should be kept, although that does not preclude other data in an accountability system. The paper also discusses the possibility of summative/accountability tests distributed over the course of a year, which can be implemented in places where there is some degree of curricular uniformity.

The end-of-year tests may have at least two major components: common tests of the common core standards; and tests of state-specific content or augmentation. The model allows states to pass up state-customized components, if they wish.

It may be premature to discuss item types before the standards are established, but enough is known about the emerging standards to make some general points. The construct and measurement needs of the system will require a range of exercise types, from selected response to short-answers, to more extended tasks. This mix of item types is likely because of the college-ready expectations that require information on students’ abilities in a variety of areas, such as problem solving and conducting critical analyses. The items and tests should be

Any assessment system is based on choices of which priorities to give precedence. This system ... places a high value on the comparability of data obtained from formal testing events, on measuring a broad range of skills including those that involve new technologies, and on using technology aggressively in an integrated assessment solution.”

Stephen Lazer
developed with an awareness of how students learn. Teaching to the test would be less of a problem if the test reflects learning progressions and models good learning and instruction, while maintaining technical quality.

According to the paper, several issues will need to be addressed during the design phase. Among these questions are: how to deal with the use of audiovisual resources and interactive tasks; how to address the standards for speaking and listening in a summative assessment; how big an item pool is needed to ensure security; and, what should be the length of individual tests at different grade levels?

Computer-Based Assessments
A major question for test designers is: how much technology and how soon? This paper calls for aggressive use of technology in the testing program and argues that standardized assessment components of the system should be computer-based, with traditional paper/pen reserved for special accommodations. There are several reasons for this, including:
- Emerging standards in math and E/LA – and eventually in science – likely will define constructs that only can be measured through the use of technology.
- Technology allows for the use of a range of forward-looking item types such as digital content and formats.
- Testing some skills such as writing on paper may yield invalid results because students are accustomed to doing their work on computers.
- Technology allows for flexible (adaptive) testing and electronic scoring of some items, which will broaden the range of items.
- Technology facilitates more effective dissemination of student responses to teachers and the use of assessment development and scoring for professional development.
- Technology speeds up access to results; makes a broad range of accommodations possible; and if it is the only delivery tool, simplifies issues with comparability.

The summative assessment system and end-of-year tests in particular should make use of adaptive testing administration for several reasons. It allows shorter testing times than linear testing and the use of assessment pools that cover more rigorous standards. It could identify standards which are particularly difficult for students, while also allowing a bigger bang for the buck from open ended/performance-based testing. It also can accommodate extended windows for testing while maintaining high security.

In an adaptive system, the use of items that require human scoring could represent challenges. But there are ways to address them, such as multi-stage testing with machine-scored testing followed later by items requiring human scoring.

The new end-of-year assessment system must be innovative, but at the same time affordable, sustainable and provide rapid scores. We need items beyond the traditional selected-response that push the limits on what can be scored electronically. Scaled scores and status indicators are needed. Full local discretion/customization in the use of formative components. Optionally could include Periodic Assessments and/or Project-Based Components which would begin as low-stakes components until ready for inclusion in summative/accountability system. High school option: either end-of-course or end-of-domain model. System could be operational within 3 to 4 years, but the use of results from the periodic assessments and project-based components may require more time prior to integration into the summative/accountability system unless there is sufficient curricular uniformity.

MODEL HIGHLIGHTS
An integrated summative/formative assessment system built on the Common Core Standards, frameworks and learning progressions. Summative components are standardized for comparability of scores, and built to provide information to the formative systems when additional diagnostics are needed. Full local discretion/customization in the use of formative components. Optionally could include Periodic Assessments and/or Project-Based Components which would begin as low-stakes components until ready for inclusion in summative/accountability system. High school option: either end-of-course or end-of-domain model. System could be operational within 3 to 4 years, but the use of results from the periodic assessments and project-based components may require more time prior to integration into the summative/accountability system unless there is sufficient curricular uniformity.

© 2010 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS).
from an assessment remain essential, but technology must allow us to develop better ways of analyzing data from assessments (e.g., the steps students take in writing essays or engaging in math simulations).

The paper argues that even given advances in electronic scoring, if the assessment is to measure key outcomes human scoring will be necessary for some items, even though it raises issues about affordability. Its added value is the professional development payoff for teachers who do the scoring, provided the system uses the professional development for maximum impact while minimizing the burden.

Measuring Student Progress and Reporting Status Indicators
Given the interest in student growth measures, the assessment system should support cross-grade comparability of scores. Various cross-grade comparability models can be used in the grade 3-8 elements of the system. Measuring growth at the high school level may be more problematic, depending on the high-school model chosen.

If status (proficiency scores) are to be used, it would make more sense from a measurement perspective to apply them to the summative system as a whole and not to specific standards. The content standards must give sufficient guidance on acceptable levels of performance.

Using Periodic Assessments and Project-based Scores in a Summative System
Combined with end-of-year testing, periodic testing should provide much richer information on what students know and can do and should have instructional value. This could be accomplished through single or multiple standardized assessments over the course of a year. Another model is to use data from standardized projects conducted over the course of study (e.g., research papers, lab reports), instead of or in addition to periodic assessments. This would encourage the use of tasks usually omitted from large-scale testing but reflective of good instruction. To the extent to which this would rely on local choice, it raises issues of comparability over time and across jurisdictions, but the problems with both kinds of assessments can be eased if the common standards lead to a common sequence of learning objectives.

The major elements of the K-8 assessment system should be applicable to the high school—an array of item types, a combination of automated and human scoring and delivery through technology. Policymakers, however, have at least two models to choose from: end-of-domain assessment or end-of-course assessment.

The end-of-domain assessment would be given at the point considered right to judge college/career readiness skills mastery (grades 11-12) for ELA and math. These tests would be adaptive in order to allow them both to cover rigorous content and skills and to report meaningful results for all students. This allows for comparisons across schools. If students are testing at different times in their high-school careers (for example, upon entry into high school and then in grade 11), this model could be used to measure student growth. Rapid reporting of scores would not be imperative, and local educators would have maximum flexibility in curriculum decisions (end-of-course models will constrain some choices). On the other hand, because they are not linked to specific courses, they will probably be less effective in providing feedback to teachers or serving as indicators in teacher accountability systems.

The end-of-course assessment model forges direct links between assessment and instruction and could drive instructional improvement. It would include end-of-course tests and could include periodic and project-based assessments and be delivered by computer (if scores are available quickly, they could become part of a course grade). A drawback is that while data are comparable at the student level, aggregation to the school level and higher becomes potentially problematic, given different course taking patterns. End-of-course assessments also create certain issues in measuring student growth data.

