Executive Summary/Introduction

To ensure that your input is fully considered, we urge you to identify clearly the specific question, purpose, and characteristic that each of your suggestions addresses and to arrange your submission in the order of the questions listed later in this notice. Please also include a description of your involvement, if any, in statewide assessment practices.

Educational Testing Service (ETS) respectfully submits this response.

Description of Involvement in Statewide Assessment Practices

ETS has been involved in K-12 assessment for decades. At the federal level, we have held contracts since 1984 to develop, administer, and report the National Assessment of Educational Progress (NAEP). Under contract to the College Board we also develop national assessments that play important roles in K-12 education. These include the Advanced Placement Program® (AP®), the SAT®, and the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT®). ETS develops national-level assessments for other clients, most notably for the Educational Records Bureau and the Southern Regional Education Board. Moreover, at the state level, we are either developing or have developed statewide assessments for California, Florida, Georgia, Indiana, Maryland, Mississippi, New Jersey, Puerto Rico, Tennessee, Texas, Virginia, and Washington. In California, we also work with the California State University system to develop the Early Assessment Program (EAP), a college readiness supplement to end-of-course components of that state’s assessment system.

ETS provides a range of services to states and other clients. These include psychometric research and statistical analysis, assessment development, program management, production and delivery, communications, and policy analysis. The types of assessments for which we contract with states include No Child Left Behind (NCLB) summative assessments, along with their alternate and modified versions; high school end-of-course assessments; high school exit examinations; and Title 3 English language proficiency tests.

In the pages that follow, we provide one set of possible answers to the questions raised in the U.S. Department of Education’s (the Department’s) Notice of Public Meetings and Request for Input on the Race to the Top assessment program. Before beginning a direct response to these questions, we offer a brief introduction.

Introduction

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 historic opportunity to change the direction of education is real, so too are the challenges inherent in any change in assessment paradigm. At the heart of these challenges is one point that is often missed: 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 they 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” types 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, local education agencies (LEAs)/districts, schools, and children, 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, therefore, is 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 conflicts 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 four principles following from this discussion:

» First, we should think of systems of assessments rather than individual tests, as this approach 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 with 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.

» Fourth, we must stage the creation of the new assessment system to accommodate reality, because even if we know what we want to do and how we want to do it, the existing assessment infrastructure in the U.S. is a limiting factor in implementation.

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 the general requirements and desired characteristics of such an assessment system. Then we considered various factors and made judgments about competing priorities. Different decisions about priorities would certainly result in different assessment designs, and we endeavored to point out places where alternate decisions might have such impact. Ultimately, some areas require further research and more thought.
Finally, we defined the desirable system as “Generation 2” and recommended a “Generation 1” transition system to achieve many of the goals of the ideal system sooner than would be possible if we waited for all elements of Generation 2 to be feasible. Because of all these considerations, it is important for readers to understand that this document is only one of a broader set of possible answers, and is meant to inform the Department’s thinking rather than to propose a single path forward.
1) Propose an assessment system (that is, a series of one or more assessments) that you would recommend and that meets the general requirements and required characteristics described in the notice. Describe how this assessment system would address the tensions or tradeoffs in meeting all of the general requirements and required characteristics. Describe the strengths and limitations of your recommended system, including the extent to which it is able to validly meet each of the requirements described in the notice. Where possible, provide specific illustrative examples.

We believe the following 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 of 2009 (ARRA) funds support only the development of the summative elements of the system, the Department should ensure that the system and system infrastructure are designed to work with formative and interim elements that are 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 assessments for grades 3 through 8 in both mathematics and reading/language arts. At high school, the system may include either “end-of-course” or “end-of-domain” assessments. The elementary- and middle-school assessments should support growth modeling and across-grade comparability. The assessments should also support within-grade proficiency standards. While we believe that 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 below).

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. However, the technology to effectively administer computer-adaptive tests on a large scale in a narrow summative assessment window is not available yet in many states. Therefore, we may need to consider the possibility that while complete technology delivery is a goal for the Generation 2 assessment system, transition to these technologies may need to be staged over the period of Generation 1 implementation.

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 assessments with data collected over the course of the year. The use of interim elements as part of a summative system could also provide ways to experiment with the use of new item types and technologies.

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.
2) For each assessment proposed in response to question 1), describe the—
   ● Optimal design, including—
     — Type (e.g., norm-referenced, criterion-referenced, adaptive, other);
     — Frequency, length, and timing of assessment administrations (including a
       consideration of the value of student, teacher, and administrative time);
     — Format, item-type specifications (including the pros and cons of using different
       types of items for different purposes), and mode of administration;
     — Whether and how the above answers might differ for different grade levels and
       content areas;
   ● Administration, scoring, and interpretation of any open-ended item types, including
     methods for ensuring consistency in teacher scoring;
   ● Approach to releasing assessment items during each assessment cycle in order to ensure
     public access to the assessment questions; and
   ● Technology and other resources needed to develop, administer, and score the assessments,
     and/or report results.

   ● Optimal Design

We believe the summative assessments should have two major components: a common-core
assessment and an optional state-specific assessment. Our understanding is that states may augment
the common-core standards with their own standards, as long as the common-core standards represent
at least 85 percent of the universe of standards in the state at any grade where common-core standards
exist. 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 could 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, LEAs, and schools. The common components of the test will be
designed to yield state, LEA, 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 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. Note that we do not mean that the same exact test form is required, but rather the same assessment, which would be available in equivalent (or adaptive) forms consistent with test security.

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 toward 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. (We comment on this further in the response to bullet 4 under this section [Technology and other resources needed to develop, administer, and score the assessments, and/or report results].) 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, provisions will need to be made for pilot/field testing new content. For the common components, this could be accomplished either through a variable section or by embedding pilot/field test items within operational sections.

One open question is how big a system (in terms of assessment exercises) would be needed to maximize 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. This answer also will be affected by the speed 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 to 120 minutes (for both common and state-specific components). High school tests could, conceivably, take between 2 and 3 hours. If extended tasks are used, assessment time may need to exceed these limits.

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 the 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.

Our proposed design 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 help 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 students should 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 summative assessment. In addition to this, the assessment will need to have 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, states may find it challenging to explain major differences between their standards and national standards. But the system needs to support these types of data.

