Thoughts on an Assessment of Common Core Standards

Stephen Lazer, Vice President, Assessment Development, ETS

John Mazzeo, Vice President, Statistical Analysis & Psychometrics Research, ETS

Jon S. Twing, Executive Vice President, Assessment & Information, Pearson

Walter D. Way, Senior Vice President, Psychometric & Research Services, Pearson

Wayne Camara, Vice President, Research & Development, The College Board

Kevin Sweeney, Executive Director, Psychometrics, The College Board

Preface

ETS, Pearson, and the College Board have formed a collaboration to explore how innovative approaches and best practices in high-quality assessments can be applied to the creation of a common assessment system. Our objective is to work with states to develop an assessment system that will improve learning. We propose to design an integrated system that can provide accountability data, instructionally actionable information, and can inform teacher professional development and evaluation. Combined, we have extensive experience in the research, development, and delivery of a wide variety of assessments. We have worked within and across all 50 states and have worked together collaboratively for many years. Our expertise includes the development of innovative computer-based assessment systems and student growth measures, and the application of a wide range of item types and scoring approaches to provide timely feedback to teachers and students.

This paper is an attempt to raise key assessment design questions and discuss some ideas for a systematic high-level assessment design that satisfies many of the needs expressed by stakeholders. It is meant only to begin discussion, and not to serve as a firm and fixed recommendation.

Introduction and Summary

American educators stand at a moment of unprecedented opportunity. With opportunity, however, comes risk: decisions we make may well affect the course of assessment in the United States for years to come. Advances in technology, coupled with innovative assessment task design and advanced psychometric and cognitive models, make it possible for us to obtain a richer, more intelligent, and more nuanced picture of what students know and can do than ever before. While the historical opportunity to change the direction of education is real, so are the challenges inherent in any change in assessment paradigm. At the heart of this challenge is one point that is too often missed in these discussions: Different stakeholders will set diverse priorities for an assessment system. Some of these stakeholders value snapshots of what students know and can do at fixed points in time and consider the use of these data for accountability purposes as the highest priority. Others value obtaining multiple points of data that can be used to evaluate schools and teachers systemically. For some, instructionally actionable data at the student level for the purpose of improved instruction is the main system goal, while others are more interested in data at higher systems levels for auditing or "return on investment" type of decisions. Most want formal assessments to be as short and inexpensive as possible, while others would trade some cost and time efficiency to have more authentic, complex, and reliable tasks. Some stakeholders require data that are unambiguously comparable across states and districts, while others would rather see some substantial state and local control over the content of assessments.

No single assessment, not even an integrated assessment system, can optimally serve all possible purposes. Any assessment design is therefore a compromise. Tests that provide optimal instructional feedback may not be the best way to get an overall snapshot of what students have learned over the course of a school year. The need for formative information is not necessarily consistent with the need for data that can be used to evaluate teacher or school effectiveness. Tasks that model good instruction are not always consistent with desires for tests to be as short as possible and for scores to be returned immediately. The desire for comparability of data across jurisdictions stands in tension with wishes to allow those jurisdictions and their teachers and curriculum specialists substantial and variable input into the form and content of assessments. The need for low operational cost may be at odds with many other goals of the system. Efficiency in the long term involves investments in technology and human capital in the short term.

Policymakers should consider the three principles following from this discussion:

- First, we should think of systems of assessments rather than individual tests, as this is likely the only way to satisfy the various information needs identified by stakeholders.
- Second, we are at a moment when new technologies and assessment methodologies provide us an unprecedented opportunity to satisfy many perceived needs in a carefully structured integrated system.
- Third, we must realize that, even in a complex system, we will need to choose among competing and conflicting priorities.

This document represents an attempt to create a high-level framework for an assessment of common core standards. We arrived at this framework in the following way: First we considered a series of questions regarding the likely design requirements of such an assessment system. Then we considered various factors and made judgments about competing priorities. This led to a high-level assessment model, along with a discussion of various matters that require further research and more thought. Different decisions about priorities would certainly result in different assessment designs, and we tried to point out places where alternate decisions might have such impact. For this reason, this document is meant to begin a conversation about not only these priorities but all aspects of such an assessment design and is not intended to provide the answer or solution. This is also meant to be a high-level design document. We will prepare additional documentation that will discuss, in greater depth, topics such as elements of the assessment system that are designed to provide instructionally actionable information, exercise types that can be used, how scoring might be accomplished, the special needs of high school testing, the assessment of students with disabilities and English-language learners, and how the assessment system might measure student growth.

