Considerations for Performance Scoring When Designing and Developing Next Generation Assessments

White Paper

Marianne Jones
Daisy Vickers

March 2011

About Pearson
Pearson, the global leader in education and education technology, provides innovative print and digital education materials for preK through college, student information systems and learning management systems, teacher licensure testing, teacher professional development, career certification programs, and testing and assessment products that set the standard for the industry. Pearson's other primary businesses include the Financial Times Group and the Penguin Group. For more information about the Assessment & Information group of Pearson, visit http://www.pearsonassessments.com/.

About Pearson's White Papers
Pearson's white paper series shares the perspectives of our assessment experts on selected topics of interest to educators, researchers, policy makers and other stakeholders involved in assessment and instruction. Pearson's publications in .pdf format may be obtained at: http://www.pearsonassessments.com/research.

Table Contents (Hyperlinked for Ease of Use)

Executive Summary

Next generation assessments will be designed to measure the full range of knowledge and skills students need to succeed in college and 21st century careers. The assessments will be more robust and include more performance items and tasks. Performance tasks in particular are desirable because students can research; apply knowledge, critical thinking skills, and analysis to solve problems; and complete authentic real-word tasks. Performance tasks provide a much more direct link between the act of learning and the assessment of learning.

To truly measure college and career readiness, next generation assessments must include an array of item types; however, item and task complexity will affect the entire assessment process, influencing cost, turnaround times, and ultimately the feasibility of the tests.

This white paper explores the interactions between test design and scoring approach, and the implications for performance scoring quality, cost, and efficiency. This paper is an extension of ideas discussed in "Considerations for Developing Test Specifications for Common Core Assessments" and further illuminates the intersection of theoretical and practical issues related to scoring performance items. It includes recommendations for how states can develop assessment programs that balance the need for richer, timelier information with the need to create affordable, sustainable assessments.

This white paper focuses on performance scoring considerations, including human and automated scoring. More detailed treatment of automated scoring may also be found in the soon-to-be-released white paper entitled "Automated Scoring and Next Generation Assessments." Access to this paper as well as other Pearson research related to next generation assessments is available here.

This paper is organized around the following topics and concludes with a summary and set of recommendations related to:

- assessment goals and scoring implications,
- measurement and test blueprint considerations,
- test and item types,
- test administration, and
- scoring requirements and strategies.

Return to Table of Contents

Assessment Goals and Scoring Implications

There is a close connection among the goals of a given assessment, planned use of its data, and strategies for performance scoring.

The first consideration relates to accountability and whether the assessment will be high or low stakes either for the educator or the student.

In a high-stakes context, quality of the scoring data must take precedence over other considerations if the scoring data are to be used for valid, reliable, and fair measurement. Scoring must follow a rigorous, disciplined, and repeatable process across the cohort of scorers, so scoring standards are applied consistently across the cohort of students.

If the scoring context is not consistent—following different methodologies for each state or district—rater effect or other variability could cloud the accuracy and utility of the data. For example, if scorer training differs or scorer calibration and management during the scoring process differ, scorers may have different perceptions of how to apply the scoring rubrics to student responses. This in turn can lead to differences in frequency distribution of scores. For a given item and rubric, it will be difficult to understand why one state's or district's frequency distribution differs from another state's or district's data if there are many variables affecting scorer judgment. Understanding the resulting differences in student performance levels would be difficult at best.

In addition to the stakes associated with an assessment, it is important to understand how the data will be used to support instruction. Will the data from performance scoring be used to provide diagnostic information about student performance for teacher use? If so, the use of comment codes by scorers may be appropriate. Comment codes extend the rubric by providing additional detail and rationale for why a given student response was awarded a particular score point.

Finally, it is important to understand what is expected from the scoring process itself. As stated earlier, if the primary goal is valid, reliable, and fair scores for student responses, then scoring process integrity and discipline are paramount. If the goal is also teacher professional development, this goal must be designed into the assessment system from the beginning. Later in the paper, we describe and make recommendations for teacher-scoring applications within a balanced assessment system.

Return to Table of Contents

Measurement and Test Blueprint Considerations

Test blueprints provide the architecture for assessments, determining how student performance will be measured against specific curricular standards. It is critical that scoring experts collaborate up front with content experts when building these blueprints, so that what is built can be successfully and efficiently delivered, scored, and reported.

The volume of standards for a given subject and grade will require states and consortia to make decisions about the test blueprint, test focus, and data aggregation and reporting; all have implications for scoring performance items and tasks.

As noted in the paper “Considerations for Developing Test Specifications for Common Core Assessments,” the number of items or tasks in a reporting category must be sufficiently large to support reliable subscores. At the same time, the number of items, particularly the number of performance items, will affect the test administration time for the student and scoring costs and schedule for the states.

The more performance items on a test, and the more complex these items are to score, the greater the cost and schedule impacts. For this reason, states have long held to a limited number of constructed- or extended-response items per test or content area. However, the inclusion of a wide variety of item types and tasks is crucial for measuring student performance against next generation, Common Core State Standards. As we explore in this section, creative test blueprints and item types, scoring technology, and scoring models can provide states and consortia a better balance. Factors include item types, test design and delivery strategies, scoring resources, and how technology choices can affect many of these factors.

Return to Table of Contents

Test and Item Types

Next generation assessment systems will include multiple measures and item types to create a balanced system, measuring student progress, while driving instructional improvement. Test types may include:

- through-course summative assessments,
- computer-adaptive assessments, and
- formative and interim assessments.

Through-course Assessments

To measure learning and provide information throughout the school year, through-course assessments are being considered, so that assessment of learning can take place closer in time to when key skills and concepts are taught and states can provide teachers with actionable information about student learning on a more frequent basis (Partnership for Assessment of Readiness for College and Careers [PARCC] Application, 2010).

Through-course models have the potential to improve the efficiency of scoring by testing and scoring throughout the year, versus the current summative-assessment model of testing, scoring, and reporting predominately in the spring, close to the end of the school year. However, the potential for schedule compression still exists if all states in a consortium test within a constrained window three times per year. Additional considerations for designing the appropriate through-course administration window will include test security, item release strategy, and curricular sequencing (aligning the assessment calendar across multiple states with different academic calendars).

Schedule compression can be mitigated if spiraling strategies are used to extend the testing windows while simultaneously addressing security concerns. For example, if a large number of districts or states test on the same day because of test security considerations, this could drive up the total number of scorers needed at any time, thus increasing scorer recruitment, training, and supervisory and quality-management costs for human-scored test items. However, if a larger item pool is created that supports spiraling of items across different test forms and staggering the testing windows, the scoring windows can also be staggered, reducing the total number of scorers required at any given time. A staggered testing approach could have the added benefit of reducing impact of the new assessments on current state preferences and practices for sequencing content during the academic school year.

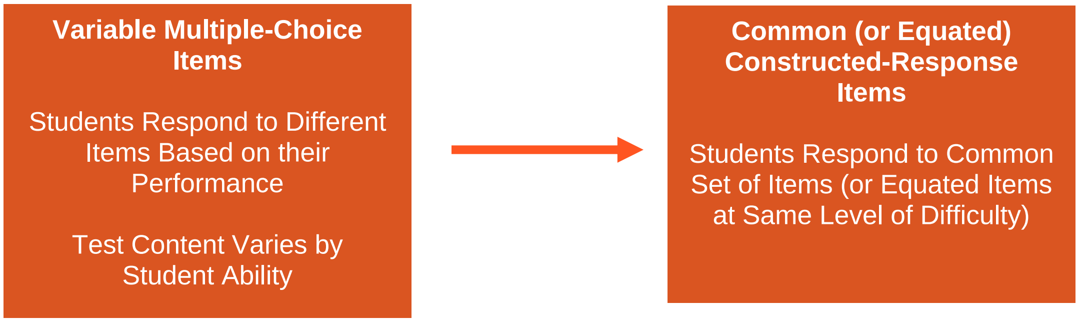
Additional considerations for through-course assessments relate to the specific item types selected (see "Item Types and Complexity" later in this paper) and scorer requirements, including teacher-scoring considerations (see "Teacher Scoring Models").