Deciding between these high school models will depend on the goals for assessment, and it might be possible to use both models in an assessment system, although this would be expensive. In any case, it is important to have research and validity evidence that supports the intended use of the scores from the assessment. Even if international benchmarks set the base, they need to be updated continually, which will change test specifications. The claim that the tests measure college readiness needs to be validated, and gathering such data should be part of the design from the beginning.

**FOR MORE INFORMATION** on this assessment system model, please see the paper by Stephen Lazer at [www.k12center.org/publications.html](http://www.k12center.org/publications.html)

**STEPHEN LAZER**
As Vice President of Assessment Development, Stephen Lazer is responsible for managing over 500 individuals who develop all assessments and related materials at ETS.

Before taking on this assignment, Lazer served as Group Executive Director of the ETS Government Research & Assessment Services unit. Between 1996 and 2004, he was the Executive Director of NAEP. Lazer joined ETS in 1985, and worked as a test developer in the College Board unit until he moved to the NAEP area in 1991. He has authored a range of publications, including NAEP reports and studies of domestic and international large-scale assessments. He most recently published, with John Mazzeo and Michael Zeicky of ETS, a chapter of the fourth edition of *Educational Measurement*. Lazer has expertise in large-scale assessment design and management and instrument development. He is a member of the American Educational Research Association and the National Council for Measurement in Education.

Lazer earned a graduate degree in political science from Princeton University. He earned his bachelor's degree (a joint honors degree in English and political science) at McGill University. Before working at ETS, Lazer served as a policy analyst for the World Policy Institute, and was a lecturer at Princeton University.
An American Examination System

LAUREN RESNICK
University of Pittsburgh

LARRY BERGER
Wireless Generation
(with Brian Junker, Carnegie Mellon University)

This model focuses on designing an assessment system that works to improve instruction. It reflects key aspects of the substantive, cognitively demanding European systems, while maintaining standards of psychometric rigor necessary to support America’s accountability, comparability, and equity agendas. The model aligns standards, assessment and curriculum through three principal innovations: distributed accountability exams which have a high degree of content and instructional validity; a system of “mass personalized” formative assessments which become an integral part of each teacher’s instructional routine; and a technology platform with broad uses, especially to help teachers manage the assessment process and have ready access to insights from the assessment data. A major aspect of the technology is the creation of a honeycomb, or interactive map, which visually explains the instruction and assessment goals in each grade as well as across grades, tracking the progress of individual students, classes, schools and districts. A substantial amount of work has already been done to develop the content and tools needed to implement The American Examination System.

The Problem

For two decades the United States has been trying to move to a standards-based accountability system, but instead it has created a test-based accountability system that does not reflect the standards we aimed for at the beginning of the 1990s, much less today’s “fewer, clearer, higher” Common Core Standards. Several studies have shown that most state tests do not measure the higher order thinking, problem solving, and creativity needed for students to succeed in the 21st century. These tests, with only a few exceptions, systematically over-represent basic skills and knowledge and under-represent the complex knowledge and reasoning we are seeking for college and career readiness. This has had negative effects on teaching and learning, especially for poor and minority students. The use of interim assessments as a check on progress tends to reinforce these suboptimal teaching behaviors.

This has created a “testing bind” that drives attention away from the intended standards. The effects are greatest in the poorest schools. And it may be lowering the learning opportunities even for many more privileged children as schools turn their energies to the test-based basic skills program.

A Solution

Testing and accountability should remain at the heart of national education policy because equity and national prosperity depend on an education system that keeps stretching toward higher achievement, but there should be new forms of assessment. The proposed American Examination System draws from some of the most successful policies/practices overseas, while maintaining goals important to this country such as equity and comparability. This model:

- Emphasizes the kind of instruction that is valued so that preparing students for tests works for – rather than against – high-cognitive demand instruction.
- Places exams within ongoing instruction so that assessments support instruction rather than distract from it.
- Ensures content and instructional validity of all assessments, avoiding the alignment problem that has plagued state testing programs.
- Provides reliable and valid accountability for student, school and educator performance.
- Includes diagnostic tools for instruction to meet individual student needs.
- Leverages advanced data collection and computational resources for mass personalization of the formative assessments for students and teachers.

By applying the tools of ‘mass personalization’ already prevalent in Internet-based commerce and social networking, we will eventually be able to personalize assessment at the individual level so that the enhanced resolution it provides is targeted to an individual student’s current learning level as well as to appropriate standards of reliability and validity.”

Lauren Resnick & Larry Berger
The American Examination System would be educative for those who use it—teachers, students and education organizations. The system will help teachers, students, and education organizations do their jobs better. It will do this through distributed accountability exams and formative assessments linked to specific content topics and managed through an online technology platform.

Distributed Accountability Exams (DAEs)

Accountability data for this system would come from exams administered throughout the school year after students have completed a unit of study on particular content and skills based on the Common Core of Standards. The exam tasks would be familiar to students because they would be similar to the instruction they received, but neither teachers nor students would know the questions beforehand. Students would take three to five Distributed Accountability Exams (DAEs) in math and literacy during each school year, with each exam assessing material covered through three to seven weeks of instruction. The specific number and timing of exams would be worked out with the states.

The DAEs would model the kind of high-cognitive demand performance embedded in the Common Core Standards, as well as test basic skills. The standards provide a foundation for a criterion-referenced exam system that is closely tied to instruction yet meets technical quality criteria. The core grade-level standards are specified at a granular size that can be used to organize meaningful units of instruction and correspondingly meaningful assessments. Tasks/items for the DAEs would go through rigorous processes, including scientific (experiment-based) tests, to establish content and instructional validity. Availability of multiple forms of the DAEs (to use as pre- and post-tests) would allow states and districts to use the content-based exams to plot student growth, as well as teacher and school effectiveness.

The DAEs would contain a mix of short constructed-response items and more extended written responses (needing human scoring), along with multiple-choice items as appropriate. The exams, given several times over the course of the year, would provide much more data than current end-of-year tests, thereby increasing test reliability.

Educative Formative Assessments

The model would foster the development and use of formative assessments for daily/weekly classroom use that in part mirror the summative assessments. These Educative Formative Assessments would be aligned with the learning trajectories from the Common Core Standards and thus aligned with what teachers need to teach. They would model approaches to how to teach and be part of teachers’ instructional routines rather than imposed as an extra testing assignment. Formative assessments that cannot be machine-scored would have simple rubrics that can be quickly analyzed. Teachers’ scoring forms would be digitalized and include samples of answers for each rubric so teachers can calibrate their analyses.