Use of Adaptive Testing

As mentioned above, we believe that the summative-assessment system should make use of adaptive administration in Generation 2. Whether or not all elements of the system can use adaptive administration in Generation 1 is yet to be determined. The answer will depend on the type of adaptive models we wish to use, and the availability of technology for universal computer administration.

A variety of adaptive testing approaches may be used when the assessment system reaches maturity (e.g., traditional item-level 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 reading/language arts with a heavy emphasis on authentic reading, we believe variations in traditional computer-
adaptive testing 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 return on the investment in 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 the system could be adaptive in the first year of administration, even for the interim-assessment system. Large-scale piloting of items would be necessary before rollout. However, given issues associated with calibrating a pool under suboptimal motivational conditions, it is likely that the rollout year of the program would require assembling a large number of linear tests and assigning these randomly to candidates. The system could, however, use adaptive administration in subsequent years.

---

Frequency, length, and timing of assessment administrations (including a consideration of the value of student, teacher, and administrative time)

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 end-of-year or end-of-course 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 assessment 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 assessment did cover the entire year’s worth of content, but 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, as well as real challenges. On the positive 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, 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 LEAs got the same average scores on the end-of-year assessment, that phenomenon would normally be interpreted to mean that those two LEAs ended that school year “in the same place.” Rating one LEA higher because of performance on intermediate ratings might be problematic.

We describe an alternate model, used in some other countries, below. There would still be an end-of-year assessment, 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. By making these sorts of tasks part of a formal accountability system, this model 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.

Our recommended transition to a new assessment system in two generations can allow for experimentation in these approaches without disrupting the utility of the accountability testing system. During Generation 1, the end-of-year or end-of-course assessment can be an “event” test, with the pressures of fast turnaround of results and the benefits of low cost, emphasizing or exclusively containing machine-scorable items. This event test can be supplemented with results from carefully controlled, but not necessarily identical, interim assessments that take place throughout the year, consisting of items in various formats; these assessments can be computer-adaptive. As we learn more and get technologies and operations in place to make the assessment system work more fluidly, we can advance the innovative item types and administration methodologies into the end-of-year assessment in Generation 2.

There are a number of issues that would need to be addressed in making such an interim-assessment system operational. It would need mechanisms to help ensure 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 tradeoffs that must be considered 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 concept deserving of thoughtful consideration. 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 the time that they can be added to the accountability system. We believe that strong, forward-looking end-of-year assessments will be part of the system.

---

Format, item-type specifications (including the pros and cons of using different types of items for different purposes), and mode of administration

Stating a firm position about item types 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 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).

During the design effort, other questions will emerge about the sorts of items and tasks that can be used. These will surround issues like the use of audiovisual stimuli (as called for in the CCSSO/NGA English-language arts 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 reading/language arts 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). Decisions will need to be made about 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 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 the key.

Our proposal to phase in the new system in two generations interacts with the question of item and task
type. During Generation 1, the tasks and items that take longer to administer and score, or require
computer administration, could be limited to the interim system, and the end-of-year assessment can
rely only on the types of items and tasks that can be scored by computer, thus hastening the availability
of results for accountability purposes. We have concerns that such limitations on the summative
assessment would narrow curriculum and teaching, but these concerns are tempered by the fact that
results from the interim assessment would also be part of the accountability system. In addition, we
would be working toward a Generation 2 system in which these performance items/tasks would also be
contained in the summative assessment event itself.

— Whether and how the above answers might differ for different grade levels and
content areas

We assume that the summative-assessment system will include end-of-year reading/language arts and
mathematics assessments 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. These end-of year
assessments may not be the only components of the summative system. At high school, we believe two
summative models are possible: either end-of-domain assessments in both reading/language arts and
mathematics that cover the knowledge and skills needed to be ready for college and career training, or a
series of end-of-course assessments. Each approach has advantages and disadvantages, depending on
the priorities selected.

One should not assume that a single assessment model or design will make sense at all grades and in all
subjects. For example, tests used at early grades will almost certainly be shorter than those used at the
high school level. It is also likely the case that the types of exercises used may vary across grades and
subjects, as may the mix of machine and human-scorable items. The amount of technology familiarity
we can expect of test takers may also not be consistent across grades. It is even conceivable that the
constructs covered at grades may vary: For example, a grade 4 reading/language arts test may focus on
reading and writing skills, while a college readiness measure may also include measures of student
abilities to listen to lectures. Certain underlying goals, like growth modeling, may be easier to achieve at
elementary than at secondary levels. Finally, even if the “end state” is to have assessments that are
similar across ages and subjects, the transition plan may not be the same. We may be “more ready” to
test grade 8 reading/language arts in a computer-based setting than we are to assess grade 4
mathematics.

We cannot determine the specific ways in which answers vary by age and grade until specific grade- and
subject-specific standards are finalized.

- Administration, scoring, and interpretation of any open-ended item types, including
  methods for ensuring consistency in teacher scoring

To optimize the speed and cost-effectiveness of scoring a range of non-traditional 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 exercises such as simulations that go beyond simple student responses. Third, while
some tasks can be machine-scorable, 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. The 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-scorable 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.
Ensuring Consistency, Reliability, and Accuracy in Scoring

Given the high-stakes nature of these proposed assessments, helping ensure reliable and accurate human scoring is critical. We propose a multilevel, multifaceted approach, because it is most effective at establishing that only the raters who learn to use the scoring rubric accurately are allowed to begin scoring, and it verifies that raters stay on track throughout scoring. The following is a list of some of the procedures that are used to monitor and train raters to help ensure consistency in scoring – regardless of whom makes up the rater pool (i.e., teachers or professional raters).

- Rater calibration — occurs prior to operational scoring and tests the rater’s abilities to appropriately apply the scoring rubric to specific items
- Response randomization — occurs during operational scoring and randomizes responses distributed to raters
- Double reads — occurs during operational scoring and requires two independent rater scores for items, if required by the design of the program
- Response distribution rules — occurs during operational scoring and helps ensure that no systematic biases are introduced into scoring
- Scoring leader backsoring and validity papers — occurs during operational scoring and helps ensure that raters are scoring according to the scoring rubric and rules
- Trend scoring and equating — occurs during operational scoring when items are reused between administrations to determine a statistical comparison

The combination of these procedures creates a strong framework of checks and balances that protects scoring fidelity, while attempting to minimize the amount of additional scoring time and cost needed to establish a strong and defensible process.