Computer-Adaptive Test (CAT) Considerations

Implementing a CAT raises interesting and complex considerations for scoring. Multiple implementation scenarios are possible, as shown below in three models, ranging from the simplest to the most complex and costly.

Model 1—Apply the CAT algorithm to the selected response (multiple-choice) items only, and have all students take a common set of constructed-response items. The constructed-response items may be spiraled for security purposes (i.e., two test takers sitting side by side respond to different items), but the constructed-response items will be at the same level of difficulty and are not based on students' responses to the selected response items. This model simplifies both test delivery and test processing and reduces the total number of items that need to be developed, field-tested, and deployed. See Figure 1 below. In this model, the selected response and constructed-response items can be interspersed throughout the test. The constructed-response items do not need to be scored immediately; however, artificial intelligence or human scoring would need to be completed before a final score was reported.

Figure 1: CAT Model 1 with Common Constructive Response Form (Items)



Variable Multiple-Choice Items

Students Respond to Different Items Based on their Performance

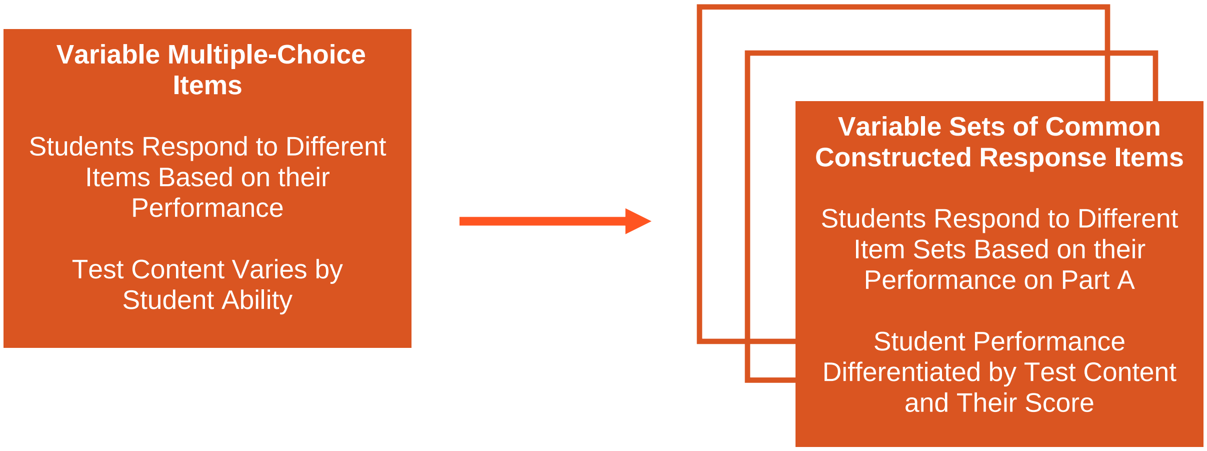
Test Content Varies by Student Ability

Common (or Equated) Constructed-Response Items

Students Respond to Common Set of Items (or Equated Items at Same Level of Difficulty)

Model 2—Use selected-response items to determine the specific constructed-response items a student will receive based on his or her multiple-choice performance. Under this design, student performance on the multiple-choice items will be used to generate an estimate of student ability. That estimate will then determine the appropriate set of items to be administered. Each set of items will vary in difficulty. States may consider three or more levels of difficulty and thus three item sets or forms. See Figure 2 below. As with model 1, constructed-response items would need to be scored using artificial intelligence or human scoring before a final score was reported.

Figure 2: CAT Model 2 with Variable Constructive Response Forms



Variable Multiple-Choice Items

Students Respond to Different Items Based on their Performance

Test Content Varies by Student Ability

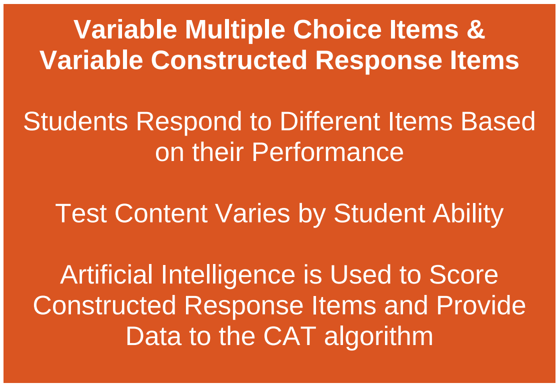
Variable Sets of Common Constructed Response Items

Students Respond to Different Item Sets Based on their Performance on Part A

Student Performance Differentiated by Test Content and Their Score

Model 3—Apply the CAT algorithm to both selected-response and constructed- or extended-response items, so that both item types are variable and personalized, depending on student performance. This is the most complex and costly, and would require automated scoring for all constructed-response or extended-response items. A significant drawback here is that with any item, there will be outliers (unusual student responses) that the scoring engine scores at a lower confidence level. In non-CAT testing environments, these responses would be identified by the automated scoring engine and routed for scoring by humans, and the resulting score would be used for reporting. On a CAT test, a process would have to be established to handle exception cases without interrupting the test taker or impeding his or her ability to complete the test (or test session) in one sitting. One solution may be to generate automated appeals after testing if a constructed-response item scores at a lower confidence level than pre-set parameters.

Figure 3: CAT Model 3 with Variable Constructed Response Items



Variable Multiple Choice Items & Variable Constructed Response Items

Students Respond to Different Items Based on their Performance

Test Content Varies by Student Ability

Artificial Intelligence is Used to Score Constructed Response Items and Provide Data to the CAT algorithm

Formative and Interim Considerations

Formative and interim assessments play an important role in the design of the overall assessment system. Formative assessment shifts the focus from measuring what students have learned to capturing what they know and can do in ways that can inform decisions about next steps in instruction (Black & William, 1998).

In terms of scoring, performance items used in formative and interim assessments are typically scored in the classroom to provide immediate diagnostic information for teachers, although automated scoring may be advantageous to use in situations where these assessments are administered online.

With formative and interim assessments, the *primary* assessment goals tie to relevance and timeliness, rather than accountability, and local, classroom-based scoring is best suited for meeting these assessment goals. Formative assessment is the most appropriate and cost-effective context for local scoring. (An interim assessment can be either formative or summative in nature, but will be given periodically throughout the year.)

Return to Table of Contents

Item Types and Complexity

Figure 4 on the following page summarizes the item types and tasks many states are planning to include in next generation assessments. Figure 4 also shows scoring models applicable to each item type, including the following:

- Machine scoring. Simple scoring rubrics are applied by computer through fixed scoring rules. Machine scoring is typically used to score multiple-choice or selected response items.
- Automated scoring. Artificial intelligence is used to score students' typed responses. The scoring engine is trained using data provided by human scorers.
- Human scoring. Complex tasks and scoring rubrics require the judgment of trained teachers or other qualified scorers.

Although benefits and trade-offs for each method are specific to each item and task developed, Figure 4 summarizes what can typically be expected for cost, complexity, and turnaround considerations.