Executive Summary

The bulk of this document describes how we answered the key design questions and explains our suggested assessment framework. Before moving to this discussion, we have included an executive summary of what we believe to be key design elements of a forward-looking assessment system:

1. The educational system needs both accountability and instructionally actionable data, and no single test will be optimal to provide both. Therefore, we believe that the goals of this new effort will be best served by an integrated assessment system that includes summative and formative or interim elements built to a common framework. If the American Recovery and Reinvestment Act (ARRA) funds support only the development of the summative elements of the system, we should ensure that the system and system infrastructure are designed to work with formative and interim elements designed and developed by others.
2. The system must measure common standards and must allow for state-to-state comparability on the common standards. To accomplish this, the new summative measures should have a set of common components assessing the common standards, and produce scores and performance indicators that are comparable across states. However, the system should also allow states to augment this core with materials of their choosing to produce separate state-specific information.
3. The summative portions of this battery will need to include, at a minimum, end-of-year tests for grades 3 through 8 in both math and English-language arts (ELA) at the elementary and middle-school levels. At high school, the system may include either "end-of-domain" or "end-of-course" assessments. The elementary and middle-school tests should support growth modeling and across-grade comparability. The assessments should also support within-grade proficiency standards. While we believe these end-of-year and end-of-course/domain assessments should be part of the system, we also believe we should consider using data collected over the course of the year as part of the summative system (see point 9 on page 6).
4. Assessment designers will likely need to incorporate international benchmarking and facilitate comprehensive alignment efforts, although the methods for accomplishing these goals have not yet been determined.
5. The tests should be delivered on computer or other similar technology. Student mastery of emerging standards can likely not be measured based on paper assessments alone. Further, summative assessments should make use of adaptive administration, although adaptive models will need to make allowances for the full range of item types needed to measure emerging constructs, including those that will be scored by humans. We envision that such a system will ultimately support the on-demand needs of a personalized education system.
6. The development of assessment tasks will be based on an Evidence-Centered Design (ECD) process that involves experts and stakeholders. To measure the intended constructs, the tests will likely need to use a range of tasks and stimulus materials, and will need to include more than traditional multiple-choice questions. Important decisions will need to be made regarding how constructed-response questions are scored, though we picture a mixed model that uses technology and professional (e.g., teachers and other subject matter experts) scoring that is supported by assessment technology infrastructure. Such a system will also provide opportunities for professional development.

7. Compared to current summative tests, items and tasks should be created based on an improved understanding of learning and development, both to promote better interaction with formative elements of the system as well as to provide models consistent with good instruction.
8. Tests should be as accessible as possible to students with disabilities and English-language learners, and designers should make use of technology to improve such accessibility.
9. Certain forward-looking ideas should be considered that may or may not be ready for operational implementation at the time of initial rollout of the new system. Perhaps most important among these considerations is that summative assessments may not be single-testing events but could augment end-of-year tests with data collected over the course of the year.
10. We should have careful plans in place to validate assessment scores and claims made based on them, as well as a long-term research agenda to continuously improve the efficacy of the assessment system for its intended purposes.

The pages that follow detail the process through which we arrived at the general parameters listed above.

1. Should we consider the test of common core standards as simply a summative assessment or as part of an integrated system that involves interim and/or formative components as well as summative assessments?

As previously stated, no single assessment can be optimal to serve all possible needs. It is possible that the United States Department of Education (USED) will use the Race to the Top (RTTT) grants to focus on the development of summative assessment systems. Summative assessments will remain a key element of an educational quality-management system, and one of the main goals of this effort is to improve the quality and efficiency of our summative systems. However, without questioning this goal, we believe that American education would be best served by an integrated system where summative and interim or formative components are built from common frameworks and cohere as an information provision system. The system, taken as a whole, should provide both accountability and instructionally actionable information without unduly or unrealistically burdening any given component (for example, summative tests should not be expected, on their own, to provide in-depth instructionally actionable data). It is not necessary for the USED common assessment grants to pay for the development of formative elements. It is essential that the summative systems be designed to work in tandem with these formative elements.