During Generation 1 of the assessment system, types of items that either require human scoring, or use of as-yet-unproven technologies for computer scoring, should be limited to the formative system and/or the interim-assessment system, with results aggregated to the accountability system. As we gain confidence in their use for high-stakes purposes, we can graduate them to the end-of-year summative assessments for Generation 2.

As mentioned in the Introduction to this response, there are competing priorities that have a lot to do with the type of scoring used, especially on the end-of-year assessment. How important is rapid turnaround of results on the year-end summative assessment? Is it important enough to limit the end-of-year assessment to machine-scorable items, or can we take the time to do human scoring of some items, adding a few weeks to the processing required before getting score reports?
General Assessment Questions

- Approach to releasing assessment items during each assessment cycle in order to ensure public access to the assessment questions

Although costly, the release of test questions is very beneficial for a number of reasons. Released items for the common assessment will enable students and educators to understand how the content standards will be measured and will give students practice with the various item formats used on the tests. This kind of information is especially valuable for open-ended questions, with their accompanying rubrics and sample responses. To the extent to which we succeed in making the exercises models of effective instruction, released exercises can provide a useful toolkit to teachers. In addition, released exercises tend to demystify assessments.

We have provided a more complete discussion of releasing test questions in response to question 3 under “Specific Technical Assessment Questions.”

- Technology and other resources needed to develop, administer, and score the assessments, and/or report results

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, as soon as it is practical, the assessment of common standards should be computer-based (or other-technology-enabled) tests in which paper is used solely for certain special accommodations. We describe this ideal version of the Race to the Top assessment program as Generation 2 throughout this document, and believe that the transitional system (Generation 1) should consist of a steady march toward the eventual goal of having almost all of the system be computer-administered.

There are several reasons for recommending that the entire system be computer-delivered in Generation 2:

- Emerging standards in both mathematics and reading/language arts 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 are 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 brings 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 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.

If the summative assessment is delivered via a technology platform based on accepted interoperability standards, it could feed data to, and receive data from, the interim and formative segments of the system, thus creating an integrated, balanced system.

Technology will continue to improve, become easier to use, and become 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 administration 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 tradeoffs between assessment purpose (like high-stakes graduation decisions) and the size of the testing window allowed. Finally, because 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.

During Generation 1, we recommend that computer administration be used as much as possible for the interim assessments and formative assessments, at least. This would allow for the use of newer item types and scenarios that measure 21st century skills and Information and Communication Technology (ICT) literacy throughout the year. It would also allow for the build-up of capability and capacity in the scores for the eventual transition of the entire end-of-year assessment to technology delivery in Generation 2. During Generation 1, the summative assessment can be administered via technology in those states and LEAs that are ready for it, but might have to be administered, at least for two years or so, by paper in other states and LEAs. While this puts severe limitations on the types of items we can include in the summative assessment during Generation 1, and slows down turnaround time for some LEAs and states, it would allow for quicker implementation of the overall assessment and avoid the problems that would occur in forcing the system into total technology administration before the infrastructure and operational base is ready.
ARRA requires that States award at least 50 percent of their Race to the Top funds to LEAs. The section of the notice entitled Design of Assessment Systems – LEA-Level Activities, describes how LEAs might be required to use these funds. What activities at the LEA level would best advance the transition to and implementation of the consortium’s common, college and career ready standards and assessments?

The Race to the Top funds provides LEAs with an extraordinary opportunity to participate in the development and implementation of next generation assessments. In particular, resources directly available to LEAs provide the chance to build capacity in ways that improve teaching and learning. Additionally, the investment in and among LEAs will help prepare all students for college and careers.

The following LEA-level activities would, in our judgment, best advance state consortia common standards and assessments.

**Development of Formative Components of Assessment Systems**

As part of a balanced next generation assessment system, LEAs can help develop formative components designed to work with the summative components that are centrally developed. These components can include rich performance tasks that are closely aligned to classroom practice, reflect learning progressions, serve to reinforce learning, and identify gaps in knowledge and skills.

While it is also possible for LEAs to be involved in the development of the interim components of the system, this is a bit more challenging if those interim components will be used for accountability purposes via combination with the summative assessments. The interim assessments would have to have some degree of standardization and security for that plan to work.

**Professional Development**

One of the most enduring and useful expenditures of Race to the Top funds at the LEA level would be to promote professional development activities related to a next generation assessment system. Without teacher and school staff involvement in and understanding of what the changes are and why they are being implemented, the new assessment elements introduced by this initiative will at best be minimized; at worst they will be frustrating to staff and eventually ignored.

The following are some of the ways that school and LEA staff can learn about and better appreciate the changes to their assessment program:

*Assessment Literacy:* A next generation assessment system will bring new terminology and concepts that could confuse and intimidate those who are put in the position of explaining the new assessments to parents and the general public — terms such as common standards, 21st century skills, international benchmarking, and adaptive testing. Assessment literacy requires an understanding of types and purposes of assessments, how to glean information from summative assessments, and the use of student achievement data in teacher performance evaluations. Teachers need to be thoroughly aware of the changes that will result from the new assessment
system because they are the front line, talking daily with parents, neighbors, and others outside of education. Their understanding and support are critical to the implementation and success of a new assessment system.

**Writing/Reviewing Test Questions and Scoring Open-ended Questions:** Training teachers and staff to both write/review new test questions and score open-ended questions has been shown to be one of the most beneficial professional development experiences for those who provide and manage instruction. Likewise, involving teachers in item and test development activities provides important opportunities for training in the writing of clear and accurate items that are aligned to the standards, as well as effective assessment design. As previously noted, teachers can also help to develop formative assessments — the perfect arena for local staff to make a significant contribution to the state’s assessment system. Because part of the cost of using teachers to score assessments is hiring substitute teachers to replace them in the classroom, this is an appropriate use of the LEA flow-through funds under the Race to the Top assessment program.