The formative assessment results should not be used in accountability reporting, but the student, class, and school results would be available to teachers and principals for use in classrooms and for professional development. Metrics of fidelity in implementing the formative assessments (and their associated instructional recommendations) could be used as part of teacher/school performance management/accountability. For

MODEL HIGHLIGHTS

An integrated system of standards, assessments (summative and formative) and curriculum, built upon the Common Core Standards. A technology platform manages all student data and provides an interactive map (the honeycomb) of academic progress. The summative/accountability assessments consist of three to five Distributed (periodic) Accountability Exams (DAEs) taken at the end of units of instruction, and scores on the DAEs are aggregated for accountability purposes. System includes high-speed scanning and distributed human scoring system for DAEs, with 24-hour turnaround stated as possible. Formative assessments include machine-scored and may include teacher-scored items, rubrics provided. Technology platform integrates formative and summative student work and data to create student learning profiles and support mass personalization. Model could be operational within 4 years.
instance, formative data can show whether teachers are doing progress monitoring with the frequency appropriate for each student. (DC Public Schools is an example of a school system that is already using these types of formative assessment metrics as part of their “SchoolStat” approach to continuous, district-wide performance management.)

Managing Mass Personalization

An advanced model of measurement can be built on great magnitude of data – both formal and informal – on each student in the course of the year so that each test enhances a picture already drawn of the student. This mass personalization, already used in internet-based commerce and social networking, eventually will be able to personalize each assessment at the individual student level. Attributes that could be the basis of personalization include past student performance on assessments, teacher and school characteristics, aggregated assessment performance of students in a school, previous effectiveness of the teacher and the specific curriculum and assessments used.

Technology makes this scalable – there is no limit to the amount of data computers can collect. The initial goal for mass personalization would be to apply it to customizing formative assessments used as part of classroom instruction.

The Assessment Platform

The assessment platform would manage both the DAEs and the formative assessment system, enabling assessment delivery, scoring, reporting and analysis. Based on widespread classroom experience with existing products and on current designs (some of which have been funded by the Gates Foundation), it would be able to handle all of these elements at scale in a cost-effective way, while minimizing additional burdens for teachers, students and administrators.

The model provides a honeycomb or an interactive map of learning trajectories. It shows the instruction and assessment that should take place across all grades, collecting data on the individual progress of students, classes of students, schools and school districts. It also would provide data to validate/refine hypotheses about the skills in the Common Core of Standards and state standards.

When the system is fully operative, educators will have formative and summative assessments for each skill step along each learning trajectory, starting with math and literacy for grades 3-10. They will be able to see exam and formative assessment results for each student and how the results match with the trajectory.

The assessment platform will: enable students to take the assessments on line or on paper; support teachers/schools in scanning and uploading paper-based assessments and other student work; manage remote scoring workflow; provide teachers with a scoring interface and dashboard tools for tracking and analyzing the progress of particular students or groups of students; provide administrators with a reporting interface that includes aggregate analysis; and allow teachers to share formative assessments with each other. The platform also will include an assignment builder, so that educators can select formative assessment items as tasks for use by the students in the classroom or as homework.

The American Examination System would not assume, at the outset, that all assessments will be conducted with students sitting at computers. Given current school infrastructures and the challenge of showing math work via keyboards, it may be more efficient to continue to rely for some time on paper-and-pencil inputs to an otherwise digital system. The continued value of these “primitive” recording tools seems especially compelling when one considers that much of the value of the new generation of assessment tasks depends on soliciting open-ended expressions of student reasoning and thinking—and in the case of math this includes drawings, graphs and explanations. So the American Examination System would include a process to enable scanning/digital photographing, uploading and archiving of very large volumes of paper-based student work, including for Distributed Accountability Exams, to enable remote scoring as well as online student portfolios. The scanning/photographing process, which has already been tested in North Carolina classrooms, puts minimal burdens on teachers or other school staff and does not require large per-school investments in hardware or network infrastructure.

Development Timelines

This assessment model is feasible because much of the work in developing the content and tools needed already has been done. The system can be fully operational within three to four years from the beginning of the process, with mass personalization of summative assessment playing a larger role at the end of that timeframe.

Costs

We estimate that a typical state would spend the same for this American Examination System as it now spends on assessments for NCLB, or $20-30 per student. The DAEs will cost more to administer than current high-stakes tests because they will be given more frequently. On the other hand, states will be able to eliminate the costs of current interim testing.

FOR MORE INFORMATION on this assessment system model, please see the paper by Lauren Resnick and Larry Berger at www.k12center.org/publications.html

LAUREN B. RESNICK is an internationally known scholar in the cognitive science of learning and instruction. She served as the director of the prestigious Learning Research and Development Center from 1977 until 2008, at the University of Pittsburgh, where she is a Distinguished University Professor in the School of Education. Dr. Resnickind is founder and director of the Institute for Learning and co-founded the New Standards Project (1990-1999), which developed performance-based standards and assessments that widely influenced state and school district practice. Dr. Resnick was the founding editor of both Cognition and Instruction and Research Points, was advisor to the first chairman of the National Education Goals Panel, and was a member of the National Council on Education Standards and Testing, the governing body that launched the nation’s standards policy. More recently, Dr. Resnick has been Chair of the National Academy of Education White Paper Project, an initiative to connect policymakers in the current Presidential administration and Congress with the best available evidence on selected education policy issues. Dr. Resnick has received multiple awards for her research, including the 2007 Award for Distinguished Contributions in Applications of Psychology to Education and Training and the 1998 E. L. Thorndike Award, both from the American Psychological Association.

LARRY BERGER is CEO and co-founder of Wireless Generation, a company that helps PreK-12 educators to teach smarter through the sensitive and innovative application of technology in the classroom. The company has developed software for mobile devices that makes formative assessment instructionally useful to teachers, “next generation” curriculum customized throughout the school year to students’ needs, and large scale data systems that centralize student information and integrate knowledge management tools to spur teacher collaborations. Berger was a Rhodes Scholar and a White House Fellow working on educational technology at NASA. He serves on the Carnegie Institute for Advanced Study Joint Commission on Mathematics and Science Education, and on the Board of Trustees for the Carnegie Foundation for the Advancement of Teaching. He is a member of the Board of Overseers for the Annenberg Institute on School Reform at Brown University and serves on the Board of Editorial Projects in Education Inc., publisher of Education Week.
An Assessment System for the United States: Why Not Build on the Best

MARC TUCKER
National Center on Education and the Economy (NCEE)

This paper describes the model that underlies NCEE’s State Consortium for Board Examination Systems. Based on exams in countries with the best-performing education systems, it consists of high school programs of study covering the whole core curriculum. Each course has a well-constructed syllabus with matching instructional materials, high quality examinations and high quality training for the teachers of the courses. NCEE has identified the world’s best board examination systems available in English for use in the United States and negotiated the alignment with the Common Core Standards. The states in the consortium will pilot the use of these exams in their high schools. NCEE will set the pass points for the lower division exams to the level of literacy required to be successful in the initial credit-bearing courses in open admissions colleges. Students who pass their lower division exams by the end of their sophomore year will certify for a special diploma and enrollment in an open admissions college without having to take remedial courses. Or they can stay in high school to take an upper division program designed to get them into a selective college. Students who do not pass will receive help in those areas in which they did not do well on the exam. The aim is for virtually all students to be ready to pass their exams and therefore ready to succeed in programs leading to industry qualifications or in four-year colleges.