There are a number of reasons to favor an integrated system. First, formative and summative components will likely both function better if built to work together. Specifically, they should be built to meet the same skills standards and to a common assessment framework. They should be constructed using open technology standards and assessment frameworks so that material can flow from one set of instruments to others. Second, an integrated system should relieve pressure from the summative tests to serve a purpose for which they are not ideally suited: to provide in-depth, reliable, and valid instructionally actionable data. This is particularly true at the level of individual standards, where coverage on any summative test will be, by necessity, limited (even in cases where, as we propose, flexible or adaptive administrations or multiple administrations throughout the school year can be used to get better information at this level). Attempts to provide such data from a summative test will increase pressure to lengthen tests—pressure that will become especially important since we believe the system should exploit technology for delivery. An integrated system should prove far more likely to meet the varied goals people have set for the assessment.

While the ability of summative measures to provide formative data is limited, one could, in a carefully designed and integrated system, view summative assessments as providers of information to formative systems, particularly for students who have “outlier performance” in some area. In these cases, summative data might focus teachers on areas where more testing or diagnosis seems indicated. This could involve thinking across grades. For example, a summative result at grade 5 could identify students who appear to be struggling in certain areas. Based on

the specific nature of the results, the system might identify "diagnostic intake test" components that would be administered at the beginning of grade 6. These would not go to all students but only to those whose grade 5 results had indicated the need for further testing.

There are, of course, a number of different models for how an integrated assessment system might provide instructionally actionable information. An integrated system can include formal elements like interim assessments, which are given throughout the year to get a snapshot of how students are doing in mastering the required skills, or diagnostic adaptive assessments, which provide more in-depth information on the gaps in student learning or performance. Both components could utilize banks of performance tasks/assignments and scoring rubrics available for teacher use. While this paper focuses on summative elements of the new system, we plan to address different models of providing instructionally actionable information in a future paper. However, any of these models assumes certain educational system requirements, including the ability to deliver various assessment components via computer, an automatic way of linking assessment results with enrollment and teacher information, and a series of connections between assessment results and curricular materials.

Formative assessment components of an integrated system may be excellent areas to allow for customization, differentiation, and local education agency involvement in development. While there are common standards, to the extent that districts and states use different curricula to address the common standards it is possible that they will prefer to incorporate different formative systems within their instructional programs.

As mentioned previously, this paper focuses on summative components of the assessment system. One open question is whether accountability data will come solely from single summative tests, or whether data gathered over the course of the year can be part of a formalized accountability system. In the latter case, we can possibly increase the amount of instructionally actionable data that comes out of summative systems (although not to the point where it obviates the need for formative systems) and improve the quality of the summative data. This will be addressed briefly below and will also be the subject of a follow-up discussion.

2. What sort of general design should the assessments that make up the summative system have?

We believe these tests should have at least two major components, although it is likely federal funding will address only the initial one. Our understanding is that states may augment the common core standards with 15 percent of their own standards. Thus the common core assessment system must provide data on the common standards that are strictly comparable across states and must allow states to measure state-specific content as needed.

Because there will be both common core standards and state additions, the tests would likely have at least two major components. The first would be the test of common core standards. This would be consistent across all participating states, districts, and schools. Note that we do not mean the same exact test form is required but rather the same assessment. The common components of the test will be designed to yield state, district, school, and individual results on the common core standards and will not include state-specific augmentation. The second

component could be composed of state-specific content or augmentations. Such augmentations could focus solely on the up to 15 percent of unique state-specific standards that are in place or provide additional measures or coverage of common core standards. These augmentations would be analyzed in tandem with common core items to yield state-specific results.

Why do we believe that the common-standards components of the summative measure should not be customizable, and that state choices should be located in state-specific sections? Comparability of results on the common core standards and test development efficiency will be high priorities of the system. Comparability across states and the economies of scale will be enhanced if there is a common assessment of the common standards. Other designs are possible if the ability of states to customize the common core assessment is viewed as desirable, but these will likely threaten comparability of results and will lead to higher cost.

In system terms, the approach we recommend means adopting a single national delivery package and permitting states (or groups of states) to add components as needed, as opposed to "opening up" the common materials for each state. Finally, this approach allows some states to decide they do not need state-specific content, without affecting the comparisons on the common components (which embedding items in the common core would risk).