**Capacity Building**

As states adopt common standards, considerable time and effort will be needed to align local curriculum, assessment, and instruction to the new standards. LEAs need resources, direction, and support to develop benchmark and formative-assessment materials and encourage the use of multiple measures that work together to achieve common goals.

Funding at the local level could help manage logistics, communication, and technical support to create sustainable programs that align curriculum and instruction at each grade level with the common and state-specific standards.

Capacity building involves purchasing instructional programs and resources, and doing the professional development that enables teaching and learning of the new standards to be most effective.

In addition, we should encourage LEAs and schools to collaborate with others and share resources and best practices, which includes a focus on capacity building at all levels. County and regional cooperatives can pool resources to purchase services that may otherwise not be available to individual LEAs.

4. If a goal is that teachers are involved in the scoring of constructed responses and performance tasks in order to measure effectively students’ mastery of higher-order content and skills and to build teacher expertise and understanding of performance expectations, how can such assessments be administered and scored in the most time-efficient and cost-effective ways?

Human scoring adds challenges for any assessment system. Human scoring means that it takes longer to release scores. It also raises assessment costs and psychometric challenges regarding data comparability, particularly over time. Of course, human-scorable items also allow for the measurement of skills beyond what is possible in a machine-scorable system. And finally, scoring itself can be a powerful professional development activity. This tension leads to a key question: How do we mitigate
the negatives and enhance the positives? How do we improve quality and control cost and time? The answer involves, at least partly, an effective use of online scoring technology.

Student written responses were traditionally scored in a face-to-face (F2F) setting, with raters assembled in a central location. This approach had some real advantages. There were few technology needs, as the actual student books were used in scoring. F2F scoring allows for personal interaction between raters and scoring leadership. However, there were two main limitations of this approach. First, real-time quality control and assurance tools in paper-based systems were limited. Second, F2F scoring required expenditure on travel, lodging, subsistence, facilities, and equipment.

Online scoring has helped address both of these weaknesses. The systems themselves provide real-time quality control and assurance tools. The systems also obviate the need for F2F training: One can create either online or virtual F2F environments and dramatically reduce costs for travel, lodging, subsistence, facilities, and equipment through the use of online distributed scoring. Distributed scoring allows participation in the scoring process by teachers with computers and internet connections from anywhere in the world.

**Approaches to Involving Teachers in Constructed Response Scoring**

We propose alternative ways to meet the Department’s goal of involving teachers in the assessment process — ways that would promote buy-in and opportunities for professional development. Employing teachers to complete the constructed response scoring of field test or operational items is one approach to achieving the goal of involving teachers in the assessment development process. The degree to which teachers are involved in scoring, however, must be balanced with regard for the apparent conflict of interest. In an accountability context where the outcomes are consequential for teachers as well as for students, a rater pool of teachers might compromise public acceptance of the validity of the outcomes. This is not to argue that teachers should not be involved, but rather that the system must help ensure that this involvement does not threaten the validity of scores.

In addition to carrying out all operational scoring, there are other types of teacher involvement that can increase confidence in test scores from a stakeholder’s perspective, and reliability and validity from a psychometric perspective. One approach would be to invite a diverse panel of teachers to evaluate the alignment between scoring criteria and the broader performance expectations described in the set of K-12 standards. During such evaluative exercises, teachers would map the scoring criteria for each content area to the performance expectation they judge to be the best fit. The degree to which these judgments align with the standards and with one another would provide information about how well the performance expectations are used by the scoring criteria. Another approach might entail teachers’ input in pilot item and scoring guide review. When constructed response items are field tested, teachers would review the items and score responses using the accompanying guidelines. The merit of the items and scoring guidelines would be judged based on the quality of responses received and how well the guidelines can be applied to those responses. Teachers participating in this activity would interact with colleagues from other LEAs and states and gain valuable professional development experience in crafting performance assessments.
Another alternative to actual scoring is to get teachers involved in a range of quality control processes such as exemplar (sample) response selection. Teachers would collaborate with assessment developers to choose the responses used for training, certification, calibration, and validity monitoring of raters. This highly deliberative session gives participants the opportunity to discuss the item, the expected performance standard, and the scoring rubric in detail. This process results in samples of student responses established as clear examples that represent each attribute of the scoring rubric. One set of exemplars is used to train raters for operational scoring; these exemplars serve as reference points during the scoring process. The others are used in rater scoring quality monitoring.

5) Given the assessment design you proposed in response to question 1), what is your recommended approach to competency-based student testing versus grade-level-based student testing? Why? How would your design ensure high expectations for all students?

The distinction between grade-based testing and competency-based testing is neither clear nor absolute. For example, standards for grade 4 students define competencies we expect students at that level of schooling to have obtained. It is possible to build assessments of those competencies. The decision about whether to administer those assessments to all grade 4 students or to allow grade 3 students to take tests if their teachers view them as ready to do so does not necessarily change the nature of the assessments. That having been said, there are ways to address the question above that help illuminate choices test developers will need to make in designing the new system.

One other point of definition: The phrase “competency-based testing” has been used in two ways. The first is to describe an assessment system in which students take tests when they are ready to show mastery of the competencies covered in that test. So in the example above, a grade 3 student might take the grade 4 competency test if ready. For purposes of discussion, we will call that “testing-when-ready.” The other meaning of “competency-based testing” tends to refer to separate testing of distinct competencies or clusters of competencies (which in the current discussion resemble standards). We will call that “assessment of specific competencies.” Given the nature of the question above, we assume the former usage is intended, although we will say a few words about both.

The summative assessments of common-core standards can combine elements of competency-based and grade-level-based testing. As mentioned above, at least in the early years of the new assessment, we will need to administer end-of-grade or end-of-course assessments to allow for the collection of system-level accountability data. We also recommend that the summative system make use of adaptive administration. In this case, off-grade content may be selected for either high- or low-performing students. However, all students would be tested with on-grade content, and the only use of off-grade content would occur as a result of adaptation. Finally, this helps ensure high expectations for all students in that all children will be evaluated against the within-grade rigorous standards, even if off-grade content is administered.