The Goals for a Testing System

For more than two decades the National Center on Education and the Economy (NCEE) has studied intensely the education systems of those countries that consistently demonstrate the best performance on international tests. The research has produced benchmarks of their education standards, instructional systems and assessments – what they do that the United States could adopt or adapt. Based on the benchmarking research, a testing system should have the following qualities:

Assessments that are conceived as part of a highly integrated instructional or learning system. The countries with the best student achievement require a core curriculum at the high school level with a well-constructed syllabus for each course, accompanied by matching instructional materials and a high-quality exam based on the syllabus. Professional development is tailored to the syllabus.

Assessments that make reasonable accommodations for the disabled. New research on the use of advanced technologies with this population should be incorporated in any assessment design.

Assessments that are balanced, in the sense that they can support instruction as it is taking place, by providing immediate, targeted feedback on student performance, as well as provide summative information to a variety of audiences, when a student has completed a course or a program of study. In addition, summative tests could combine results of a timed test with assessments of a portfolio of student work performed during the school year. This would return classroom-based assessment to its former status.

Assessments that cannot be test-prepped. Good assessments make the invidious practice of “test-prep” unrewarding because they give students questions that do not look like the ones they practiced on, but which call for the understanding and skill they should have mastered.

Assessments that are reliable and valid. The United States has emphasized reliability (the scoring is technically correct and consistent) and downplayed tests that accurately measure the more complex Standards (validity). Other countries have obtained reliability through professionally-scored exam systems, combined with collections of student work that are scored by teachers and checked by professional scorers.

Assessments which, when combined with the accountability system, do not produce a very narrow curriculum. The consequence of the emphasis on basic skills under the No Child Left Behind Act has been to narrow the core curriculum for all students to math and English/language arts. A new national assessment system must provide incentives to states and schools to teach the whole core curriculum.

Assessments that are affordable. The countries with the best student achievement use multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that promote and measure critical thinking, strong analysis and real imagination and creativity. No high-achieving country uses multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that are balanced, in the sense that they can support instruction as it is taking place, by providing immediate, targeted feedback on student performance, as well as provide summative information to a variety of audiences, when a student has completed a course or a program of study. In addition, summative tests could combine results of a timed test with assessments of a portfolio of student work performed during the school year. This would return classroom-based assessment to its former status.

Assessments that are reliable and valid. The United States has emphasized reliability (the scoring is technically correct and consistent) and downplayed tests that accurately measure the more complex Standards (validity). Other countries have obtained reliability through professionally-scored exam systems, combined with collections of student work that are scored by teachers and checked by professional scorers.

Assessments which, when combined with the accountability system, do not produce a very narrow curriculum. The consequence of the emphasis on basic skills under the No Child Left Behind Act has been to narrow the core curriculum for all students to math and English/language arts. A new national assessment system must provide incentives to states and schools to teach the whole core curriculum.

Assessments that are affordable. The countries with the best student achievement use multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that promote and measure critical thinking, strong analysis and real imagination and creativity. No high-achieving country uses multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that are balanced, in the sense that they can support instruction as it is taking place, by providing immediate, targeted feedback on student performance, as well as provide summative information to a variety of audiences, when a student has completed a course or a program of study. In addition, summative tests could combine results of a timed test with assessments of a portfolio of student work performed during the school year. This would return classroom-based assessment to its former status.

Assessments that are reliable and valid. The United States has emphasized reliability (the scoring is technically correct and consistent) and downplayed tests that accurately measure the more complex Standards (validity). Other countries have obtained reliability through professionally-scored exam systems, combined with collections of student work that are scored by teachers and checked by professional scorers.

Assessments which, when combined with the accountability system, do not produce a very narrow curriculum. The consequence of the emphasis on basic skills under the No Child Left Behind Act has been to narrow the core curriculum for all students to math and English/language arts. A new national assessment system must provide incentives to states and schools to teach the whole core curriculum.

Assessments that are affordable. The countries with the best student achievement use multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that promote and measure critical thinking, strong analysis and real imagination and creativity. No high-achieving country uses multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that are balanced, in the sense that they can support instruction as it is taking place, by providing immediate, targeted feedback on student performance, as well as provide summative information to a variety of audiences, when a student has completed a course or a program of study. In addition, summative tests could combine results of a timed test with assessments of a portfolio of student work performed during the school year. This would return classroom-based assessment to its former status.

Assessments that are reliable and valid. The United States has emphasized reliability (the scoring is technically correct and consistent) and downplayed tests that accurately measure the more complex Standards (validity). Other countries have obtained reliability through professionally-scored exam systems, combined with collections of student work that are scored by teachers and checked by professional scorers.

Assessments which, when combined with the accountability system, do not produce a very narrow curriculum. The consequence of the emphasis on basic skills under the No Child Left Behind Act has been to narrow the core curriculum for all students to math and English/language arts. A new national assessment system must provide incentives to states and schools to teach the whole core curriculum.

Assessments that are affordable. The countries with the best student achievement use multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that promote and measure critical thinking, strong analysis and real imagination and creativity. No high-achieving country uses multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that are balanced, in the sense that they can support instruction as it is taking place, by providing immediate, targeted feedback on student performance, as well as provide summative information to a variety of audiences, when a student has completed a course or a program of study. In addition, summative tests could combine results of a timed test with assessments of a portfolio of student work performed during the school year. This would return classroom-based assessment to its former status.

Assessments that are reliable and valid. The United States has emphasized reliability (the scoring is technically correct and consistent) and downplayed tests that accurately measure the more complex Standards (validity). Other countries have obtained reliability through professionally-scored exam systems, combined with collections of student work that are scored by teachers and checked by professional scorers.

Assessments which, when combined with the accountability system, do not produce a very narrow curriculum. The consequence of the emphasis on basic skills under the No Child Left Behind Act has been to narrow the core curriculum for all students to math and English/language arts. A new national assessment system must provide incentives to states and schools to teach the whole core curriculum.