This approach has other advantages: Even if a single consortium develops the common core assessments, states would be free to work with whomever they wished for state-specific components. If developers of the common core components of the system were to work to some open and shared standards for test material, packaging, and delivery, all components could be delivered as a single test by any number of assessment-delivery systems. Alternately, the developers of the common core assessment could build some special components that could be used at state discretion.

Note that in any of these models, provision will need to be made for field testing new content. For the common components, this could either be accomplished through a variable section or by embedding field-test items within operational sections.

One open question is how big a system (in terms of assessment exercises) would be needed to ensure security. The answer will depend on the length of the test window, which in turn depends on the number of students who can be tested at any time. It will also be affected by the rapidity with which test developers can rotate content, or the number of different aggregations of content we can provide.

A second open question concerns the length of the individual tests. It is likely that tests at grades 3 and 4 will be limited to 50 minutes, while tests at grades 5 through 8 will take 60 – 120 minutes (for both common and state-specific components). High school tests could, conceivably, take between two and three hours. If extended tasks are used, assessment time may need to exceed these limits.

3. What grades and subjects?

We assume that the summative assessment system will include end-of-year ELA and math tests at grades 3 through 8, all of which need to produce individual scores as well as aggregate scores

and will need to work together to track student growth. As discussed under point 9 on page 16, these end-of year tests may not be the only components of the summative system. At high school, we believe two summative models are possible: either end-of-domain tests in both ELA and math that cover the knowledge and skills needed to be ready for college and career training, or a series of end-of-course tests. Each approach has advantages and disadvantages, depending on the priorities selected.

Annual testing between grades 3 and 8 will be an optimal way to support student growth modeling, which we believe to be a key goal of the new system. It also provides data at fixed points, which should be usable by parents, teachers, and policymakers.

One assumption we make is that these tests could replace the current generation of No Child Left Behind (NCLB) assessments. Through use of technology, we believe we will be able to provide a state-of-the-art range of accommodations to students who need them. We also believe that through use of computer administration, we may be able to tailor tests to individual students. Such personalized assessment may even cause us to reevaluate the need for modified (or "2 percent") assessments. Additionally, it would be appropriate to think of 1 percent or Title 3 tests as part of a common assessment system that shares data among components.

Closing comments in this area: End-of-year testing at grades 3 through 8 is likely necessary given an educational system that is still organized by grade and which needs annual accountability data. However, just because students are "housed" into educational institutions based on this classification system, it does not mean that this should restrict how we teach and assess these students. For example, the system we propose here could evolve into an on-demand system that will make sense as school schedules and student needs continue to evolve. It also would allow for a system in which students take tests when they are ready based on their personalized instructional paradigm. Second, as mentioned above, one could consider systems in which accountability data are not solely the province of the end-of-year test (see point 9 on page 16). This would not, of course, necessarily obviate the need for the end-of-year snapshot of what students know and can do.

4. Cross-grade or within-grade scaling and reporting?

Given the overall interest in student growth metrics (and the use of such metrics in teacher evaluation), the assessment should support cross-grade comparability, and the assessment will need to be set up to allow for such comparisons. This work will, of course, be greatly facilitated if the content standards and expectations are coherent across grades. In addition to supporting growth modeling, cross-grade comparability facilitates another element we view as desirable in the system: the ability of flexible administration engines to select "out-of-grade" content for either advanced or struggling students. We assume that this out-of-grade content will mirror the instruction the student has received regardless of his or her grade level or age. Note that use of off-grade content is forbidden under current rules of NCLB, and USED would have to facilitate dispensation.

While we believe we need cross-grade comparability, we will also need to have within-grade performance levels. This does not pose a problem but simply must be considered as part of the work planning.

There are interesting questions that will need to be answered in this area. For example, while it is likely that some constituents will want to see tests at grades 3 through 8 on a vertical scale (perhaps mistakenly thinking vertical scales are required for growth measures), it is not at all clear that high school tests should (or need to be) placed on such a scale. Frankly, the notion of comparing performance in various high school subjects, such as chemistry and Algebra II, is problematic in itself. In the past, states have not tended to require this, and high school content may not be as friendly to cross-grade comparability. But there is a real need for data on whether or not high school students are proceeding as necessary.