Figure 4: Item Types and Scoring Methods

Item Types and Scoring Methods		Benefits and Trade-offs				
Student Response Type	Testing Method	Scoring Method	Ability to Measure Critical Thinking & CCSS	Development Cost & Complexity	Scoring Cost & Complexity	Rapid Turnaround Potential
Selected Response (Multiple Choice)	Paper or Online	Machine	Low to High	Low to Medium	Low	High
Computer-Enhanced Selected Response • Drag–and-drop elements (sequencing and classifying) • Select one or more elements • Hot spots (mark one or more locations) • Digital media (audio and/or video)	Online	Machine	Medium to High	Medium to High	Variable, depending on student response formats	High
Constructed Response • Written text (e.g., essay, short answer) • Graphing • Equation/formula construction	Paper or Online	Human or Automated	Medium to High	Medium	Medium	Variable by Scoring Model
Computer-Enhanced Constructed Response • Inserting, correcting, substituting, or completing text • Graphical modeling • Formulating hypotheses or generating examples	Online	Human or Automated	Low to High	Medium to High	Medium	Variable by Scoring Model
Performance Tasks • Immersive, interactive problems • Extended essays • Multi-step problems • Outcome based responses	Paper or Online	Human or Automated	Medium to High	Medium to High	Medium to High	Variable by Scoring Model

Selected Response (Multiple Choice)

Selected-response items, whether delivered on paper or via computer, can be machine scored and thus are very efficiently processed. Significant enhancements continue to be made in the quality and richness of traditional and computer-enhanced selected response items to measure a broader array of cognitive attributes. Selected response items contribute to test reliability and validity, and are well suited to measure some of the standards found within the Common Core State Standards (CCSS).

At the same time, as British anthropologist and scientist Roger Lewin once noted, "Too often we give children answers to remember rather than problems to solve." This parenting advice also applies in the assessment world—the more an item does, the less the child will have to do. Selected-response item types reduce the testing process to one of *selection* versus *creation*. No new student ideas or work are created as a result of a selected-response item. Going forward, high-quality assessments designed to measure critical thinking and support instructional improvement must go "beyond the bubble" (Duncan, 2010).

Computer-Enhanced Selected and Constructed-Response Items

With online testing, states have assessment options that go beyond multiple choice and human-scored constructed response. Innovative items, such as those with drag-and-drop features, hot spots, and other types of interactivity, can be developed with scoring rules that allow the items to be computer-scored. Moreover, these online items are engaging for students, allowing them to interact with content in ways that motivate performance and correlate more strongly with how 21st century students learn.

Computer-enhanced items include selected response, constructed response, and performance tasks. There is a wide range of how the technology is used for each of these item types.

Computer-enhanced items, while more costly to develop, can be rapidly and cost-effectively scored. Items using drag-and-drop features, hot spots, and other types of interactivity can usually be machine scored, similar to how a selected-response item is scored. Additionally, it is possible to place more computer-enhanced items on the test blueprint without negatively affecting the test administration time for the students and scoring schedules, which in turn supports better subscore reporting. Minnesota, among other states, has moved to this type of item to preserve the benefits of constructed-response items without the scoring cost. Click here to view examples of innovative Minnesota science items.

Constructed-Response Items

In the past few years, some states have eliminated constructed-response items from their state assessment programs because of budgetary and other constraints. The single largest driver of scoring cost for constructed-response items is the length of time it takes to score an item, referred to as the scoring rate. Score rate directly affects the number of scorers needed, regardless of scoring window length. A slow rate means more scorers, increased training time, and the probable need for increased supervisory staff.

The more complex the constructed-response item and rubric, the slower the item will typically score, with portfolios often being the most complex (for example, scoring a series of artifacts for one student).

For human-scored constructed-response items, Pearson has developed a scoring complexity index (SCI) that uses a mathematical algorithm to assign a value for different types of items. This complexity value helps to predict the effort that will be needed to score constructed-response items.

Within the SCI, complexity is defined as the difficulty of assigning a score to a student response consistently and reliably. Factors built into the algorithm include score points (rubric), number of traits, response length, and quality metrics (scoring reliability and validity). Each of these factors influences the complexity and therefore the rate and efficiency of scoring. An SCI or an equivalent measure could be used at the time of test blueprint to assess impact of item complexity on scorer training development, training time, and scoring.

Scoring rate will be a factor for either professionally scored items or teacher-scored items. When employing teachers as scorers, long scoring windows will not only affect cost and schedule, but will also affect the amount of time a teacher needs to be away from the classroom in order to participate in the scoring process.

An item's suitability for automated scoring is another key cost consideration. Automated scoring is in many cases less expensive than human scoring and applications of automated scoring to extended response items—both essays and more content-based prompts—is quite mature. To take full advantage of the benefits automated scoring can provide, test designers should include artificial intelligence considerations at the beginning of the test development process. Further, technology and scoring experts should be involved in item and rubric review to optimize the quality of automated scoring and reduce (where possible) the number of outliers that cannot be scored at a high enough confidence level by the engine.

The state of the industry is fully explored in our “Automated Scoring” white paper. However, in general, automated scoring of very short constructed-response items and mathematics items is more challenging than automated scoring of extended responses.

Performance Tasks

Performance tasks have been used by some states and international assessment communities for many years, and will be the bedrock of next generation U.S. assessments developed by states and consortia:

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

Although higher order thinking skills can be measured by well-designed selected response items, performance tasks allow for measuring multiple components of knowledge and skills within the structure of the same task. Performance tasks are more complex and may take place over multiple days or class periods.

Tasks are crucial for next generation assessments, but the design of these items will greatly affect scoring feasibility.

For formative assessments, we recommend classroom-based scoring done by the teacher or through automated scoring technology (trained by teachers). In fact, teachers are the best authors and evaluators of this content.

Technology can greatly aid the scoring of formative assessments, whether done online or via paper. Desktop or open source, Internet-based applications enable teachers to:

- design items that align to common standards;
- share items across classes, schools, or even districts;
- publish tests in the manner most appropriate for the assessment (paper or online); and
- leverage artificial intelligence (for online tests) to automatically score essays and provide instant feedback to students.

Teachers’ feedback on formative assessments can also be captured in a variety of ways, including innovative strategies like creating personalized podcasts for each student.

On the other hand, for high-stakes summative and through-course assessments, large numbers of student responses need to be scored in a short period with a high degree of consistency. This warrants a different approach, so these complex performance tasks can be scored accurately, consistently, and efficiently, within the likely scoring timelines for next generation summative tests.

One option for states to consider is the development of performance tasks requiring students to research, engage in real-world tasks, do experiments and so forth, but create a series of more standardized response mechanisms. For example, the New England Common Assessment Program (NECAP) assessments include performance-based science items wherein students collaborate to conduct science experiments, make observations, and record data. The students then use their data sheets to individually complete a series of traditional constructed-response items to measure the application of their knowledge to problem solving. While it would be costly to video tape student science experiments, it is completely feasible to score these constructed-response items predicated on science experiments and classroom collaboration. Click here to view sample NECAP science inquiry tasks.

For high school assessments, states may also look at postsecondary models for performance tasks. On the City University of New York (CUNY) Proficiency Exam, formerly used as a matriculation requirement from the sophomore to junior year (or to award students an associate's degree), students were required to read specific texts prior to sitting for the exam. During the test administration, the students were required to read additional text and synthesize their reading while responding to a variety of constructed-response items, which could be traditionally scored by professional scorers.

Another very useful postsecondary model is the Council for Aid to Education (CAE's) Collegiate Learning Assessment (CLA). In "Architecture of CLA Tasks," CAE describes its test strategy in this way:

> The CLA uses direct measures of skills in which students perform cognitively demanding tasks from which quality of response is scored. All CLA measures are administered online and contain open-ended prompts that require constructed responses. There are no multiple-choice questions. The CLA tasks require that students integrate critical thinking and written communication skills. The holistic integration of these skills on the CLA tasks mirrors the requirements of serious thinking and writing tasks faced in life outside of the classroom (Architecture of CLA Tasks, 2010).

The CAE is scored on a 6-point analytic scale. For many years, it was scored by human readers. The effort was time and cost intensive. More recently, however, CAE has adopted automated scoring, with outliers (unusual responses) scored by trained and qualified human readers.

More information on CAE and use of performance tasks may be found here.