Assessments that are affordable. The countries with the best student achievement use multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that promote and measure critical thinking, strong analysis and real imagination and creativity. No high-achieving country uses multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that are balanced, in the sense that they can support instruction as it is taking place, by providing immediate, targeted feedback on student performance, as well as provide summative information to a variety of audiences, when a student has completed a course or a program of study. In addition, summative tests could combine results of a timed test with assessments of a portfolio of student work performed during the school year. This would return classroom-based assessment to its former status.

Assessments that are reliable and valid. The United States has emphasized reliability (the scoring is technically correct and consistent) and downplayed tests that accurately measure the more complex Standards (validity). Other countries have obtained reliability through professionally-scored exam systems, combined with collections of student work that are scored by teachers and checked by professional scorers.

Assessments which, when combined with the accountability system, do not produce a very narrow curriculum. The consequence of the emphasis on basic skills under the No Child Left Behind Act has been to narrow the core curriculum for all students to math and English/language arts. A new national assessment system must provide incentives to states and schools to teach the whole core curriculum.

Assessments that are affordable. The countries with the best student achievement use multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.

Assessments that promote and measure critical thinking, strong analysis and real imagination and creativity. No high-achieving country uses multiple-choice, machine-scored tests anywhere near as much as the United States. Such tests cannot measure skills/attributes that are important to student success.
Superior assessments do cost more than the United States spends on per pupil assessments, but this consortium offers a strategy for offsetting this increase with other cost reductions.

**Getting Started on a High School Exam System**

NCEE has established a State Consortium for Board Examination Systems. Instead of building an assessment system with the attributes just described from scratch, consortium members would start by adopting the one or more of the world’s best instruction and examination systems, modified as necessary to reflect the Common Core Standards. These systems meet all the criteria just described. They are curriculum-based; offer complete instructional systems; and provide syllabi with matching materials, exams, professional scoring and teacher training. They address both basic skills and the full range of desired knowledge and skill such as critical thinking and analysis. They include tools for formative assessments as well as summative assessments. They make clear to students the level of performance needed for them to move on. To get started:

- States would require their high schools to adopt at least one of the instructional/exam systems that NCEE has identified for use in the first two grades of high school (QualityCore program from ACT; the International General Certificate of Secondary Education from the University of Cambridge International Examinations; and the International General Certificate of Secondary Education from Pearson/Edexcel). All of the programs would be set to a single pass point calibrated to expectations in credit-bearing courses of the first year of open-admissions colleges.
- Students who pass the exams at the end of their sophomore year could enroll immediately at open admission 2- or 4-year campuses or remain in high school for studies designed to help them gain admission to competitive colleges/universities. High schools would provide customized education for those students who did not make the standard for the exam, and the students could retake the exams as often as they wish. This is a move-on-when-ready system.
- The experience of other countries indicates that this system would be immediately motivating to high school students. It would encourage them to take tough courses and study hard so they could leave early, and the money saved in the upper grades could be used to help students who fail the exams.
- All of the organizations that provide the lower division exams have agreed to modify them as necessary to comply with the Common Core Standards and to adjust any language/spelling differences. With these modifications, the United States could adopt the world’s best instructional systems and use the best assessments at a fraction of the cost of developing them from scratch and have them available much sooner.

**A High School Assessment System That Improves on the World’s Best**

Once these standard-setting instructional systems are in place, they can be improved upon, largely through the use of advanced technology. Computer-based systems using simulations and other dynamic models can require students to demonstrate higher order skills. As this field is developed, grading of certain qualities can become more precise than now, and the costs held down. Advances in instructional technology could enable almost all

---

**MODEL HIGHLIGHTS**

Based on the adaptation and adoption of existing world-class instructional systems, aligned to the Common Core Standards, that include course syllabi, instructional materials, examinations and teacher training. Schools and districts would select from a menu of approved curricula, and give the accompanying examinations; scores from the various exams will be placed on a common scale to ensure comparability of scores and expectations. All examinations include multiple item types, and higher percentages of extended response items and tasks, and all summative assessments are to be either machine scored or professionally scored, with results within 10 days. Model calls for major advances in technologies for instruction, assessment and scoring. This model does not include annual summative/accountability testing in grades 3 through 8, but end-of-cluster assessments at grades 3, 5 and 8.
The K-8 Assessment System

NCEE’s consortium will not offer assessments in the K-8 arena, but NCEE’s international benchmarking research provides a K-8 assessment design. In almost all high-performing countries, there are no stakes or very low stakes for K-8 teachers, although in some countries student assessment results are used as a signal to authorities that a school needs a visit by school inspectors. National tests or exams usually come at the end of grade bands and are mainly scored by the students’ teachers (teacher scoring is often considered an important part of professional development). Most assessment is done with prompts calling for essay-type responses. In contrast, the United States is almost alone in developing a testing system designed to be insensitive to the curriculum students have taken.

The proposed K-8 assessment system would divide the grades into bands at important junctions. Students would be given a diagnostic assessment on entering Kindergarten and then be given summative assessments at the end of grades 3, 5, and 8. The assessments would reflect a natural learning progression in the curriculum. The summative assessments would include computer-scored, multiple-choice items, though fewer than is now typical, as well as more and longer constructed-response items. Similar items would also be available for use by teachers during the year. Scores on extended assignments would be included in the final grade for the year, along with the score on the final exams.

The learning progression for each major subject in the K-8 system curriculum would be used to construct a curriculum framework for that subject, specifying which topics are to be taught, grade by grade. Publishers of texts and other instructional materials would use these frameworks to guide the production of instructional materials. This means that tests would measure a smaller number of topics for each grade, but at greater depth. This approach could adopt the British system for pre-secondary exams or draw on the best in this country such as the New England Common Assessment Program. The data from the high school design and K-8 system would enable the public to hold educators responsible for the proportion of students who are college-ready, subject by subject and grade by grade in high school, and at the grade bands for the lower schools. While it cannot provide accurate measures of the contribution of each teacher to his/her students’ growth, the proposed system can provide growth measures for a school as a whole. The state would establish strong incentives for school faculties to work hard to get their students college ready. Similarly, data indicating that schools were not doing a good job of preparing their students for college and work would trigger visits by school inspectors to determine the causes and take appropriate measures.

Governing and Regulating the System

NCEE’s consortium of states interested in implementing the high school design has created a technical advisory committee to help design the necessary research needed to array the selected exams along a common scale and to ascertain the level of literacy needed to succeed in first-year, credit-bearing college courses. The consortium will have a governing board to make essential decisions.

This model, in sum, allows swift implementation of the world’s best instruction and assessment systems, as well as the tools they use to raise student performance, without the development cost. With the addition of technology know-how, the United States could even get ahead of the pack. It can afford to implement this system, despite the hard economic times, by combining the proposed high school exam system with the proposed move-on-when-ready system, using the money saved to boost the chances of all students to meet the performance standards.