It is worth mentioning that there are several ways to produce measures of growth and cross-grade comparability. How the requirements of specific growth models affect the system will need to be studied, and we plan to devote more thought to this topic as follow up to this paper.

Two closing points: First, the need for cross-grade comparability is likely to be required for the common core standards. State-specific augmentations may or may not need to support such cross-grade comparability.

Second, given the number of standards and the pressures on assessment time available, it would make the most sense from a measurement standpoint to establish any passing scores on the summative system as a whole and not just at the level of specific standards. We will almost certainly need to produce sub-score and collateral information as well as disaggregated performance by standard (and other breakouts), and the presence of an underlying comparability paradigm would facilitate all these purposes. Such system wide comparability may also be used to guide any adaptive administration and an integrated system to improve the quality of the standard-level data. Reporting meaningful information at the standard level will become easier if new standards are fewer and more cognitively distinct.

5. National or state-specific scales and performance levels?

The system must support both common and state-specific performance levels. A comprehensive system might work as follows: There could be a single-scale score and a set of achievement levels on the common test component. This would allow for comparisons among participating states and placement of individual scores in the context of the common standards. Recall that this is possible because each state in a consortium is taking the same assessment on the same standards.

The common core standards assessments will likely need to be internationally benchmarked. The easiest way to accomplish this is through judgmental processes: either through the use of the internationally benchmarked standards as key descriptors of goals in a level-setting process, or through some assurance from an independent body that the standards themselves conform to international best practice and that the assessment is aligned with the standards. Alternately, the system could rely on statistical linkages to international studies such as Trends in

International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). Regardless, a key step involves meeting with stakeholders to determine the specific uses stakeholders wish to make of the international benchmarks.

This paper assumes that the new assessments will have performance standards. Therefore, using appropriate methods and sources of information to set standards will be of key import. Standard setting is often not considered when designing an assessment, but the validity of claims made based on the assessment will be no stronger than the performance standards allow. Assessment designers should ensure that crucial evidence is brought to bear regarding topics such as what successful students around the world know and can do in different grades, and what sorts of texts should students be prepared to encounter to succeed at the next grade. Overall, we should have a solid evidentiary basis for stating that students have reached a level that will allow them to succeed in future education.

The comments above relate to the scale and performance levels for the common core components of the assessment. In addition to this, there will need to be separate state-specific scales and levels for states that augment the common core with their own materials. In all likelihood, these would be based on state-by-state analyses of the conjoined sets of items (that is, common plus state specific). In practical terms, it may be hard for states to explain major differences between their standards and national standards. But the system needs to support these types of data.

6. The use of technology in delivery

One of the major questions facing the designers of a common-standards assessment is "how much technology, how soon?" Certainly, the current state of technology availability in many states and the current price structures of testing programs would argue that an assessment system should offer a paper-based test, or at least a program that could be administered on paper as well as online. In spite of this, we believe that the assessment of common standards should be computer-based (or other technology-enabled) tests in which paper is used solely for certain special accommodations. There are several reasons for this:

- Emerging standards in both mathematics and ELA define constructs that can only be measured through the use of technology. This is likely to be true in subjects such as science as well. Maintaining parallel paper and computer systems on which results were supposed to be interchangeable would effectively prevent measurement of such skills. This "assessment tail wagging the education dog" has been a large criticism of education reform efforts in the past, and we want to avoid this.
- Technology allows for the use of a range of forward-looking exercise types, including item types that ask students to engage with digital content and formats, and bring to bear skills that wouldn't (and couldn't) be invoked on a paper test.
- Testing some skills on paper may simply yield invalid results in the future.
- Technology allows for flexible (adaptive) and on-demand testing, which we believe should be a part of this design.
- Technology allows for electronic scoring of some sorts of items, and thus for use of a broader range of items than does paper-based testing. Technology also facilitates the

distribution of student responses to teachers, monitoring the quality of teacher scoring, and increased opportunities for professional development in terms of assessment development and scoring.