The state of Ohio provides another example. Ohio has collaborated with the Stanford Redesign Network and other partners to explore development and implementation of performance tasks. Schools selected for the two-year Ohio Performance Assessment Pilot Project use research-based best practices for developing performance assessments in English, mathematics, and science. Similar to the CUNY example cited above, sample items from this project demonstrate that student responses can still be captured in a simple response document, and thus efficiently and effectively scored. Complexity in the task presented to the student does not have to result in complexity of the response artifact and ensuing scoring process.

Finally, international assessments make extensive use of both constructed-response items and performance tasks. In the UK, for example, the General Certificates of Secondary Education (GCSEs) typically include both internal (controlled) assessments and external exams. The controlled assessments in particular are heavily predicated on performance tasks. The rubrics (mark schemes) provide extensive guidelines to scorers, including sample response content.

Within Pearson's Edexcel products, controlled assessments are replacing curriculum-embedded assessments for the GCSE. The primary difference is that students produce their work under controlled conditions, within a specified amount of time. Control can be applied to the task setting, the task taking, and the task scoring (marking). The nature of control varies according to the content area in order to secure the validity, reliability, and manageability of the assessment.

Students have access to the tasks, based on scope and sequencing of instruction, prior to exam day. Students' preparation for the tasks can include reading a range of texts from a pre-defined syllabus and accessing a variety of media (Internet, TV, videos, etc.). Teachers may support students through the preparation process, and students may also collaborate in groups; however, they must provide an individual response to the task under controlled testing conditions (as in the NECAP example cited above). The nature of the controlled assessments makes them more conducive to consistent, reliable test administration and scoring than the former curriculum-embedded assessments.

Scoring Rubrics

Rubric development will occur simultaneously with content development for next generation assessments.

Rubric ranges and complexity are a key consideration when balancing the measurement and feedback objectives of the assessment with cost, schedule, and quality concerns.

Tenets of good rubrics are similar, whether the rubric is used for instruction or scoring:

- Scoring rubrics must use clear language.
- There must be distinct differences between score points, and each score point must have its own description. We strongly guard against using a single description for multiple score points (for example, 1-2 points for "response demonstrates minimal understanding," 3-4 points for "response demonstrates adequate understanding," and so forth). Rubrics with multiple score points per description allow too much personal discretion and variability in scoring, compromising scoring reliability and validity.
- Score point descriptors should be parallel from one score point to another. In other words, criteria used to evaluate a response should be represented in each score point.
- Careful consideration should be given to the range of score points. Too wide a range of score points may make it difficult to create meaningful and discernable differences between each point, and may make it difficult to find clear exemplars of a sufficient number at each score point to train scorers. An overly lengthy or complex rubric will slow scoring without adding value to the measurement goals of the item. On the other hand, a rubric with a very compressed scale may not sufficiently distinguish between different levels of performance.

- Typical score point ranges are 0 – 2 for short-constructed-response items and 0 – 4 or 1– 4 and 1– 6 for extended-response items. For performance tasks, the score point ranges may be similar to those for an extended response item but may be applied to several domains or aspects of the task.
- The rubric should clearly articulate criteria for awarding each score point and for awarding any points at all. Is "0" an earned score point or a blank response? For essays, how should off-topic or off-mode essays be treated (for example, the student writes a letter instead of a narrative)? Depending on the item or prompt, a separate condition code rubric may be warranted to clearly define what constitutes a scoreable or non-scoreable response, and how non-scoreable student responses will be handled.
- Rubrics should be accompanied by examples at every score point. If the constructed-response item is one in which students may take a positive or negative position on a topic or approach the response in several ways, sample responses should be written to reflect these alternatives.

Rubrics may be holistic or analytic, depending on the item or task. An analytic rubric will often specify the number of examples, reasons, or pieces of evidence along with other criteria, including correctness of response (mathematics or science items). A holistic rubric is applied to the student response as a whole and uses language to describe levels of performance, for example, "clearly explained," "partially explained," or "minimally explained."

The use of the data from a test will determine the type of rubric appropriate for the item or task. For a large-scale, high-stakes assessment, the holistic rubric is often the rubric of choice. It gives a snapshot of overall performance. Additionally, holistic rubrics support an efficient, reliable scoring process.

For states and customers requiring more detailed diagnostic information, comment codes can accompany the holistic scores. Comment codes are standardized annotations that can be applied during the scoring process to provide rationale for the particular score a response receives.

Alternatively, for an assessment that gives diagnostic feedback, an analytic rubric may be more appropriate. It better pinpoints the areas of instruction where improvement is needed. Analytic rubrics are often a very effective tool for classroom use.

States have recently been exploring changes to the rubrics, including moving from complex, multi-trait writing rubrics to holistic rubrics, which are highly effective from both a measurement and scoring perspective. The Arizona Instrument to Measure Standards (AIMS) program has moved from a 6 X 6 model (six traits, six points each) to a holistic scoring rubric after a year-long project involving research, teacher input, national and state advisory groups, and a review of rubrics used in state assessments and rubrics used for the SAT and ACT essays. AZ concluded that:

- holistic rubrics allow scoring to focus on an assessment of writing skill overall instead of each individual trait of writing;
- not all traits are equal in value, based on research by Paul Diedrich, but were treated as such in the trait-based rubric (for example, conventions is as important as ideas and content); and
- the trait rubric was best suited for classroom use, whereas the new holistic rubric based on those traits was best suited for summative AIMs scoring (Beach, 2010).

More information on the existing and new AIMS writing rubrics may be found here.

Rubrics may be item- or task-specific, or generic and broadly applicable across a range of items of tasks within a given content area. Item-specific rubrics provide more precise guidance to raters and can lead to efficient, consistent scoring. Item-specific rubrics are well suited to short constructed-response items. The only cost trade-off is the requirement to develop item-specific training for every item on a test.

For portfolios or complex performance tasks, common rubrics may be more appropriate. Scorers internalize these rubrics and can apply them to range of tasks within a content area. In this context, scorers can be certified on a rubric instead of on every item, saving training development time and cost while *enhancing* scoring quality and consistency, because scorers become very familiar with the rubrics and adroit at applying them to student work. There are additional benefits as well, which is why the SMARTER Balanced Assessment Consortium (SBAC) plans to use task templates and common rubrics:

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

When common or generic rubrics are used, they must be accompanied by item-specific, benchmark student responses (exemplars) to illustrate how to apply the language of the rubric (general performance level descriptors) to actual, task-specific student responses.

On international assessments, it is standard practice for the scoring rubrics (marking grids) to be common from one assessment to another, so teachers or scorers can internalize the standard and apply it with consistency. This is a tried and tested practice internationally that could be adopted for next generation assessments in the US.

Stimuli and Passages

Constructed-response items often include stimuli such as reading passages, maps, graphs, pictures, charts, or problem statements. Using the stimuli, the student may be prompted to compare and contrast; show cause and effect; identify conflicting points of view; categorize, summarize, or interpret information; or develop generalizations, explanations, or evidence-based conclusions. Student responses demonstrate complex thinking skills (Darling-Hammond & Pecheone, 2010).

In a traditional constructed-response item, prompts that are linked to stimuli are self-contained; by design, they are not dependent on any information beyond what is contained in the stimuli to elicit the students' opinions or prior knowledge. For curriculum-embedded performance tasks, this is not necessarily the case. The student may be asked to synthesize, analyze, or form opinions about stimuli from their coursework in conjunction with a passage on an assessment. Curriculum-embedded tasks, scored by teachers in the classroom and used for formative purposes, can make extensive use of these connections.

For summative assessments, however, careful consideration needs to be given to the depth and breadth of material (stimuli) used in order to support consistent, reliable, and efficient scoring. The scorers need to be well-versed in these materials in order to score them accurately, and scorers need to understand the specific parameters and options afforded to the student. For example, in the UK, a student may be required to conduct advanced research or reading, but the work will be done from a prescribed syllabus of academically rigorous and appropriate material.

Return to Table of Contents

Test Development Strategies

Pilot and field testing of next generation assessment items provides a crucial opportunity to test item reliability and validity, and also to secure feedback from scoring professionals about how an item will perform within a scoring context.

