

# **Next Generation Assessment Systems Comparison of the Four Assessment Systems**

Stanley Rabinowitz  
Martin Orland  
Elizabeth Berkes

WestEd

At the National Conference on Next Generation K – 12 Assessment Systems in March 2010, papers on four comprehensive assessment system models to support improved student achievement were presented by some of the leading minds in the fields of assessment design and educational measurement:

- Lauren Resnick, University of Pittsburgh, and Larry Berger, Wireless Generation
- Linda Darling-Hammond and Ray Pecheone, Stanford University
- Marc Tucker, National Center on Education and the Economy
- Stephen Lazer, Educational Testing Service

The ETS K-12 Center for K–12 Assessment and Performance asked these authors to address in their papers several key points about the structure and implementation of their assessment systems, as follows:

1. rigorous standards and good instructional practices
2. technology
3. measuring growth and projecting readiness
4. accessibility
5. technical quality
6. reporting
7. informing instruction and leadership
8. leveraging common standards and assessments
9. implementation timeline
10. cost
11. limitations
12. value vs. burden

The table below presents a side-by-side summary of these key points.

## 1. Rigorous Standards and Good Instructional Practices

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• Summative exams distributed over the year—each one following the relevant curriculum unit.</li> <li>• Formative assessments designed to model how to teach and enable teachers to learn to use those methods as part of their instructional routines.</li> <li>• Exams include high-cognitive-demand tasks and tests of basic procedural skills.</li> <li>• Content and sequence of exams based on Common Core Standards learning progressions.</li> <li>• Scientific (experiment-based) validation process to ensure that exams measure good instruction and what should be taught according to the Common Core Standards.</li> </ul>	<ul style="list-style-type: none"> <li>• Assessments will be designed to reflect learning progressions embedded in the Common Core Standards.</li> <li>• 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 that reflect college- and career-ready standards.</li> <li>• Ongoing formative and interim assessments mapped to curriculum frameworks and standards for use across participating states.</li> </ul>	<ul style="list-style-type: none"> <li>• Integrated system of formative and summative components (not solely a single end-of-year test).</li> <li>• Use of evidence-centered design (ECD) processes to ensure alignment with standards.</li> <li>• Use of mixed item types (selected-response, short answer, and extended answer)</li> <li>• Use of adaptive test administration.</li> <li>• Banks of tasks, assignments, and scoring rubrics coded to specific learning outcomes available for teacher use.</li> </ul>	<ul style="list-style-type: none"> <li>• Builds on highly respected instructional systems currently available in English and modified as needed to reflect the Common Core Standards. The selected existing systems include course syllabi, materials, summative and formative tests, and teacher training.</li> <li>• Aligns lower-division (typically Grades 9 and 10) instructional systems and exams with college- and career-readiness.</li> <li>• Move-on-when-ready design restructures high school to allow students to choose, after they complete the lower division, either an open-enrollment to 2- or 4-year college or additional coursework to prepare for a selective college.</li> </ul>

## 2. Technology

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• Assessments employ computer-based and paper-and-pencil formats, with paper-based work digitized so that both types of work are available for raters and teachers to inspect and annotate online.</li> <li>• Interfaces for remote human scoring and annotation of open-ended expressions of student reasoning; all annotation available for teacher (and parents) to review online or in print as part of individual student profiles.</li> <li>• Mass personalization engine draws on student, class, and school data to recommend which formative assessment to give and when.</li> <li>• Interactive map of standards-based learning goal trajectories facilitates instructional use of assessment data.</li> </ul>	<ul style="list-style-type: none"> <li>• Goal is computer-delivered assessment with digitized student responses for both on-demand and curriculum-embedded assessment components, (with paper-based tests as an option in the shorter run).</li> <li>• Adaptive delivery for on-demand components.</li> <li>• Web-based interface for students, allowing for retrieval, uploading, and storing of work products.</li> <li>• Web-based interface for teachers to train and certify for scoring, facilitate scoring and auditing, and support instructional practice (through ready access to curriculum materials and formative assessments).</li> <li>• Electronic system management and reporting.</li> <li>• Intelligent technologies include formative and summative student performance data for classroom, school, and district information and summative scores for state and cross-state reporting.</li> </ul>	<ul style="list-style-type: none"> <li>• Computer-based testing and flexible (adaptive) administration.</li> <li>• Electronic scoring of some items in conjunction with teachers and/or other human scorers.</li> <li>• Open architecture and standards to facilitate transfer of materials and state/consortium customization.</li> <li>• Integrated data-management system to facilitate periodic assessments and project-based components.</li> <li>• Technology to support a range of accommodations for students with disabilities and English language learners.</li> <li>• Use of computer-based testing to assess technological skills as reflected in the new standards.</li> </ul>	<ul style="list-style-type: none"> <li>• All of the selected systems:               <ul style="list-style-type: none"> <li>○ Use technology to capture student work and answers to exam questions, ship data to scorers, do quality control on scoring, analyze scores, and present data anywhere, anytime.</li> <li>○ Offer web-based access to curriculum designs, course syllabi, lesson plans, and instructional materials on-demand.</li> <li>○ Are built on plans by curriculum specialists.</li> </ul> </li> <li>• Technology provides access to world-wide interactive networks of teachers teaching the same courses, access to experts, and formal training via the web.</li> </ul>