- Rapid return of scores and seamless data/information interchange is facilitated by technological delivery.
- It is easier to see the summative test (or tests) as part of an integrated-assessment system if it is built around a technology platform based on accepted standards for content and data transfer.
- We assume technology will continue to improve, become easier to use and more common in the future such that our proposed system will be operationally feasible.
- Technology allows for provision of a range of accommodations for students with disabilities and English-language learners that might not otherwise exist.
- Using technology as the single delivery paradigm simplifies issues with comparability.

This decision, of course, has major operational implications. Even with expanded technology access we cannot rely solely on mass administrations, so scheduling becomes essential. Testing windows will need to be open long enough to accommodate test takers, and exercise pools will need to be large enough to protect test security. The final system must allow for trade-offs between assessment purpose (like high-stakes graduation decisions) and the size of the testing window allowed. Finally, since it is likely that state-specific content will be developed by a number of different entities, we would need a set of data transfer and delivery protocols that could be used by all involved.

As mentioned previously, we believe that the summative-assessment system should make use of adaptive administration. A variety of approaches may be used for this purpose (e.g., traditional computer-adaptive testing, multistage testing, variable or fixed-length testing). The appropriate adaptive testing solution will depend on the content and structure of the exams.

Some arguments in support of adaptive testing follow:

- It allows for on-demand testing.
- It allows for somewhat shorter testing times than linear testing, which helps from various perspectives, particularly if access to computers is an issue.
- It allows us to measure the “higher” standards, while at the same time gaining some meaningful information about what lower performers know and can do.
- Considered appropriately, it may allow us to identify standards on which students are struggling without unduly lengthening tests. Particularly in ELA with a heavy emphasis on authentic reading, we believe variations in traditional CAT approaches (e.g., section-based or passage-based adaptivity) can be implemented in an advantageous manner. Again, this will allow for far more personalization than traditional assessments.
- It will allow us to get better “bang for the buck” out of open-ended/performance-based testing.

One possible challenge is the use of items that require human scoring in an adaptive system. There are in fact ways to use such items. In a multistage system, for example, routing decisions

can be made based on a machine-scorable stage, with performance or open-ended exercises requiring human scoring administered during later stages.

While we believe the assessment should be adaptive, it is not certain we will be able to make it adaptive in the first year of administration. We would, of course, do large-scale piloting of items before roll-out. However, given issues associated with calibrating a pool under sub-optimal motivational conditions, it is likely that in the roll-out year of the program we would assemble a large number of linear tests and assign these randomly to candidates. The system could, however, use adaptive administration in subsequent years.

7. What item types should we assume?

This question is in many ways premature: Final internationally benchmarked standards do not exist at all grades. Decisions about the sorts and arrays of tasks that ought to be included on these assessments should be the result of a careful Evidence-Centered Design (ECD) process in which we gather expert groups, review research, and identify the sorts of behaviors that would convince us that students have reached the stated standards. Simply stated, we want to use the assessment task or item that most appropriately measures the construct desired.

However, we need working assumptions. Our task design should be guided by the general goal of measuring each construct as validly, effectively, and thoroughly as possible. This will certainly involve a range of exercise types that move well beyond traditional multiple choice. These may include, though not be limited to, scenario-based tasks, long and short constructed responses, tasks that involve the exercise of technology skills, and simulations. This is particularly true given the general goals of providing college readiness information, eliciting more than content mastery information (i.e., problem solving and critical analysis), and exploiting the assessment medium (namely online technology).

To optimize the speed and cost-effectiveness of scoring these items, we should be prepared to adopt a range of strategies. First, we may need to push the limits of what can be scored electronically: machine scorable must not equal multiple choice. Computerized-scoring systems are getting more effective all the time. Second, we can and should develop better ways to analyze data obtained from simulations that go beyond simple student responses. Third, while some tasks can be machine scored, we must realize that emerging standards will likely necessitate the use of items that, given the current state of scoring technology, will require human scoring for some number of years. If this is true, we will have to find ways to balance the need for these items with other imperatives. We will also need to make effective use of technologies for distributing responses for scoring, and for monitoring and assuring the quality of such scoring. To summarize, we believe it is likely that the new assessment system will need to make use of three types of scoring: simple-machine scoring using online testing, intelligent scoring using online technologies, and human scoring using online technologies.