Downstream scoring plans should influence the field test plan. For example, if an item is to be scored operationally using artificial intelligence, states may want to consider 100% second scoring (two independent readings) for the field test. The field test scoring rules should also include resolution scoring when the readers disagree by more than one point. The additional data provided by reliability and resolution scores are useful for training the scoring engine on score-point boundaries. For instance, a response that receives two “2s” by human readers or two “3s” by human readers creates clear definitions of those score points. However, a response that receives a “2” by one reader and a “3” by another starts to describe papers that are on the border. The engine uses this data to learn to score just as human readers do.

This field test data will also be used to determine which items are suitable for automated scoring. Assessing the confidence levels of automated scoring should be done before final selection of items for operational tests. While 100% second scoring of field test items may add scoring costs during test development, this data is crucial for successful implementation of automated scoring, which in turn reduces operational scoring costs.

Equating plans will also influence field test scoring. If operational scores are to be equated using field test data (in other words, if the forms are to be pre-equated), the field test should be scored with a higher percentage of second reads, including 100% second scoring of more complex items or writing prompts.

If items are to be post-equated (equated during or after operational scoring), the equating sample size and specific sampling requirements must be carefully managed throughout the assessment life cycle (from test design through receipt and processing) to mitigate impact to the scoring schedule and costs. For example, if the sample size for equating is large (50% or more of the total student work), additional scorers may be required in order to score this large volume quickly enough for the psychometric work to be done during equating and prior to data aggregation and reporting. Additionally, the sampling plan needs to be structured to provide demographic representation and a likely range of student performance in the early testing/early returning districts. Without a well-defined plan, states and consortia may need to score a majority of responses in a very tight scoring window simply to capture the right mix of students and student performance for equating.

Return to Table of Contents

Test Administration

Scoring and Reporting Schedules

The need for rapid report turnaround and thus a compressed scoring window (total days to score) can drive dramatic increases in the number of scorers required, thereby increasing training costs and potentially adding cost to quality management (i.e., more scorers to supervisors, potential for lower reliabilities).

Given that one- or two-week scoring windows are highly likely for Common Core assessments, one way to mitigate schedule impact on scoring is to stagger test administration windows across a consortium, as described earlier in this paper. One drawback may be equating across states, although having some early testers and early returns from across the states may make this a non-issue.

Administration method will also have some influence on turnaround times for scoring and the overall reporting schedule. For example, paper-and-pencil administrations will require shipping and processing time. Conversely, online administration eliminates document transit and allows scoring to start more quickly.

Online Testing

There are numerous benefits to the scoring process when items are tested online, including the possibility of using artificial intelligence to score appropriate items. Additional benefits include the significant schedule reduction between test taking and scoring. By eliminating paper shipping, receipt, and processing, online student responses can be made available to scorers within a very short timeline after the student takes the test. All other conditions being equal, this facilitates quicker reporting timelines relative to paper-based administrations and also allows more time to be allocated to critical scoring tasks and less time to steps of lower value in the testing process.

The most time-efficient and cost-effective models will involve online test administration coupled with online scoring processes. A combination of online, human, and automated scoring will:

- promote scoring efficiency, quality, and consistency;
- support significant improvements in test-item variety and complexity, including expanded use of performance-based items, similar to international models; and
- involve teachers in the development and scoring of test items.

Many states have started the transition to online testing but are still developing and administering their exams in both modes—paper and online. In order for online testing to be truly time-efficient and cost-effective and in order for it to support the problem solving and critical thinking skills implicit in the Common Core standards, it needs to be the single mode of administration for a given content or grade. As long as states are supporting both paper and online testing, the anticipated savings will not be recognized, in part because of duplicate processes for item development, review, administration, and scoring.

Online testing is a prerequisite to realizing the benefits of automated scoring. Automated scoring is particularly well suited to English language arts (ELA) test items and most content-based items, passage-based items, and essay responses. Automated scoring engines combine background knowledge about English in general and the assessment in particular along with prompt- or item-specific algorithms to learn how to match student responses to human scores. The scoring algorithm is adapted for each prompt, based on the types of answers students write in response to that prompt and to the way in which human readers score those answers. Research has demonstrated both the accuracy and efficiency of automated scoring engines (Pearson, 2009).

Paper-Based Testing

Even paper-based test taking can translate into highly efficient, digital scoring with image-scanning technology and an internet-enabled scoring platform. Research conducted in 2002 showed that digital scoring was 15% more efficient than paper-based or booklet scoring (Nichols). With current digital scoring capabilities, this efficiency factor is likely even greater than 15% today.

As states and consortia consider paper-based testing for lower grades or during the transition to online, it will be important that scanning operations meet industry standards. Constructed-response items need to be image-captured with high-quality and appropriate metadata, crucial for managing and controlling the workflow through the scoring process.

There are numerous other technical considerations for scanning that directly affect scoring. For instance, item-level scoring is more efficient than booklet-based scoring. When scorers focus on scoring one type of item at a time, they become very proficient, which positively affects both scoring quality and rate. Additionally, item-level scoring avoids the potential for "halo" effect wherein one scorer judges all constructed-response items for a student and their perceptions of a student's ability are unduly influenced by the other items they have read for that student. In order for item-level scoring to be possible, the form design and scanning solution need to support image "clipping" of individual items. Image clipping is the process of taking several constructed-response items from a student-response page and creating separate images of each response, so they can be separately routed for scoring.

One other critical point for states and consortia to study is whether it is preferable to offer mixed-mode testing for the lower grades, or to offer a completely paper-based test for the lower grades and a completely online test for grades 6 and above. As stated earlier in this paper, mixed-mode testing within a subject and grade will require states to invest in the infrastructure to support online testing, without being able to realize all benefits of online student responses, including processing speed, opportunities for automated scoring, and other benefits. Where partial adoption of online testing makes sense, states and consortia may consider making grade-specific decisions (for example, grades 3 through 5 tested entirely on paper, but grade 6 and above tested entirely online) or even content- and grade-specific decisions (for example, certain grades of mathematics initially on paper and English language arts online). This approach would be most cost-effective and avoid complexity and comparability issues for mixed-mode testing within a given grade or content area. The benefits of this approach must be weighed against the complexity it adds to administration for districts, schools, and classrooms.

Return to Table of Contents

Scoring Requirements and Strategies

Scoring Models

Various scoring models can be deployed for human scoring of next generation assessments, including distributed, regional, and local scoring, based on assessment goals and priorities.

Distributed Scoring

Distributed scoring is an Internet-based scoring platform, enabling participants to score from home or the classroom/office. It has been called out as a requirement by states and consortia as part of the solution for next generation scoring. It is an operational model used for scoring high-stakes assessments today, including college entrance exams and state assessments. Distributed scoring eliminates cost, schedule, and logistical barriers associated with “bricks and mortar” scoring wherein teachers or professional scorers travel to scoring centers to participate in scoring large-scale operational assessments. An effective distributed scoring system also eliminates technical barriers to participation, so that teachers, retired teachers, academicians, and trained professional scorers with a very basic computer and Internet access can score from home or school anywhere in the country.

Distributed scoring is crucial to support the scale of next generation assessments and the likely production requirements (scoring level of effort and schedule). The model also offers a key advantage in terms of diversity, so that the pool of scorers is as geographically and demographically diverse as those taking the tests. Finally, distributed scoring is an on-demand system. Volumes can be handled particularly well when a nationwide pool of scorers is screened, qualified, and ready to score. For example, Pearson has a database of more than 52,000 qualified applicants and can handle changes in volume or schedule more easily than when dependent on local or regional markets for hiring of teachers and scorers.

Regional Scoring

Regional scoring, where scorers convene in centralized scoring centers, is a good model for low volume scoring or paper projects, including paper-based alternative assessment portfolios. As portfolios transition to online, either a regional or distributed model will be effective.