MARC TUCKER is the President and Chief Executive Officer of the National Center on Education and the Economy. Mr. Tucker is an internationally recognized expert in academic and occupational standards and assessment. In addition to NCEE, Tucker has been instrumental in creating the National Board for Professional Teaching Standards, the Commission on the Skills of the American Workforce, the New Commission on the Skills of the American Workforce, the New Standards Consortium, the National Skills Standards Board, America’s Choice and the National Institute for School Leadership. He has also served as author, co-author or editor of many articles and several books and reports, including, America’s Choice: high skills or low wages!, Standards for Our Schools: How to Set Them, Measure Them and Reach Them; Thinking for a Living: Education and the Wealth of Nations; The Principal Challenge; and Tough Choices or Tough Times. Mr. Tucker has testified frequently to the U.S. Congress and state legislatures.
Educational Urgency in a Flat World

The intellectual commodity of the future will be individuals’ imagination, Thomas Friedman, a Pulitzer prize-winner and columnist for the New York Times, told the opening session of the conference on next-generation assessments. As author of the best-seller The World Is Flat, he traced the evolution of the nexus of globalization from nations in the 1900s, to corporations in the 20th century, and now to individuals. Because of technology, Friedman observed, everyone can now be a journalist, a film maker, a blogger or an entrepreneur, and can connect with the content and ideas produced by others at virtually no cost.

This shift from silos of information with limited access to a global flow accessible to anyone with access to the Web, he argued, is the biggest event to change how human beings interact since Gutenberg invented the printing press in 1450. “The global economic playing field is being leveled,” Freidman was warned in an interview with a prominent Indian entrepreneur, “and you, Americans, are not ready!”

The most important competition edge today is one’s imagination, because Friedman observed, “what you can imagine now as an individual, you can now act on farther, faster, deeper, [and] cheaper than ever before.” The successful countries, therefore, will be the ones that actively nurture their people’s ability to imagine and innovate – their students’ ability to put together ideas from disparate disciplines and, snap! create new ideas. This, Friedman argued, requires a strong liberal arts curriculum that includes not only advanced math and language arts skills, but also history, science, languages and the arts. Inspiration and new advances are to be found in the intersections of the disciplines.

Technological advances also mean, however, that people’s lives are creating “digitalized” footprints. Such transparency will make character – having sustainable values when conducting business or work rather than merely situational values – more important than ever before. This also applies to schools and learning, according to Friedman. As educators, we have an obligation to help children develop strong character, as their “hows” will matter more than ever in the world of twittering, blogging and YouTube. Educators, as well as parents, will need to help students build and understand the need for strong character so their “hows” will not harm their future opportunities.

Mr. Friedman also shared with the audience his deep commitment to public education. His wife is an elementary reading teacher in a public school in Maryland, and his daughter teaches in Washington DC Public Schools with the Teach for America program.

THOMAS L. FRIEDMAN is a world-renowned author and three-time Pulitzer Prize winner whose foreign affairs column for The New York Times is syndicated to one hundred other newspapers worldwide.

The Future of Assessment and Lessons Learned Internationally

s skill demands and educational attainment increase around the world, old measurement systems no longer are adequate, Andreas Schleicher, head of research for the OECD Directorate of Education, advised in a luncheon presentation at the National Conference on Next Generation Assessment Systems. One of the reasons policymakers are looking for new assessment designs is because “the yardstick for success is no longer just improvement by national standards, but is the best performing education systems globally,” he said.

For more than a decade now, the OECD has been conducting a survey assessment of the knowledge and skills of 15-year-olds involving more than 60 countries. The trends in performance over time reveal strongly differing rates of educational improvement across nations, with several countries moving rapidly to the front of the pack and surpassing the United States. Further analysis by OECD has shown, explained Schleicher, that growth in national economies is not simply a matter of the number of years of schooling, but is also clearly related to the quality of education. That is why, in his view, it is so important to develop assessments that tell us something about the quality of learning occurring in our schools.

Schleicher described two dominant approaches to the purpose of summative assessments internationally. Most northern European countries utilize their assessments for diagnosis and improvement purposes. Results are not used for school or educator accountability. Because of this, they typically assess a sample of schools and students and utilize more complex tasks. England, Latin America and United States, on the other hand, utilize large scale assessments primarily for accountability purposes, which require the assessment of every student in a highly reliable, cost-efficient manner. Despite these differences, however, Schleicher noted several global trends in assessment designs.

First, they are multi-layered, coherent systems that emphasize not only basic subject matter content, but also require the capacity to integrate, synthesize and creatively apply knowledge to novel situations. They incorporate principles of learning and cognition, make students’ thinking visible, allow for divergent thinking and are responsive to new developments. They also add value to teaching and learning because they produce information that can be acted upon by students, teachers and administrators. Finally, these new assessments are part of comprehensive systems that include instructional materials and teacher training.

The first issue to get right when designing next generation systems, according to Schleicher, is the balance between the focus on outcomes/ accountability and the focus on tools and processes that enable improvement. He further elaborated a set of criteria to be used in developing new systems, found in his presentation at www.k12center.org/publications.html. The Race to the Top Assessment Program provides a terrific opportunity to design comprehensive, integrated systems to support both purposes well, and to design these systems themselves for continuous improvement.

ANDREAS SCHLEICHER is Special Advisor on Education Policy to the Secretary General of the Organisation for Economic Co-Operation and Development (OECD) and is responsible for the development and analysis of benchmarks on the performance of education systems internationally and the impact of knowledge and skills on economic and social outcomes.
Comparison of Four Models for Next Generation Assessment Systems