Human scoring is, of course, in many ways a positive. It allows items that are not constrained by limits of the current electronic-scoring systems. Use of teachers in the scoring process would also represent a powerful professional development activity. Teacher scoring in a system that will also be used for teacher evaluation will necessitate careful safeguards. Therefore, any final

design will need to find ways to use human-scored items in ways that optimize the instructional and professional development impact of those items, without placing undue or unrealistic burdens on the system. We should also be prepared to make aggressive use of emerging computer constructed-response scoring technologies, to make sure that teacher involvement is in fact professional development and not solely additional labor. We believe there are ways to involve teachers in scoring, without necessarily expecting them to conduct all the scoring (at least of the common core standards components that require rapid score turnaround). The good news is that much progress has been made recently in using automation in human scoring in ways that improve quality and professional development potential.

During the design effort, other questions will emerge about the sorts of items and tasks that can be used. These will surround issues like use of audiovisual stimuli (as called for in the Council of Chief State School Officers-National Governors Association ELA standards), as well as interactive tasks involving spreadsheets and databases. One interesting matter that will need to be resolved early in the process concerns the inclusion of tasks that measure ELA standards for speaking and listening (if these are in the final version of any set of standards). This is not uncommon in current state standards, but these skills are rarely if ever covered in assessments (which are normally limited to reading and writing). We will need to decide how to assess in these areas as this has broad implications for test design and administration. One possible approach is to include listening and speaking in the individual score portions of high school tests (which can be longer), and only assess these skills at state discretion in tests at earlier grades depending upon the goals of assessing listening and speaking or the outcome measures desired in these domains.

If we are to do something new and different, it is necessary that our items and tests be developed with an awareness of how students learn. A test built around an understanding of available learning progressions is likely to be a better provider of information to formative components of the system. Items that model good learning and instruction should make "teaching to the test" less of a problem. Of course, this sort of thinking cannot mean that we fail to meet psychometric standards for quality, score comparability, and fairness, particularly given the high-stakes nature of the potential use for high school graduation, college readiness/college placement and possibly college admissions. Finding the appropriate balance will be key.

8. Pre-equating or post-equating?

Given the discussion immediately above (that is, a desire to use adaptive testing), one might assume we would also recommend a pre-equating approach. It will certainly be necessary to calibrate the items to allow routing decisions. But, if the testing windows are at all long, and vary by states, post-equating might make some states wait rather long for scores. Therefore, we believe the system will eventually need to be geared toward pre-equating as allowed. One complexity associated with pre-equating, however, is the use of human-scored items. Pre-equating will only work if we can ensure that the scoring of the responses is of the same effective rigor as that used to calibrate the items; this will require very careful control over the human-scoring process.

Finally, it is almost certain that some form of post-equating and post-calibration will be needed during the first year of the program.

9. Should the summative assessment be a single test or use multiple sources of data?

In the previous sections, we have for the most part discussed the tests as if they were given at fixed points during some course of study (either the end of a school year or the end of high school). Furthermore, we believe that such tests should be part of any coherent system of assessments. However, this is not the same as arguing that they should be the *only components* of a summative system.

There are several ways in which one could consider other "assessment events" or data sources to be formalized parts of the summative-assessment system. In one family of approaches, there would be multiple assessments over the course of the year whose results would be aggregated into a summative score or scores. Such an approach could conceivably take one of two general forms. In the first, a larger assessment that would theoretically cover the entire year would be broken into component pieces covering different, and possibly non-overlapping, sets of content and skills. For example, a three-hour test might be broken into three one-hour tests that would be given over the course of the year. In this conception, the end-of-year test would essentially cover the last third of the year. A similar possibility is to build assessments around discrete instructional units (even if those were not equally spaced over the course of the year).

A variant on this approach is a system in which the end-of-year test did cover the entire year's worth of content, but that earlier standardized tests covered content from the first part of the school year in more depth. This is similar to the "midterm-final" approach used in many universities and high schools, in which scores from midterms and finals are averaged according to some preset weights and often combined with other information to derive a final grade.