Regional scoring is also often selected when online training is not feasible or cost effective. For example, for field-test administrations the volume of items is high but the student count is low. Developing item-specific online training for field tests does not have the same return on investment as developing online training for large-scale operational administrations.

Local Scoring

Local scoring is classroom-based and typically used for formative or interim assessments for which immediacy is key, teachers drive assessment content, and scoring and student data are not aggregated for district or state comparisons.

Figure 5 on the following page describes and articulates the advantages and trade-offs among various scoring models, including distributed, regional (centralized), and local scoring.

Figure 5: Next Generation Scoring Models

Scoring Model	Description	Recommended Applications	Benefits	Challenges or Limitations
Distributed Scoring	Distributed scoring is a secure, Internet-based scoring platform, enabling teachers and professional scorers to score from home or the classroom/office.	● Any assessment type, but typically large-scale summative assessment ● Any test delivery mode that can be digitized (paper-and-pencil tests that are image scanned; online tests, and digitized portfolios with images, video or audio components)	● Promotes inclusion and diversity in the scoring pool (low barriers to participation for teachers or those living in rural areas) ● Large scoring pool promotes rapid turnaround of results	● Paper-based student responses must be digitized ● Scorer training must be suitable for online delivery
Regional (Centralized) Scoring	Centralized scoring occurs when scorers gather in one or more regional centers	● Any assessment type ● Field test scoring ● Teacher workshops and scoring institutes ● Any test delivery mode (paper-and-pencil, online, portfolio, audio, video, etc.)	● Provides cost-effective scoring method when instructor-led training is required ● Enables item review and collaboration for item try outs and field tests	● Cost and schedule challenges for scorers who must travel to regional centers ● Cost and schedule challenges associated with shipping materials to a central location to score ● Scoring schedule constraints resulting from the limited scoring pool ● Opportunity for regional or site-based scoring variability in instructor-led training or management
Local Scoring	Scoring from the classroom	● Formative assessment ● Scoring of locally developed tasks ● Low stakes applications	● Immediacy of results ● Teacher involvement and control	● Potential for variability of scoring outcomes

Teacher Scoring Models

Teachers are crucial stakeholders in educational reform and in building and scoring next generation assessments. Teacher scoring can be deployed using any of the scoring models above (distributed, regional, or local models). The specific ways in which teachers participate in the scoring process, however, will vary depending on the purpose or use of a given assessment. Figure 6 on the following page summarizes our recommendations.

Figure 6: Teacher Scoring Models

Scoring Model	Description	Recommended Applications	Benefits	Challenges or Limitations
Teachers Score Their Own Students' Work	Teachers directly score their own students' work. Student responses are neither anonymous nor randomized. Teachers may score multiple items for a given student in one sitting (e.g., performance tasks or portfolios).	• Formative assessments • Interim assessments • Teacher- or locally developed performance tasks	• Immediacy of results • Meaningful diagnostic information about teachers' own students to help guide instruction and intervention • Teacher professional development • Teacher buy-in	• Requires controls to ensure scoring reliability and fairness (e.g., audit scoring or moderation) • Possibility of halo effect (the student's general classroom performance affecting his or her score on an assessment) • Burden on teachers
Teachers Score Randomized Student Work	Teachers score a randomized and anonymous selection of student work	• Summative assessments • Performance assessments to demonstrate teacher effectiveness and teacher contribution • Teacher professional development	• Provides reliability and validity (assuming best practices of 10% to 100% second scoring) • Optimizes work load (teachers are not locked into scoring a set amount) • Exposes teachers to broad range of student work	• Standard bias prevention training will need to be included • Burden on teachers
Teacher Moderation or Auditing	Teachers provide the first score on student work, and professional scorers audit this activity (anywhere from 5% to 100% sample size)	• May be applied to any assessment type	• Increases score reliability and validity	• Cost and complexity
Teachers Train Automated Scoring Engine	Teachers score responses that are used to in turn train scoring engines to score student responses using computer algorithms	• May be applied to any assessment type	• Significant cost and schedule benefits • Teacher involvement	• Not all items are suitable for automated scoring

Teacher scoring can provide significant professional development opportunities. However, several strategies need to be considered to support effective and efficient scoring in the teacher-scored model:

- Level of district engagement
- Incentives for teachers to participate (to encourage that they not only train and qualify or certify, but also continue to scoring)
- Use of distributed scoring to remove barriers to participation, eliminate commuting times, and enable teachers to score from school or home
- Use of distributed scoring to support critical workflow and routing rules across scoring pools/states, such as consortium requirements that teachers not score the work of their own students for summative assessments
- Use of substitute teachers (if the scoring pool will comprise only active/current classroom teachers, states should consider substitute teacher pay to enable teachers some time away from the classroom to score)
- Complementing the pool of teacher scorers with experienced, qualified professional scorers, retired teachers, and substitute teachers (all of whom may be able to contribute more scoring hours per day than active teachers)
- Teacher qualification or certification process, and the financial and policy impacts when individual teachers fail to qualify
- Use of moderation or auditing (for example, use trained and certified professional scorers to apply a second score or an audit score to monitor consistency and quality of teacher scores if used for accountability purposes)

States and consortia need to make decisions about the priorities for teacher professional development and the best use of teachers' time. Classroom teachers clearly derive great benefit from scoring formative and interim assessments, as do their students. Classroom teachers also derive benefit from understanding and participating in large-scale summative assessments. However, should that participation be voluntary and variable depending on a teacher's workload, or should it be a mandatory requirement? In a production-scoring environment for high-stakes tests, all scorers will be required to score a certain number of hours per week so that scoring remains consistent, scorers retain their training, and scoring can be completed on the required timeline. In this environment, will mandatory teacher participation be more of a burden than an opportunity?

We recommend a voluntary system of engagement, where teachers are encouraged to participate in scoring and systems like online training and distributed scoring are provided to enable that participation. However, the production scoring burden for large-scale summative assessments should not be placed entirely on active teachers, but rather a mix of teachers and highly trained professional scorers. This is particularly important for through-course assessments, which will be administered throughout the school year and scored in very compressed timeframes.

Scorer Training

Training development costs can be a key cost driver, but high quality scorer training is foundational to scoring reliability and validity. The following are some key considerations when designing scorer training:

- Online, asynchronous training can greatly enhance the efficiency and consistency of training results. Online training can be used in any of the models presented in Figure 5 above. Online training achieves results equal to instructor-led training and is more efficient for the participants (Wolfe, Vickers & Matthews, 2009).
- Training and qualification by rubric (vs. by item) can further enhance training consistency (across items, administrations, and years) and training efficiency
- Common or generic rubrics for performance tasks or complex items can increase training efficiency while still meeting quality goals and enhancing consistency from item to item or year to year.
- When teachers participate in the scoring process, training considerations include
  - elaborating on the key differences between grading and scoring;
  - tailoring bias prevention training to address particular preconceptions teachers may bring to the scoring activity; and
  - adding professional development elements to the training, for example, training on the assessment process overall to promote assessment literacy and providing additional content on how performance assessments can drive instructional improvement.

Scoring Technology

The National Education Technology Plan 2010 (NETP) calls for revolutionary, not evolutionary, transformation and calls for the educational system to redesign structures and processes for effectiveness, efficiency, and flexibility, and continually monitor and measure performance. The plan recognizes that technology is at the core of education reform, and that technology-based learning and assessment systems will be pivotal to improving student learning and generating data that can be used to continuously improve the education system at all levels (Transforming American Education: Learning Powered by Technology, 2010).