While the system will begin as a grade-based assessment regime, it could easily evolve into one in which people test when they are ready. For example, the system could include an end-of-grade 5 mathematics test, which will measure student mastery of the appropriate standards. Once we achieve the eventual
goal of a completely computer-based adaptive system, there is no reason that this could not evolve into an on-demand system in which students test when they are ready. So in this example, if a student in grade 4 felt ready to show mastery of grade 5 content and skills, he or she could. Of course, this system loses some of the data advantages of having fixed snapshots in time (for example, this may make it harder to make school comparisons, or implement value-added models). In addition, this sort of flexibility in testing has implications for instructional management in schools. Policymakers will need to determine whether these sorts of challenges are worth it given the advantages of an on-demand system.

We also believe that this system can accommodate the other meaning of competency-based testing, that individual testing or test-based events focus on individual competencies or clusters of competencies. This is not to recommend such an approach, but simply to state that it is possible. We argue that the summative system may not be a single testing event, and may rather be a combination of testing events that occur over the course of the year. In a system where there are multiple tests given at fixed intervals, these tests can in some cases focus on specific competencies. This implies that we can agree on the clusters of competencies that should be combined in the intermediate tests. Alternately, it implies a library of competency tests that states and LEAs can select and use with some discretion, although such a system would certainly reduce the comparability of results.

The use of a testing approach in which the summative system uses data from multiple tests does not necessarily assume competency-based testing of this sort, of course. There are several reasons not to consider separate assessments of specific competencies. Such testing has been criticized for encouraging inappropriate disaggregation of skills that should be viewed and assessed as integrated. Additionally, such approaches are easier to implement in mathematics than in reading/language arts.

In summary, the high-level system design has elements of grade-based testing, but could evolve into one that includes competency-based testing, whichever way one defines that term.

6) Given the assessment design you proposed in response to question 1), how would you recommend that the assessments be designed, timed, and scored to provide the most useful information on teacher and principal effectiveness?

Student and school effectiveness cannot be gauged based simply on percentages of students who reach standards. Different schools face different levels of challenge, and different teachers add varied levels of value. A system that has the measurement of teacher and school effectiveness as a goal requires data on the amount and nature of student improvement over time. In other words, if we are to use student performance information as a source of data on teacher and school effectiveness, we must have data on student growth.

Given the overall interest in student growth metrics (and the use of such metrics in teacher evaluation), we believe the assessment system 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.

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 assessments 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 assessments 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, but such considerations are beyond the scope of this response.

A well-structured student assessment system can be one source of data to be used in evaluating teacher effectiveness. One thing policymakers will need to consider is how to use these data in conjunction with other relevant pieces of information.
Specific Technical Assessment Questions

1) What is the best technical approach for ensuring the vertical alignment of the entire assessment system across grades (e.g., grades 3 through 8 and high school)?

From a technical standpoint, a vertically aligned assessment system is best developed within the context of coherent, vertically articulated content standards and performance expectations. An effective system includes within- and cross-grade alignment of standards-based instruction that is informed by assessment results and supported by ongoing professional development. Vertical articulation across grades is evident in performance level descriptors and the cut scores established to differentiate proficiency levels at each grade.

The assessment system would support within-grade proficiency measures as well as cross-grade comparisons of individual student performance. The cross-grade comparisons are desired to measure growth and determine at each grade level tested whether a student is on track toward college or career readiness by the time of high school completion. Recommendations for the best technical approach to helping ensure vertical alignment of the assessment system, and for scaling and reporting to support cross-grade inferences, will depend on a number of factors, as they pertain to the adopted standards and performance levels.

The provision of coherent, vertically articulated content standards and performance expectations could permit the construction of cross-grade or vertical scales. A vertical scale entails the notion of learning or growth across time. Because the common assessment tests students across a grade range with articulated content, longitudinal types of inferences based on scale scores would be possible if a vertical scale were implemented.

It should be noted that developing a vertical scale is technically complex, and may not be feasible in all situations. In these situations, use of within-grade or horizontal scales does not preclude measuring student growth. Although horizontal scales support direct comparisons of scale scores between same grade cohorts only, additional statistical procedures may be used to track individual growth over time.

We describe considerations for vertical scaling and other options to estimate student growth across grades in the context of vertically aligned content standards and performance expectations below.

Using Cross-grade Scales (Vertical Scaling)

As stated above, one approach to facilitate measurement of student growth is to develop a vertical scale. In a vertical scale, for each content area (e.g., reading/language arts or mathematics) scale scores run continuously from the lowest grade tested to the highest grade tested, with substantial overlap of the scale scores produced by adjacent grades. On the vertical scale, “Proficient” might be a scale score of 350 in grade 3, 380 in grade 4, 400 in grade 5, and so forth. The difference in a student’s scale scores at adjacent grades is a measure of the amount of academic growth achieved by that student.
Specific Technical Assessment Questions

With a vertical scale, an ideal goal is to have scale scores that have the same meaning if they are obtained from different test levels (e.g., a 400 “means the same thing” or represents equivalent knowledge or achievement whether it comes from the grade 4 test or the grade 5 test). Also, differences between scores are ideally comparable for gauging amounts of academic growth. For example, a student who grows from a scale score of 350 to 370 (20 points) would be demonstrating the same amount of growth as a student who grows from 390 to 410 (20 points).

In addition, for ease of interpretation and use in measuring growth, vertically scaled item pools allow the use of adaptive administration engines to select out-of-grade content to provide additional diagnostic information for either advanced or struggling students. Note that under current interpretations of NCLB, the content of assessments used for accountability must be aligned to grade-level content standards; however, the diagnostic out-of-grade content could be administered as an augmentation to the grade-level test. In the context of interpreting individual academic growth for students, the construction of a vertical scale is critical for these types of inferences.

Listed below are some technical considerations in producing the vertical scale:

» To produce a vertical scale, it is assumed that the tests at adjacent grades have substantial overlap or articulation in content, and that a single, major dimension (e.g., overall mathematics achievement) explains most performance differences. An additional consideration is that vertical scaling makes the implicit assumption that the same construct is being measured at the top and bottom of the score scale, and this assumption may be difficult to justify when the vertical scaling includes many grades. Caution is needed in comparing growth in different parts of a vertical scale, whether comparing growth of low-achieving and high-achieving students or students in substantially different grades.