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RIGOROUS STANDARDS &amp; GOOD INSTRUCTIONAL PRACTICES</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Assessments will be designed to reflect learning progressions embedded in the Common Core Standards.</td>
<td>- Integrated system of formative and summative components</td>
<td>- Summative exams distributed over the year – each following the relevant curriculum unit.</td>
<td>- Builds on “world’s best” instructional systems currently available in English, modified as needed to reflect the Common Core Standards.</td>
</tr>
<tr>
<td>- Summative assessments include a ‘relatively lean’ end-of-year assessment with analytic multiple-choice and constructed-response items, and a small number of curriculum-embedded performance tasks.</td>
<td>- Use of “evidence-centered design” (ECD) processes to ensure alignment with Common Core Standards.</td>
<td>- Formative assessments designed to model how to teach and enable teachers to learn to use these methods in instruction.</td>
<td>- These existing systems include course syllabi, materials, exams, and teacher training.</td>
</tr>
<tr>
<td>- On-going formative and interim assessments mapped to curriculum frameworks and standards for use across participating states.</td>
<td>- Use of mixed item types (selected response, short answer, and extended answer)</td>
<td>- Exams include high-cognitive-demand tasks and tests of basic procedural skills.</td>
<td>- Aligns Lower-Division systems (typically 4th and 10th grade) with college- and career-readiness.</td>
</tr>
<tr>
<td><strong>USE OF TECHNOLOGY</strong></td>
<td>- Computer-based testing and flexible (adaptive) administration.</td>
<td>- Content and sequence of exams based on Common Core Standards learning progression.</td>
<td>- Move on when ready” design restructures high school to allow students to choose, after Lower Division, either an open-enrollment 2- or 4-year college or additional coursework to prepare for selective college.</td>
</tr>
<tr>
<td>- Goal is computer-delivered assessment with digitized student responses for both on-demand and curriculum-embedded assessment components.</td>
<td>- Electronic scoring of some items in conjunction with teachers and/or other human scorers.</td>
<td>- Scientific (experiment-based) validation process to ensure that exams measure the Common Core Standards.</td>
<td></td>
</tr>
<tr>
<td>- Adaptive delivery for on-demand components.</td>
<td>- Open architecture and standards to facilitate transfer of materials and state/curriculum customization.</td>
<td>- All of these selected “world’s best” systems:</td>
<td></td>
</tr>
<tr>
<td>- Web-based interface for students, allowing for retrieval, uploading, and storing of work products.</td>
<td>- Integrated data-management system to facilitate periodic assessments and project-based components.</td>
<td>+ use technology to capture student work and student answers to exam questions</td>
<td></td>
</tr>
<tr>
<td>- Web-based interface for teachers to train and certify for scoring, facilitate scoring and auditing, and support instructional practice (through access to curriculum materials and formative assessments)</td>
<td>- Technology to support a range of accommodations for students with disabilities and English-language learners.</td>
<td>+ use technology to ship data to scorers, do quality control on scoring, analyze student scores, and present data.</td>
<td></td>
</tr>
<tr>
<td>- Electronic system management and reporting, including formative and summative data for classroom, school, and district information, and summative scores for state and cross-state reporting.</td>
<td>- Use of computer-based testing to assess technological skills as reflected in the new Standards.</td>
<td>+ offer web-based access to curriculum designs, course syllabi, lesson plans, and instructional materials.</td>
<td></td>
</tr>
<tr>
<td><strong>MEASURING GROWTH &amp; PROJECTING READINESS</strong></td>
<td>- Annual testing between grades 3 - 8 to support data-driven decision making.</td>
<td>- Use of equivalent pre- and post- criterion-referenced tests to enable measures of student growth and of teacher/school effectiveness.</td>
<td>- Technology provides access to world-wide interactive networks of teachers teaching the same courses, access to experts, and formal training via the web.</td>
</tr>
<tr>
<td>- Evaluates growth along well-defined learning progressions assessed using multiple assessment formats.</td>
<td>- Cross-grade comparability of scores for grades 3-8 facilitated by coherent standards and expectations across grades.</td>
<td>- Common Core Standards organized as a set of learning trajectories displayed in “Honeycomb” as a way for teachers, students and parents to track progress in achieving these standards.</td>
<td>- Realizes some savings into aggressive development of technologies for learning, assessing and reporting.</td>
</tr>
<tr>
<td>- Evaluates students’ movement along a vertically aligned continuum using, to the extent practicable, computer adaptive assessment techniques.</td>
<td>- Growth scores likely appropriate for grades 3-8 but may be problematic at high school.</td>
<td>- Allows identification of specific skill deficits, and whether students are on track toward meeting academic skills and career readiness goals.</td>
<td></td>
</tr>
<tr>
<td><strong>REPORTING RESULTS</strong></td>
<td>- Broad range of item types and difficulties built into adaptive pool, enabling system to be used for measuring growth while still measuring attainment of high standards.</td>
<td>- Technology platform delivers student work to scorers within seconds. If professional scorers are used, scoring can be done within 24 hours. If teachers do scoring, turnaround time depends on the state/district expectations set for them.</td>
<td>- Uses end-of-course test with pass points set at college level readiness.</td>
</tr>
<tr>
<td>- Uses inter-operable electronic platform to provide scores within weeks for both performance measures and on-demand standardized assessments.</td>
<td></td>
<td>- Student growth reflected in student grades on internationally benchmarked courses in the core subjects in the curriculum and in progress toward pass points defined in terms of likelihood of success in initial credit-bearing courses in open-admissions colleges.</td>
<td></td>
</tr>
<tr>
<td>- Reporting functions can include student, class, school, district, state, parent, and community summaries.</td>
<td>- Use of equivalent pre- and post- criterion-referenced tests to enable measures of student growth and of teacher/school effectiveness.</td>
<td>- ACT QualityCore scores by course reported within 2 weeks and delivered online. State to receive additional statewide analysis report.</td>
<td></td>
</tr>
<tr>
<td>- Includes release of items and tasks to students, teachers, administrators, and parents for greater understanding of expectations.</td>
<td>- Common Core Standards organized as a set of learning trajectories displayed in “Honeycomb” as a way for teachers, students and parents to track progress in achieving these standards.</td>
<td>- University of Cambridge International Examinations promises to issue online student and school reports within 10 business days.</td>
<td></td>
</tr>
<tr>
<td><strong>INFORMING INSTRUCTION &amp; LEADERSHIP</strong></td>
<td>- Comparability of scores and sound measurement practices are high priorities of the model.</td>
<td>- Allows identification of specific skill deficits, and whether students are on track toward meeting academic skills and career readiness goals.</td>
<td>- Pearson/Edexcel promises to issue online reports within 10 business days. Reports include individual scores tied to skill maps and delineated by item, in addition to school- wide, course, cohort, gender, and national comparisons.</td>
</tr>
<tr>
<td>- Release of a significant number of items and tasks to students, teachers, administrators, and parents intended to provide a concrete understanding of how standards are reflected in assessments, as well as feedback for improved teaching and learning.</td>
<td>- Goal is to provide fastest possible reporting turnaround given the need for human scoring. (score turn-around will not be immediate, especially in first year).</td>
<td>- ACT QualityCore scores by course reported within 2 weeks and delivered online. State to receive additional statewide analysis report.</td>
<td></td>
</tr>
<tr>
<td>- Capacity to show evidence of students’ learning progress over time—both during the school year and across years.</td>
<td>- Advocates maximum use of electronic scoring and, for items that cannot yet be scored, use of distributed scoring to speed results.</td>
<td>- University of Cambridge International Examinations promises to issue online student and school reports within 10 business days.</td>
<td></td>
</tr>
<tr>
<td>- Strategic use of teachers in the item development and scoring processes viewed as an important professional development tool.</td>
<td>- System should allow for all appropriate levels of data aggregation.</td>
<td>- System allows teachers, principals, and districts to generate custom reports in real time on demand.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Technology platform delivers student work to scorers within seconds. If professional scorers are used, scoring can be done within 24 hours. If teachers do scoring, turnaround time depends on the state/district expectations set for them.</td>
<td>- The summative and formative assessments are intended to provide actionable diagnostic data for teachers, and to model quality instruction.</td>
<td>- Board examinations are complete instructional systems that allow internalization of standards by providing standards narratives. Library of previous questions, and examples of top student work Board systems provide formative assessment support.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Reports how students are progressing along the learning continuum to meet Common Core state standards.</td>
<td>- By providing robust curriculum support, examinations retain validity through “opportunity to learn” principles.</td>
</tr>
</tbody>
</table>
An Assessment Agenda