Much as online testing capabilities are transformative for the range and type of items that can be included on tests, so too is scoring technology transformative in improving the quality, cost, and range of scoring options. Scoring technology can be leveraged to:

- Support nationwide, distributed scoring
- Enhance automation and quality management for human scoring applications, such that quality management tasks are systematic, predictable, and enacted with minimal human intervention (thus minimizing human error in monitoring)
- Monitor and report on human scoring activities on a near real-time basis
- Support artificial intelligence scoring, including scoring both text and speech (particularly applicable for speaking and listening tasks)

Of particular interest for the states and consortia are emerging capabilities in ePortfolio applications that support the management of student (or teacher) artifacts. Managing complex performance tasks or portfolios of student work in an efficient and effective manner can be a major challenge and cost driver. Currently, most paper-based portfolios or video- or audio-based scoring must take place in centralized locations, using a variety of disparate media and data capture solutions to record scores. ePortoflio applications enable the scoring process to be media agnostic (for example, the use of viewer technology enables many different media types to be viewed) and the scoring process to be location agnostic (e.g., Internet-based).

Scoring Rules

Scoring rules determine the number of scores that will be applied to a single student response. For constructed-response items, the greater the complexity, the more reads that are typically required to ensure acceptable reliability. For instance, writing prompts are often 100% second scored, in other words, scored independently by two readers. On the other hand, a simpler mathematics item with a limited score-point range (0-2), may be 10% second scored, with the second score only used to capture reliability statistics and monitor scoring quality.

There are many national and international examples of performance assessments that are scored with no second scoring. We do not recommend this approach for both scoring and measurement reasons. Second scores provide a check of the reasonableness of the scores and the adequacy of the scoring. Including some percentage of second scoring allows for the measurement of scorer reliability. This is especially important for rubrics that are more complex and when score points are difficult to differentiate. Reliability ratings will show the extent to which all raters are applying the rubric in similar and consistent ways.

A 2007 College Board research study on Advanced Placement (AP) scoring examined strategies for monitoring scorer performance and discussed in detail the benefits of appropriate scoring rules to support the twin goals of measurement and reader monitoring:

> If the connection among Readers in the rating design is not adequate, then there will be disconnected subsets of Readers in the analysis, which will be problematic for the measurement model as it carries out its estimation process. When there are disconnected subsets of readers, Reader severity and level of student achievement are confounded, leading to ambiguity in the calibration of both the Readers and the students. That is, if the average rating for a Reader is lower than the average rating of other Readers (i.e. the Reader appears to be “severe” in comparison to others), the measurement model will not be able to determine whether the Reader tended to assign systematically lower ratings than other Readers, or whether the set of student essays the Reader evaluated tended to be lower in quality than other students’ essays. In order for the measurement model to be able to disentangle differences in levels of Reader severity from differences in levels of student achievement, the various subsets of Readers need to be connected (Wolfe, Myford, Englehard, Jr., page 34).

The higher the second-scoring percentage, the more costly the scoring effort is with 100% second scoring being the most expensive. However, 100% second scoring provides a critical quality-management function, and it is often selected for writing or high-complexity projects.

Where 100% second scoring exists and the two scores are averaged or added, we would further advocate for quality metrics based on adjacent score agreement. By focusing on adjacent agreement, rather than exact agreement, states and consortia will avoid the pitfall of striving for an artificially high perfect agreement between two reads. The best score for a paper, for example, may be a “3.5” (by averaging a score of a “3” and a “4”), rather than forcing a score of “3” or “4.” In this case, a 3.5 reflects the fact that in the opinion of certified scorers, the paper has some characteristics of both score points.

Where 100% second scoring is used, states and consortia should also consider including both scores in the final score calculation, for example, (R1 + R2)/2 or R1 + R2 = student score). This provides more data to support reporting and analysis and more detailed feedback for teachers and students.

The scoring rules need to be established to balance the need for quality monitoring with cost and schedule considerations. Items with a rubric range of less than four points can typically achieve the desired quality metrics without costly 100% second scoring.

Scoring rules and final score calculations should be developed in collaboration with both psychometric and scoring experts during the test development phase. Scoring rules have a significant impact on the number of scorers required for constructed-response items, and ultimately the cost and duration of scoring.

Scoring Quality Practices

Scoring rules are one important aspect of scoring quality management, because the additional reads on a paper support the capture of inter-rater reliability statistics.

Another important consideration is the use of benchmark or validity papers. Validity responses are actual student responses, scored by experts, interspersed among live student responses. Scorers score these papers without knowing they are validity responses. While reliability statistics show the rate of agreement between two scorers, validity statistics show the rate of agreement between scorers and expert scorers who assign true scores on the validity papers. The College Board research study cited above (Wolfe et al) discusses the importance of both second scoring and the use of benchmark or validity papers to provide critical data on scoring accuracy.

Validity is used to monitor scoring in both professional-scoring and teacher-scoring applications. In both cases, feedback to the scorers on their validity performance is important to prevent scorer drift and promote scorer consistency. In addition to providing scorers access to validity and reliability metrics, annotated feedback on inaccurately scored validity papers can provide key remediation to scorers.

Other remediation strategies include the use of calibration (refresher training, particularly on score point boundaries), backreading by supervisors, and scorer messaging and intervention. Pearson's scoring system has built-in automation features that monitor scorer performance and send interventions and additional training to scorers falling below quality standards. This type of functionality greatly enhances scoring consistency and adherence to scoring quality goals.

As states and consortia consider local versus centralized management of the scoring effort, it will be important for them to consider best practices in quality management. Enterprise management has the key advantage of ensuring that every scorer receives the same cohort of validity papers, for instance, as well as standardized, consistent training, feedback, and remediation. Subtle variability in scorer-management practices from one state or district to another can lead to unintended variability in scoring outcomes. With current scoring technology, enterprise-management systems support local involvement, because distributed scoring systems allow "anywhere" access to management functions, including validity, calibration, and reporting. It will be up to the states and consortia to decide the right level of local participation versus local control, carefully weighing trade-offs in scoring consistency, schedule, and administrative burden for school, district, and state staff.

Return to Table of Contents

Summary and Recommendations

Test design and scoring design are inexorably linked. Test blueprints and scoring approaches need to be contemplated as part of the overall assessment design so that the solution produces valid and reliable results for the stakeholders, within the context of a scalable, practical, and efficient assessment delivery system.

To support state decision making, we posit the following summary recommendations:

- **Recommendation 1**—Collaborate broadly during test design and blueprint development with psychometric, content, scoring, and technology experts.

- **Recommendation 2**—Model the level of effort required for various test design alternatives, so operational implications, including estimated costs and schedule, can be analyzed during test development rather than waiting until operational planning.

- **Recommendation 3**—Design items, tasks, and rubrics in consultation with scoring experts, in order to appropriately minimize scoring complexity and cost without compromising on the quality and authenticity of the prompts.

- **Recommendation 4**—Assess carefully the role of teachers in the scoring process and the trade-off between teacher professional development and production scoring requirements.