» Beyond the expert judgment involved in establishing vertical articulation of content, a vertical scaling requires special data collection and analysis. (In the data collection, students take test items that measure content from adjacent grades in addition to content from their own grade; their relative performance on the two sets of items determines the relative difficulty, or scaling, of the different sets of items.) This is of particular concern for the high school assessments under the current paradigm. For example, administration of comprehensive assessments in grades 9, 10, or 11 could perhaps support cross-grade scaling if based on content that builds sequentially from lower grades, such as might be the case in certain versions of integrated mathematics curricula. On the other hand, administration of end-of-course assessments under traditional curricula (such as algebra I, geometry, algebra II or biology, chemistry, and physics) might not articulate. Of course, the fact that an approach might not work in high school does not render it ineffective at earlier grades. The new system need not rely on a single approach.

» A similar issue for consideration involves the state-specific content, and the degree to which it is a) aligned to the common-core standards and b) vertically aligned across grades within a state. It is likely that state-specific content may differ substantially across states; thus, it may be that a vertical scale is developed for the common-core component only, for the purpose of cross-state comparisons. In other words, the vertical scale based on the common core will...
measure growth only on the common core. If a state wanted to include their state-specific augmentation in the vertical growth measurement, then a state-specific vertical scale would be needed.

**Using Within-grade Scales (Horizontal Scaling)**

As stated previously, vertical alignment of the assessment system is best realized in the context of vertically articulated content standards and performance expectations. Although appealing for reasons listed above, vertical scales are not the only way to measure student growth. When assessment systems have separate within-grade scales, whether by design or because vertical scales are not feasible, alternatives that do not require a vertical scale may be used to estimate student progress across grades. For example, regression-based techniques to estimate growth percentiles or growth trajectories may also be used to determine whether students are on track to achieve the desired level of performance at various designated time intervals.

In contrast to vertical scaling, use of separate within-grade scales is not predicated on the assumption that a single major dimension is measured across grades; thus, it allows more flexibility in content differentiation. This is an advantage when considering how to assess students at the high school level where the content domain is more varied and course-specific.

Selection of statistical methods for estimating student growth should take into account the transparency of the statistical method used and its interpretability for the public and educational decision makers.

**Summary**

Recommendations include coherent, vertical articulation of content and performance standards across grades, with instruction aligned to standards and assessment; design decisions that support test properties adequate for the intended use; and the use of vertical scales if feasible. It is expected that the assessment system would support within-grade (horizontal) proficiency measures as well as cross-grade comparisons of individual student performance. Student growth measures may be obtained with or without a vertical scale, and there are some limitations associated with both options.

While beyond the scope of this response, one cautionary note is in order. Regardless of the ultimate assessment design and scaling procedures, given the high stakes associated with test results, we recommend careful monitoring of continued alignment. As with any educational reform involving new content standards, there may be changes resulting from clarification of standards and/or implemented curriculum, with implications for parameter estimation, and the stability of scales and the validity of performance standards established in the early years. To this end, we recommend periodic studies to evaluate whether revisiting performance standards and/or resetting baseline scales may be warranted.
2) What would be the best technical approach for ensuring external validity of such an assessment system, particularly as it relates to postsecondary readiness and high-quality internationally benchmarked content standards?

**International Benchmarking**

The phrase “international benchmarking” has several possible meanings. It can be defined as a content activity aimed at determining what leading students around the world are being taught. International benchmarking can be a level-setting activity, in which we set performance expectations (and related cut scores on tests) at points where they suggest readiness to compete in a global economy. Finally, international benchmarking sometimes refers to studies that compare actual educational attainment in different countries. TIMSS, PIRLS, and Program for International Student Assessment (PISA) are examples of such studies. These definitions of “benchmarking” are not contradictory: Individual studies can undertake all three sets of activities.

What is the value of international benchmarks? They allow countries as well as local educators and educational researchers to better understand the relative strengths and weaknesses of their education systems and possibly to identify best practices and plan appropriately.

If we are to make international benchmark data integral to the new assessment of common standards, we must have an ongoing plan to maintain the validity of those standards. This plan must cover the three benchmarking definitions described above. The Department should consider periodic support for curriculum reviews in high-performing countries, to make sure that “common-core standards” are keeping pace with international advances. In addition, as expectations around the world change, it may become necessary to update cut scores on the common-standards assessments. Finally, the Department should promote some linkage between the common-core assessments and international surveys (possibly through studies linking NAEP and these surveys, since all states already participate in NAEP). This linkage would likely be conducted through special studies in which some students take both the international survey and NAEP, or in which randomly equivalent samples do the same.

**Postsecondary Readiness**

Another issue relates to the evaluation of postsecondary readiness and external validity of the common assessment. Both predictive and convergent/discriminant validity can be used to support the assessment system in the context of postsecondary success. Predictive validity is concerned with the use of test scores to predict a variable of interest. Not surprisingly, students with higher levels of measured academic skills are more likely to have higher grades and graduate from college than their less able peers. Likewise, if students with low common assessment scores have to take college remedial courses, then this is an indication that these students were not adequately prepared for college-level material. A number of studies have investigated the predictive validity of the SAT or ACT® on grade point average (GPA) in the first year of college (e.g., Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008) and for other types of college outcomes. In a similar fashion, the common assessment could be used to predict first year college GPA or other outcome variables such as college graduation status. Because SAT and ACT test scores are the second-most important factor in college admissions decisions after high school
grades (Hawkins & Lautz, 2005), another option is to use the common assessment to predict college entrance performance on the SAT or ACT examinations.

There are numerous indicators of postsecondary success, including postsecondary institution enrollment, persistence, remediation, and degree completion; employment; military enlistment; and earnings that also can be used to demonstrate convergent and discriminant validity. These types of studies would require a concerted data collection effort but not one that is unprecedented. As of 2008, 20 states were providing postsecondary feedback reports on high schools, including information about each high school graduate’s participation in postsecondary education (Hanover Research Council, 2008). Unfortunately, state strategies for implementing the postsecondary feedback report — including how they collect data, produce the indicators, and report information — vary significantly. For example, while most states include enrollment and remediation data in their feedback reports, only a handful report persistence rates and first year college GPA, and even fewer provide data on degree completion or workforce participation. For the common-core assessments, there would need to be more coordination in the definition and organization of these types of information in order to support these types of validity inferences.