There are obvious advantages to such approaches and real challenges as well. On the plus side, one would get some early-warning data on students from the summative system itself; students might be able to retake modules they have failed over the course of the year. Because such systems would allow more aggregate data, they might give more stable results. On the other hand, the challenges are real. Such a system almost certainly involves making decisions about the ways content and skills are to be ordered (or at least combined) in the curriculum, and this may be beyond what is possible. While the aggregate data may be solid, the reliability of the periodic measures may be lower than one might like, which will be a problem if those data are used on their own for high-stakes purposes. Finally, in the second of these models, the system would need to be prepared to deal with a possible conundrum. If two districts got the same average scores on the end-of-year test, that would normally be interpreted to mean that those two districts ended that school year "in the same place." Rating one district higher because of performance on intermediate ratings might be problematic.

An alternate model, used in some other countries, is described below. There would still be an end-of-year test, but accountability scores would also use data from standardized projects conducted over the period of the course of study (for example, research papers, laboratory

reports, or book summaries). Scores from these projects would represent a fixed percentage of the final summative score.

This model would have clear advantages and disadvantages as well. Through making these sorts of tasks part of a formal accountability system, it encourages the use of tasks that are elements of good instruction and learning. In addition, this approach avoids the problem that usually keeps these sorts of tasks out of large-scale testing: they simply take too long to be included in a fixed-event assessment. These kinds of tasks might also provide a logical place to rely on teacher scoring and to enjoy the professional development benefits attendant upon it. Finally, centrally designed tasks and scoring guides may be able to mitigate certain comparability issues.

There are a number of issues that would need to be addressed in making such a system operational. It would need mechanisms for ensuring that students themselves completed the tasks. While steps might be taken to standardize task protocols and scoring rubrics, short of adoption of a common curriculum, some choice of tasks would need to be provided at the local level. Even with the best safeguards in the world, such choice, combined with local scoring, will almost certainly call into question the strict comparability of results both over time and across jurisdictions. This is not a reason to reject such approaches, but rather represents the sorts of trade-offs that must be considered carefully and suggests the sort of research that is necessary. It may be possible to find interesting compromise positions: we might conceptualize an accountability system in which not all data elements are used for cross-jurisdiction comparisons, for example.

The use of assessments or projects conducted over the course of the year as part of a formal summative-assessment system is a major and important idea. There are challenges to be met before such a system could be implemented, and the existence of such a system presupposes infrastructures for data maintenance and transfer that are currently beyond the scope of many states. Thus it is possible that these assessment features will begin as part of the state augmentations described above, until such time as they can be added to the accountability system. We believe that strong, forward-looking end-of-year assessments will be part of the system. We also believe that they may not be the only elements and that the system available on day one may not be the final system. We will consider this more thoroughly in follow-up discussions to this paper.

10. How do we help ensure that the assessment results validly support claims being made about students, teachers, and schools?

We must consider the need for provision of research evidence that supports intended uses of scores from the assessment system. Even if we start with internationally benchmarked standards, we will need an ongoing method for checking and updating these standards, and for making attendant changes to test specifications. We may also not be able to simply rely on those standards: Since the high school tests will claim to measure college readiness, we should plan to have some data validating that claim. There are various ways to obtain these data; the key point is that some plan to gather validity data should be part of the design from the beginning. Discussions of validity data are beyond the scope of this paper; we will come back to this topic in a later paper.

Conclusion

We stand at a moment of unprecedented opportunity. Improvements in methods and technology, possible agreement on a set of common standards, combined with a generous commitment of federal resources, should allow us to build assessment systems that provide accountability data and instructionally actionable information. However, these opportunities will surely be wasted if we do not carefully consider the trade-offs inherent in any large-scale assessment design. We must, and can, ensure that a new generation of assessments is innovative and meets all pertinent psychometric standards for quality, fairness, and best practice. This paper represents a first attempt to consider the trade-offs and to set up a "straw design" consistent with those trade-offs.

While there is reason for caution, the opportunity far surpasses the potential problems. We believe that we can create a summative assessment system that uses innovative exercise types and computer adaptive delivery to measure depth of student understanding and track student growth. The system can be designed in ways that allow it to work hand-in-hand with formative assessment elements to produce instructionally actionable data. We can provide solid data on common core standards while giving states a chance to add their own augmentations. We can do this in a way that is operationally and economically feasible.

ETS, Pearson, and the College Board are excited to be part of the national discussion of new assessment systems. This paper represents an attempt to begin discussion by laying out key questions and central elements of a possible assessment system. We plan to write further papers examining specific topics in more depth. We hope others will join in this conversation: only through open communication will the country build the assessment system it needs.