State initiatives to use common core standards and develop the next generation of common assessments as part of the federally funded RTTT Assessment Program open up an unprecedented opportunity for U.S. public education to move ahead forcefully. That is not all that is happening, however. Much more is being asked of the uses of assessment, at the same time that research on student learning progressions and the application of new technologies are expanding opportunities for the development of more useful tools and systems.

In this document we share descriptions of four potential designs for next generation assessment systems. Each design has the potential to move us forward, and each incorporates a particular set of decisions about the role of each level of the educational system, the features that are essential, and the trade-offs that are acceptable.

Moreover, each paper points out the opportunity and need for significant advances in several areas of measurement and assessment design, including:

- Development of evidence-based learning progressions, and how they can be used to inform, measure and predict student learning;
- Development of more valid and fair content assessments for English language learners;
- Development of interactive computer-based items, including complex simulations and technology-based tools, to expand the range and complexity of skills that can be assessed and electronically scored;
- Use of performance tasks and distributed assessments for summative/accountability purposes.

- Measurement of student growth and attribution of responsibility in teacher and principal evaluation systems.

The mission of the new Center for K-12 Assessment & Performance Management (K-12 Center) is to create timely events where conversations on new assessment designs can take place and to make available the best thinking and research on the range of issues facing policymakers. Created in 2009 as an independent resource under the aegis of the Educational Testing Service, the K-12 Center already has produced two sets of research analyses for the conversations ahead.

First, eight papers and discussants’ responses on two major assessment system design goals – the measurement of individual student growth and assessment that informs instruction – are available on the K-12 Center website at www.k12center.org. The papers provide guidance on these critical measurement demands required by the national reform agendas. Leading educational researchers presented the papers at an Exploratory Seminar in December 2009 and discussed their recommendations regarding the R&D work required to meet the new design goals, including the projection of college- and career-readiness, the measurement of teacher effectiveness and the provision of individually diagnostic feedback. One important conclusion from the Seminar was that while RTTT will fuel important advances in the field, the Administration and/or consortia will also need a parallel evaluation plan to determine if each component and the system as a whole function as intended and serve the defined purposes in defensible ways, and how the system can be further improved to support student achievement.

Second, the full papers and the presentations on the four next generation assessment designs presented at the National Conference on Next Generation Assessment Systems are described in this Supplement and are also accessible on the K-12 Center website. Two important conclusions came from the Conference: (1) To more powerfully support instructional and systemic improvements, the next generation of assessment systems will require further research and development on learning progressions and the use of technology to produce more timely and actionable information for educators, students and parents, and (2) the new common assessments will require a clear definition of the priority policy objectives of the system, as well as the inherent limitations and trade-offs involved in any given design.

The K-12 Center website also provides a video of highlights of the national conference, a comparative analysis of the four models, and other materials. The papers and models will be produced in a book later this year.

The Center for K-12 Assessment & Performance Management will continue to support the work of states and school districts in their development and implementation of assessment policies to enhance teaching and learning. The Center will also be active in producing timely and useful policy analyses to inform national, state and local decision makers. Interested partners are encouraged to contact us.
Message from ETS President Kurt Landgraf

An Invitation to the Future

IF YOU WORK IN EDUCATIONAL ASSESSMENT, CONGRATULATIONS: You have been invited by history to help shape the future of public education, and to do so in ways that strengthen your country, improve its communities, and enhance the lives of your 300 million compatriots.

That’s a pretty good invitation. Do I overstate the opportunity? Not at all. As the analyses in this supplement discuss, we are at a unique moment of history. The federal government’s “Race to the Top” education reform initiative, combined with states’ adoption of common academic standards, puts us on the verge of creating a new generation of comprehensive assessment systems to replace the state-by-state mix of tests.

But an invitation only gets us to the dance. Our job now is to convert this public policy opportunity into classroom reality through the design, dissemination and advocacy of assessment systems. ETS and its new Center for K–12 Assessment & Performance Management are in the vanguard of the effort.

Led by Pascal Forgione Jr., the Center embodies ETS’s nonprofit mission to advance quality and equity in education. It leverages ETS’s expertise in K–12 assessment to serve as both resource and catalyst for innovative, research-based assessment and performance management systems — systems characterized by validity, fairness, usefulness, accuracy and timeliness of data.

We are not alone in our commitment to this cause. Representatives from 38 states — policymakers, state and urban district leaders and many of the nation’s leading assessment experts — attended our National Conference on Next Generation Assessment Systems in Washington, D.C., which we sponsored with the Education Commission of the States and the Council of the Great City Schools as co-hosts.

As discussed in these pages, conference participants examined four proposals for the next generation of assessment systems. There were differences in the details, of course. And important questions remain: Should the priority of an assessment system be on providing individual scores that are comparable among states? Or should states’ budget constraints be the highest priority? Should assessment systems stress instructional practice, curriculum rigor and U.S. economic competitiveness? Or should they stress measuring student progress?

Those are important questions. But they are not threshold questions that will decide whether or not we go forward in developing a new generation of assessment systems. The fact is, there is already broad agreement on the need for improved assessment systems. A common set of fewer, clearer and more rigorous standards is the opportunity to put them in place.

“Working together, states may be able to realize economies of scale and other efficiencies that allow the use of more innovative exercise types and the pursuit of a more robust research agenda,” Stephen Lazer, ETS’s Vice President of Assessment Development, notes in his proposal for a model assessment system.

“Furthermore,” Lazer adds, “advances in cognitive science, task design, psychometrics, and natural language processing combined with the wide availability of technology make it possible to assess a more meaningful array of skills and knowledge than ever before.”

Those are ambitious goals. Given our moment in history, they constitute an ambition well within our reach, and an invitation we should accept.