### 3. Measuring Growth and Projecting Readiness

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• Use of equivalent pre- and post- criterion-referenced tests to enable measures of student growth and of teacher/school effectiveness.</li> <li>• Common Core Standards organized as a set of trajectories displayed in <i>honeycomb</i> as a way for teachers, students and parents to track progress in achieving standards.</li> <li>• Allows identification of specific skill deficits, as well as identification of whether students are on track toward meeting academic skill and college- and career-readiness goals.</li> </ul>	<ul style="list-style-type: none"> <li>• Evaluates growth along well-defined learning progressions assessed using multiple assessment formats.</li> <li>• Evaluates students' movement along a vertically scaled continuum utilizing, to the extent practicable, computer adaptive assessment techniques.</li> </ul>	<ul style="list-style-type: none"> <li>• Annual testing between Grades 3-8 to support student growth modeling.</li> <li>• Cross-grade comparability of scores for Grades 3-8 facilitated by coherent standards and expectations across grades.</li> <li>• Growth scores likely appropriate for Grades 3-8 but may be problematic at high school.</li> <li>• Broad range of item types and difficulties could be built into adaptive pool, enabling system to be used for measuring growth while still measuring attainment of high standards.</li> </ul>	<ul style="list-style-type: none"> <li>• Uses end-of-course test with pass points set at college-level readiness.</li> <li>• 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.</li> <li>• Does not provide growth data for each student by course.</li> </ul>

#### 4. Accessibility

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• Assessments designed for all students with flexibility in test administration, modality, and item type.</li> <li>• Assessments for students with low-incidence disabilities may deviate from learning trajectories but will still focus on academic content.</li> <li>• Technology platform and the exams are informed by principles of Universal Design for Learning.</li> </ul>	<ul style="list-style-type: none"> <li>• Uses principles of Universal Design for Learning.</li> <li>• Includes field-testing of all components to evaluate the validity of items and tasks for measuring the content knowledge and skills of special populations.</li> <li>• Creates a blue-ribbon panel to help develop and oversee new directions for the development of tools, strategies, and specific accommodations/modifications for English language learners, students with disabilities, and others from culturally and economically diverse families.</li> </ul>	<ul style="list-style-type: none"> <li>• Through the use of technology, the assessments will provide a wide range of accommodations to students, such as voiced administration.</li> <li>• Computer-adaptive administration allows tests to be tailored to individual student needs.</li> <li>• All items reviewed to ensure elimination of unnecessary linguistic complexity.</li> <li>• Adaptive administration could allow the new assessments to replace some currently mandated special population assessments (No Child Left Behind tests and the 2% tests).</li> </ul>	<ul style="list-style-type: none"> <li>• The selected systems offer accommodations, which could be expanded.</li> <li>• Pearson/Edexcel offers an English as a second language (ESL) course and other second-language courses, and participates in England’s Joint Council for Qualification on Access Arrangements, Reasonable Adjustments and Special Considerations.</li> <li>• University of Cambridge International Examinations offers an ESL course, extra time, adapted test forms, and reading/writing assistance.</li> <li>• ACT/Quality Core provides large print, Braille, reader scripts, and audio-cassettes, but no ESL accommodations.</li> </ul>