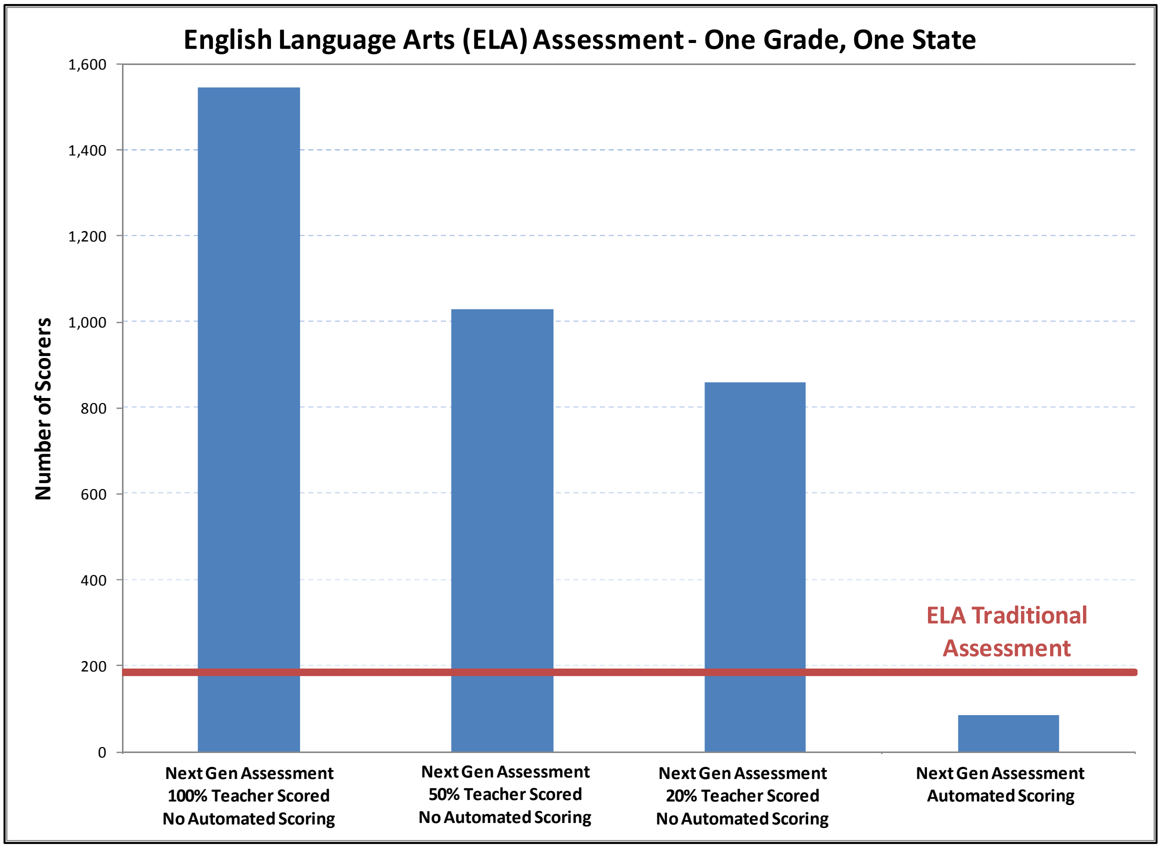
- **Recommendation 5**—Leverage technology to the greatest extent possible to enable the inclusion of more performance items and tasks without unnecessarily increasing cost or compromising quality.

For the purposes of illustrating the trade-offs and recommendations described in this paper, Pearson has created a modeling tool to look at the implications of various test designs, including comparing what a traditional test might require to score versus a higher quality, next generation assessment.

Figure 7 on the following page shows the scoring level of effort for a single grade of English language arts (ELA) for a single state. In this scenario, a traditional assessment has one extended-response writing item scored by professional scorers in a five-day scoring window. For this representative state, scoring would require fewer than 200 scorers to score the test (see horizontal line in Figure 7). On the other hand, for a next generation assessment with more complex items (in this scenario, two extended-response prompts and one performance task scored during a three-week scoring window), the number of scorers for that single grade could grow to more than 1,500, even more if the scoring window is compressed to five days or if scoring is done only by active teachers (who can score for a limited period each day, e.g., two hours in this example). By “active” we mean teachers who are currently working in the classroom (i.e., not retired or former teachers) and teachers who are required to maintain their existing workload (i.e., the district has not engaged substitute teachers to allow staff more time to score).

However, steps can be taken to optimize the scoring workload and reduce the total number of scorers needed. If professional scorers are used to complement and reduce the scoring burden on active teachers, or if districts use substitute teachers so that active teachers can score for more hours per day, the number of required scorers decreases. Figure 7 shows the impact of changing the mix of scorers to 50% active teachers and 20% active teachers. The scorer count further decreases if technology (automated scoring) can be used to apply the first score. Under this scenario, the number of required scorers for a next generation assessment falls well below that of a traditional assessment.

Figure 7: Scoring Level of Effort under Different Scoring Models



This chart illustrates a few of the points made in this paper, but all aspects of test design and scoring approach have implications for scoring feasibility, reliability, cost, and turnaround times. Each could be modeled in a similar way. Additional cost analysis considerations may also be found in a series of papers produced by the Stanford Center for Opportunity Policy in Education (SCOPE). Click here for more information.

A new day in testing is here, and states and consortia should be unwavering in their desire to produce much higher quality assessments, indicative of best practices around the world. To meet these goals, states and consortia do not have to make artificial compromises. By bringing content, psychometric, technology, and scoring experts together with educators at the start of the test-design process to model decisions and trade-offs, states and consortia can create a sustainable assessment system to support next generation needs, drive instructional improvement, and create the kind of rigorous teaching, learning, and assessment system our students and children deserve.

Return to Table of Contents

Web Links

For readers who choose to review this document on paper, we are including hyperlinked URLs in the order in which they appear.

Figure 8: Web Links

Reference	URL
"Considerations for Developing Test Specifications for Common Core Assessments"	http://www.pearsonassessments.com/testspecifications
Pearson next generation research	http://www.pearsonassessments.com/NextGeneration
Sample innovative items from Minnesota	http://www.pearsonassessments.com/MinnesotaInnovativeItems
Sample NECAP science inquiry tasks	http://www.ride.ri.gov/Instruction/science.aspx
CLA Performance Task Academy	http://www.claintheclassroom.org/
Ohio Performance Assessment Pilot Project	http://www.escofcentralohio.org/Pages/PBA.aspx
New AIMS rubric information	http://www.ade.az.gov/standards/aims/aimswriting/
SCOPE papers on performance assessment	http://edpolicy.stanford.edu/pages/pubs/pubs.html

Return to Table of Contents

References

Bay-Borelli, M., Rozunick, C., Way, W., & Weisman, E (2010). *Considerations For Developing Test Specifications For Common Core Assessments.* Retrieved from http://www.pearsonassessments.com/TestSpecs.

Beach, M. (2010, November). *AIMS writing assessment 2010*. Session presented at the *2010 Mega Conference,* Arizona Department of Education, Phoenix, AZ. Retrieved from https://www.azed.gov/asd/megaconf/WednesdayPDFs/W-210E.pdf.

Black, P. & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 139-148.

Council for Aid to Education. (n.d.). *Architecture of the CLA Tasks*. Retrieved from http://www.collegiatelearningassessment.org/files/Architecture_of_the_CLA_Tasks.pdf.

Darling-Hammond, L., & Pecheone, R. *Developing an internationally comparable balanced assessment system that supports high-quality learning.* Retrieved from http://www.k12center.org/rsc/pdf/Darling-HammondPechoneSystemModel.pdf

Duncan, A. (2010, September 2). Beyond the bubble tests: The next generation of assessments—Secretary Arne Duncan's remarks to state leaders at Achieve's American Diploma Project leadership team meeting. Retrieved from http://www.ed.gov/news/speeches/beyond-bubble-tests-next-generation-assessments-secretary-arne-duncans-remarks-state-l.

Lane, S. (2010). *Performance assessment: The state of the art.* (SCOPE Student Performance Assessment Series). Stanford, CA: Stanford University, Stanford Center of Opportunity Policy in Education.

Nichols, P. (2002, June 24). *Paper versus image-based scoring.* Presented at the Council of Chief State School Officers' Large-Scale Assessment Conference, Palm Desert, CA.

Pearson (2009, December 2). Race to the Top Assessment Program Public Input. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/CF3F0357-1B0E-4460-96DB-6F680994ADFC/0/RacetotheTopAssessment.pdf.

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

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

U.S. Department of Education, Office of Educational Technology. (2010). *Learning powered by technology*. Retrieved from http://www.ed.gov/sites/default/files/netp2010.pdf.

Wolfe, E., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning, and Assessment, 10(1)*, 16-17.

Wolfe, E., Myford, C., Engelhard Jr., G., & Manalo, J. (2007). Monitoring reader performance and DRIFT in the AP® English literature and composition examination using benchmark essays (Report No. 2007-2). New York, NY: The College Board.

Return to Table of Contents