3) What is the proportion of assessment questions that you recommend releasing each testing cycle in order to ensure public access to the assessment while minimizing linking risk? What are the implications of this proportion for the costs of developing new assessment questions and for the costs and design of linking studies across time?

We believe that the release of exercises should take place in three stages: at the prototype stage, after initial pilot testing, and on an ongoing basis after operational testing. Because released items are an important part of a testing program, a number of prototype items should be disclosed as soon as possible after the assessments have been designed, even before any items have been pilot/field tested. The prototype items should be widely available on the Internet so that students and teachers can begin to understand the overall assessment design as well as the item formats.

After items have been pilot/field tested, a predetermined number of pilot/field tested items should also be released, possibly with one or more statistical characteristics (e.g., item difficulty) accompanying each item. Both of these releases should include sufficient numbers of items to demonstrate how key standards will be tested and to illustrate the complete set of item types that will be used.

Traditionally, the items released after pilot/field testing comprise either a representative set, perhaps the equivalent of 15 percent to 25 percent of the items an individual student may take, or the number of items that appear in a full-length testing session. We recommend the latter approach, given the importance of this assessment.

After the program is up and running operationally and pools have grown to a “steady-state size,” we believe that each year new content addition should be more or less balanced with the release of exercises from the operational pool. So for example, if we add 100 new grade 4 mathematics items to an adaptive pool, we would release roughly 100 for use by educators. We would further propose that the
released exercises be accessible in at least two ways. First, teachers who wish to use these exercises for classroom purposes should be able to retrieve them from the Web. Second, students and teachers should be able to access the items in a way that emulates an actual administration (in other words, that is adaptive and that produces a scaled score, if teachers are willing to score the open-ended items). Such uses would be low-stakes, of course.

This sort of an aggressive release plan has the advantage of giving teachers and students more complete information about the knowledge and skills to be assessed as well as the level and type of performance that are expected. In addition, the release of forms rather than single items avoids an appearance of secrecy regarding the tests. However, it can be argued that releasing one or more complete or multistage forms increases the likelihood of teachers focusing on test items rather than student learning. Also, the cost of releasing intact forms, even smaller multistage forms, is much greater than the cost of releasing representative sets.

The plan for releasing items must be fully crafted prior to beginning item development. It is important to develop a sufficient number of items, with the customary overages, to allow for the planned releases. A careful plan will permit the release of items without any risk to linking items.
High School Assessment Questions

Provide recommendations on the optimal approach to measuring each student’s college and career readiness by the time of high school completion. In particular, consider:

1) How would you demonstrate that high school students are on track to college and career readiness, and at what points throughout high school would you recommend measuring this? Discuss your recommendations on the use of end-of-course assessments versus comprehensive assessments of college and career readiness. (Note: If you recommend end-of-course assessments, please share your input on how to reconcile the fact that college and career ready standards might not include all of the topics typically covered in today’s high school courses.)

It would be ideal to measure readiness in grades 10 and 11, so that students are encouraged to use their last two years of high school, and especially grade 12, to increase their mastery of the readiness standards.

To demonstrate that high school students are on track to college and career readiness, there are two available models:

» States would administer comprehensive reading/language arts and mathematics assessments in grades 10 and 11, and these assessments would be designed to directly measure the college and career readiness standards. Cut scores would be established to indicate degrees of readiness, with sufficient time for students to increase their readiness.

» States would administer end-of-course assessments (in such courses as English II and III, geometry, algebra II, biology, and chemistry) and supplement those assessments with sets of questions that specifically measure readiness standards. For the purposes of giving readiness scores, items from the end-of-course assessments would be combined with the readiness items to form a reliable full-length assessment. This model has been in use in California’s Early Assessment Program (EAP) since 2005. The state administers a set of readiness items along with their grade 11 end-of-course assessments: English grade 11, algebra II, and high school summative mathematics. Under this model, a similar supplement could be developed for grade 10 and for science, which currently are not part of the California EAP.

Because the material typically covered in today’s high school courses is usually broader than the requirements of the college and career readiness standards, California selects half to two-thirds of the items on the end-of-course tests to use for the EAP measure. The selected items meet specific statistical and content specifications.
Some educational leaders in the U.S. have recommended that we employ a system of extensive examinations in high school, similar to that used in other countries, like the United Kingdom. There are many advantages to this proposal, especially in that it drives the high school curriculum and high school teaching to prepare students for extensive, multifaceted assessments that are worth teaching to. The type of system we have proposed here is not incompatible with such an approach, especially if the system we propose is used for grades 3 through 8 and another system is used for high school. We assume that the proponents of the examination approach are submitting their comments and advice to this request for information separately, so we will not deal with this subject in detail here.
Questions on the Assessment of English Language Learners

1) Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners. How would you recommend that the assessments take into account the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills in core academic areas? Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

Importance of Assessing English Language Learners

English language learners comprise a large and rapidly growing portion of the K-12 population in the U.S. According to recent statistics from the National Clearinghouse for English Language Acquisition (2007):

- Currently 1 in 9 K-12 students in the U.S. is an English language learner; by 2025, as many as 1 in 4 students may be.
- English language learners are prominent in our largest states. For example, in California, more than 25 percent of the student population are English language learners; in Texas, more than 15 percent.
- Many states that have not traditionally had large numbers of English language learners have experienced rapid large-scale influxes in recent years.

The import of these data for the next generation of common summative assessments is clear: In order for it to successfully serve the purposes for which it is designed, the new assessment system must provide fair and valid information about the skills and abilities of English language learners.

Factors Influencing the Assessment of English Language Learners

The challenges of designing and implementing an assessment system that is appropriate to both the general population and to English language learners involve factors related to language, to educational background, and to culture. These factors are briefly outlined below. For a more detailed discussion, see Pitoniak et al. (2009).