## 5. Technical Quality

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• Multiple (i.e., three to five) exams over the year produce a valid estimate of content mastery relative to standards.</li> <li>• Content validity established by closely matching exams to content of model instructional units aligned to the standards (with independent content and instructional experts validating matching and alignment).</li> <li>• Instructional validity established through in vivo (live classroom) studies establishing connection between good instruction and improvements in exam performance.</li> <li>• Reliability on constructed-response items established through explicit scoring rubrics, training of teachers to be scorers, and selective grading of some exams twice by state education agency and/or additional teachers.</li> </ul>	<ul style="list-style-type: none"> <li>• Items, tasks, and scoring rubrics will be constructed from common templates representing the constructs to be measured.</li> <li>• Qualitative and quantitative data collected in pilot years to study items/tasks.</li> <li>• Results from large-scale field tests of items and tasks provide all traditional metrics of reliability and validity.</li> <li>• Scorers certified based on their ability to score to benchmarks accurately; human scoring model includes training, moderation, and auditing to ensure consistency of results.</li> <li>• Data collection design includes links between items and tasks using common students to examine task comparability.</li> <li>• Bias and fairness analyses conducted by adding parameters for differential task function to the item response model, as well as by conducting close-in field trials.</li> </ul>	<ul style="list-style-type: none"> <li>• Comparability of scores and sound measurement practices are high priorities of the model.</li> <li>• Assessment tasks would result from an evidence-centered design process.</li> <li>• Needs to include separate state-specific scales and levels for states that augment the Common Core Standards.</li> <li>• Item calibration to allow adaptive engine’s routing decisions and eventually possibly geared for pre-equating.</li> <li>• Some form of post-equating and post-calibration needed during the first year.</li> <li>• Human-scoring carefully controlled.</li> </ul>	<ul style="list-style-type: none"> <li>• Examination providers are highly regarded and experienced and employ large teams of research scientists to ensure assessment products are fair, valid, and reliable.</li> <li>• Highly experienced National Center on Education and the Economy (NCEE) Technical Advisory Committee.</li> </ul>

## 6. Reporting

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pechone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• Technology platform delivers student work to scorers within seconds. If professional scorers are hired, scoring can be done within 24-hours. If teachers do the scoring, it will depend on the state/district expectations set for them.</li> <li>• Summative results reported to all stakeholders within a few days of scoring.</li> <li>• Results can be aggregated to classroom, teacher, school, district, and state levels, or according to demographic or other subpopulations within levels or across those levels.</li> <li>• System allows teachers, principals, and districts to generate custom reports in real time on demand.</li> </ul>	<ul style="list-style-type: none"> <li>• Uses interoperable electronic platform to provide scores within weeks for both performance measures and on-demand standardized assessments (including open-ended and constructed-responses).</li> <li>• Reporting functions can include student, class, school, district, state, parent, and community summaries.</li> <li>• Includes release of items and tasks to students, teachers, administrators, and parents for greater understanding of expectations.</li> </ul>	<ul style="list-style-type: none"> <li>• 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).</li> <li>• Advocates maximum use of electronic scoring and, for items that cannot be so scored, use of distributed scoring platform to speed results. System should allow for all appropriate levels of data aggregation.</li> </ul>	<ul style="list-style-type: none"> <li>• ACT/Quality Core reported scores by course delivered within 2 weeks online. State to receive additional statewide analysis report.</li> <li>• University of Cambridge International Examinations promises to issue online student and school reports within 10 business days.</li> <li>• 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.</li> </ul>

## 7. Informing Instruction and Leadership

Resnick & Berger (An American Examination System)	Darling-Hammond & Pecheone (Balanced Assessment)	Lazer (An Assessment of Common Standards)	Tucker (Building on the Best)
<ul style="list-style-type: none"> <li>• The summative and formative assessments are intended to provide actionable diagnostic assessment data for teachers and to model quality instruction.</li> <li>• Reports on how students are progressing along the learning continuum to meet Common Core Standards.</li> <li>• Reports for education leaders track student performance by class, by teacher, and by other subgroups.</li> <li>• Use of teachers in cross- or within-school scoring process viewed as important professional development tool.</li> </ul>	<ul style="list-style-type: none"> <li>• 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.</li> <li>• Capacity to show evidence of students' learning progress over time—both during the school year and across years.</li> <li>• Use of teachers in the item development and scoring processes viewed as an important professional development tool.</li> </ul>	<ul style="list-style-type: none"> <li>• Items and tests developed with an understanding of learning to provide information to formative components and more meaningful summative results.</li> <li>• Items model good learning and instruction and therefore make preparation for tests a valuable instructional activity. Summative data may point to areas where further formative diagnostic testing is required.</li> <li>• Use of teachers in the item development and/or scoring processes viewed as important professional when burden is minimized.</li> </ul>	<ul style="list-style-type: none"> <li>• 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.</li> <li>• Board systems provide formative assessment support.</li> <li>• By providing robust curricular support, examinations retain validity through <i>opportunity to learn</i> principles.</li> </ul>

## 8. Leveraging Common Standards and Assessments

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>Exams are directly based on the learning sequences embedded in the Common Core Standards and state standards. Emphasis on one-to-one mapping of assessment items to a particular standard.</li> </ul>	<ul style="list-style-type: none"> <li>Builds on leading work that states and assessment developers have already conducted, accelerating improvements in quality, richness, timeliness, and instructional relevance.</li> </ul>	<ul style="list-style-type: none"> <li>Adoption of common standards and assessments would reduce aggregate test development, psychometric analysis, and system management, allowing savings to support an improved assessment and instructional system</li> </ul>	<ul style="list-style-type: none"> <li>Adoption and modification of existing very high quality, internationally benchmarked examination systems to reflect Common Core Standards is viewed as the most effective and least expensive way to affect classroom instruction and build a comprehensive assessment system.</li> </ul>

## 9. Implementation Timeline

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• Intended to be fully operational within 3–4 years from the beginning of the process.</li> <li>• State level training in Year 1. Teacher education begins in Year 2.</li> <li>• Development of secure and scalable version of initial platform within 6 months of startup with comprehensive system at scale in use beginning in 18 months.</li> </ul>	<ul style="list-style-type: none"> <li>• Year 1: Evaluate existing work to be built upon. Design and develop assessment instruments (for Wave 1).</li> <li>• Year 2: Small-scale pilots and redesign (Wave 1); design and develop assessment instruments (for Wave 2);</li> <li>• Year 3 and Year 4: Field trial, refinement, and capacity for building for scale-up.</li> <li>• Year 5: Scale-up.</li> </ul>	<ul style="list-style-type: none"> <li>• Basic operational system implemented within 3–4 years, including:               <ul style="list-style-type: none"> <li>○ computer-based, adaptive end-of-year tests,</li> <li>○ full accommodations,</li> <li>○ automated and distributed scoring,</li> <li>○ growth measures,</li> <li>○ performance levels,</li> <li>○ libraries of formative materials, and</li> <li>○ periodic assessments and project-based components ready for initial roll-out.</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• NCEE consortium states will begin implementing board examination systems in high schools in fall 2011.</li> </ul>

## 10. Cost

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• No overall estimate of system development costs provided.</li> <li>• Projected ongoing costs of \$15–\$20 per student per subject per year, lower than most current No Child Left Behind tests.</li> <li>• Expenses related to current interim exams (typically \$15–\$20 or more per student) eliminated.</li> </ul>	<ul style="list-style-type: none"> <li>• No overall estimate of development costs provided.</li> <li>• Cost study cited showing that a combination of economies from state consortia, expansive use of technology, and teacher-moderated scoring could result in a per-pupil cost for a high-quality assessment of \$10–\$20 per pupil, depending on the teacher-scoring model used. Projected costs to be no more than the estimated cost of the typical summative state test that provides much less rich information to teachers, schools, districts, and states.</li> </ul>	<ul style="list-style-type: none"> <li>• No overall estimate of system development costs or ongoing maintenance costs provided.</li> <li>• Predicts substantial start-up costs leading to later efficiencies.</li> </ul>	<ul style="list-style-type: none"> <li>• Initial cost to states will be more than current system, but at scale will be less expensive due to more upper-division high school students advancing to colleges by junior year.</li> <li>• Advocates cost savings be used to fund examination systems, teacher incentives, and college scholarships for students and greater instructional time for disadvantaged students.</li> </ul>

## 11. Limitations

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pecheone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>Summative assessments will not be used for accountability until Year 2 or 3 of startup.</li> </ul>	<ul style="list-style-type: none"> <li>Growth vs. grade-Level standards: System emphasizes the measurement of learning and growth over time but can accommodate the measurement of grade-level standards.</li> <li>Learning value vs. speed of reporting: system favors the quality of items/tasks and involvement of teachers in scoring a small number of tasks over immediate return of all scores.</li> <li>Local management vs. efficiency of moderation: For summative assessments, system would develop cost-efficient and comparable modes of teacher scoring (e.g., distributed computer-based scoring) that rely less on local management of scoring than some current systems.</li> </ul>	<ul style="list-style-type: none"> <li>Favors measuring broad range of skills using technology and comparability over fastest-possible score turnaround.</li> <li>Assessments may not be adaptive in the first year of administration.</li> </ul>	<ul style="list-style-type: none"> <li>Does not provide growth data for each student by course.</li> <li>Potentially slower reporting time.</li> <li>Potential modest diminution of reliability in exchange for increase in validity.</li> <li>Will need to invest in improvement of existing systems to ensure most up-to-date features are included, but this development time should be equal to or less than time required to build from scratch.</li> </ul>

## 12. Value vs. Burden

<b>Resnick &amp; Berger (An American Examination System)</b>	<b>Darling-Hammond &amp; Pechone (Balanced Assessment)</b>	<b>Lazer (An Assessment of Common Standards)</b>	<b>Tucker (Building on the Best)</b>
<ul style="list-style-type: none"> <li>• Overarching value lies in refocusing accountability towards improving instruction. The current emphasis on strict standardization is replaced by greater precision and usefulness.</li> <li>• Burden of a distributed assessment system is reduced by online interfaces that make the data usable and actionable to an unprecedented extent.</li> <li>• Remote scoring by teachers increases burden but provides a valuable form of professional development.</li> <li>• Mass personalization reduces burden on students, decreasing the number of assessments needed and items necessary to inform instruction and provide accountability information.</li> </ul>	<ul style="list-style-type: none"> <li>• The range of components included creates some costs and implementation burdens that are the necessary cost of achieving a coherent system that fully represents the standards and supports instructional improvement.</li> <li>• System will require significant investment of resources for professional development, but the subset of teacher-scored items and curriculum-embedded tasks will stimulate change for teachers and students.</li> <li>• The development of a technology platform is challenging but critical for managing costs and building a system that can go to scale. Much work has already been done on the development of the platform.</li> </ul>	<ul style="list-style-type: none"> <li>• Periodic standardized tests and project-based elements coupled with end-of-year test may increase testing time, but this burden is balanced by exams that are instructionally useful and result in improved data and more actionable feedback.</li> <li>• Teacher involvement in scoring can be a burden but, if well planned, can be valuable professional development.</li> <li>• Other benefits include:               <ul style="list-style-type: none"> <li>○ Improved growth data.</li> <li>○ Better comparative data across states.</li> <li>○ Ability to test emerging skills, including those that involve the use of technology.</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Burden varies by board examination system chosen.</li> <li>• States will require their high schools to offer complete core programs, as opposed to current system of picking and choosing certain courses from a system.</li> <li>• States will need to establish an executive authority to manage all aspects of the examination system.</li> <li>• While burdens are not modest, they should be balanced against improvements in student learning, dropout rate reductions, and better preparation for students to succeed in postsecondary education.</li> </ul>

## **For More Information**

For more information on each of these assessment systems, please see the following papers:

Darling-Hammond, L., & Pecheone, R. (2010). *Developing an assessment system of, as, and for learning.*

Available from <http://www.k12center.org/publications.html>

Lazer, S. (2010). *High-level model for an assessment of common standards.* Available from

<http://www.k12center.org/publications.html>

Resnick, L., & Berger, L. (2010). *An American examination system.* Available from

<http://www.k12center.org/publications.html>

Tucker, M. (2010). *As assessment system for the United States: Why not build on the best?* Available

from <http://www.k12center.org/publications.html>

For more information on the Conference on Next Generation K – 12 Assessment System, please see:

<http://www.k12center.org/events.html>