Language

- **Different language backgrounds**: While approximately 80 percent of English language learners come from Spanish-speaking backgrounds, it has been estimated that more than 400 languages are spoken by English language learners nationally.
- **Varying levels of proficiency in English**: English language learners can range from true beginners with minimal English skills to students with levels of fluency approaching that of native speakers. A particular challenge for assessment is posed by the fact that there is no predictable relationship between age (or grade) and level of English proficiency; one English language learner in grade 10 might be a true beginner just learning the English alphabet,
while another might be highly fluent in English. Students at these two extremes — and at all levels in between — will interact very differently with content assessments.

» **Differing profiles of English language proficiency**: English language learners may have varying degrees of relative proficiency in oral versus written English and in interpersonal versus academic English. Students who are able to converse fluently in English may not have the literacy skills required to negotiate a standardized test.

» **Varying levels of literacy in native language**: As English language learners vary in the degree to which they can read and write in their native language, it is important not to assume that they will be able to understand written test directions or other test content in their native language.

**Educational Background**

» **Varying degrees of formal schooling in native language**: In addition to native language literacy levels, the degree of formal schooling in the native language also affects English language learners’ content-area skills and knowledge.

» **Varying degrees of formal schooling in English**: English language learners vary both in the number of years they have spent in English-medium schools and also in the type of instruction received there (e.g., bilingual, full English immersion, English as a second language).

» **Varying degrees of experience with standardized testing**: Any testing format — including multiple-choice items, constructed response items, and computer-based administration models — will likely impact different cohorts of the English language learner population differently depending on their degree of previous exposure.

**Culture**

» **Varying degrees of acculturation to the U.S. mainstream**: Students who are unfamiliar with American culture may be at a disadvantage relative to their peers because they may hold different assumptions about the testing situation or the educational environment in general, have different background knowledge and experiences, or possess different sets of cultural values and beliefs, and therefore respond differently to test directions and questions.

**The Goal**

This question calls for recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners, as well as for the broader population.

The primary goal of any effort to make content-area tests valid and reliable for English language learners is to reduce the level of construct-irrelevant variance stemming from English language learner status. That is, we should work to help ensure that the tests are accurate assessments of what English language learners know and can do in reading/language arts and mathematics. Of course, the elimination of construct-irrelevant variance assumes an understanding of “construct relevance.” This careful definition of the constructs of interest is a first important step. The relationship of the construct to English
language learner status is, of course, different for reading/language arts than for mathematics. For reading/language arts, proficiency in English is a fundamental and essential enabling skill. For mathematics, the situation is a bit more complex. If the construct of interest is *mathematical skill exclusive of language skill*, we should provide English language learners with as pure a measure of mathematical skill as possible. If the construct of interest is defined as *the ability to perform and communicate about mathematics in an English-medium classroom*, however, the task is more subtle: We should make sure that the English language load is low enough to minimize construct-irrelevant variance, even while recognizing that some degree of communicative competence in English is part of the construct.

Several additional goals can also be identified as elements that will support the general goal above:

» Maintain the inclusion of English language learners in the nation’s assessment and accountability system and continue to use the progress of English language learners toward meeting standards as an essential criterion by which the success of schools, LEAs, and states is measured.

» Provide assessments that will give meaningful information about all English language learners, including those whose performance levels are currently well below grade-level expectations. These assessments should provide accurate information both about English language learners’ current skill levels and about the relationship between those skill levels and grade-level expectations.

» Minimize the need for lower-proficiency English language learners to take tests that are inappropriate for them, ones on which they cannot perform meaningfully and will have a frustrating and discouraging testing experience.

» Minimize the need for English language learners to be double-tested by, for example, taking both an English language proficiency assessment with substantial reading and writing components, and also a separate reading/language arts assessment focused on reading and writing (unless there are meaningful differences in the information provided by the different assessments).

**Recommendations**

To help ensure that the assessment system enables English language learners to demonstrate their knowledge and skills in core academic areas, we recommend the following:

» Set the same standards for English language learners as for all other students in terms of what constitutes “proficient” or “advanced” performance. Do not, in any way, lower the level of assessments or expectations for English language learners.

» In the guiding documents for the assessment system (curriculum standards, framework documents, test specifications, item specifications, etc.) include construct definitions that provide clear operational definitions of the role of English language proficiency in the construct. For example, the September 21, 2009, draft of the *College and Career Readiness Standards for Mathematics* contain several standards implying that communicative skill is
part of the construct of interest (e.g., Core Practices 2, “Construct viable arguments;” Statistics Core Skills 3, “Interpret data displays and summaries critically; draw conclusions and develop recommendations.”). For these standards to be used in an effective assessment of English language learners’ skills, such statements should be elaborated to clearly define the level of linguistic skill called for.

» Enlist the meaningful participation of teachers experienced in working with English language learners in the development of all aspects of the assessment system including the standards, the framework documents, the test specifications, and the item specifications. Make sure that such experts participate on item review committees (for content and for fairness); in item analysis; and in differential item functioning (DIF) analysis.

» Establish a policy to help ensure the use of accessible language, providing guidance on how to minimize construct-irrelevant variance. The policy should specify that non-construct parts of the assessment (e.g., directions, many elements of the mathematics assessment, questions — though not passages — on the reading/language arts assessment) be phrased in language that is as clear and accessible as possible.

» Include English language learners as a distinct subgroup in all item tryouts (including one-on-one tryouts of item types), small-scale pilot tests, and large-scale pilot tests.

» Consider carefully the role of constructed response items requiring written responses in the test design. Because writing is the most challenging language skill for most English language learners, using tasks that require written responses to assess anything other than writing skill itself poses a considerable risk of introducing construct-irrelevant variance.

» When constructed response items are included, consider the needs of English language learners in designing the items, developing the scoring rubrics, and selecting responses for training and for public release.

» When scoring constructed response items, train raters to distinguish between construct-relevant and construct-irrelevant ways in which English language learner responses tend to differ from the responses of native speakers, and document those procedures.

» In designing the assessment system, include formative assessments sensitive to the particular (and varied) developmental needs of English language learners.

» Consider appropriate accommodations for English language learners (including English language learners with disabilities) throughout the test design and development process, including the design of the test delivery format. (See response to question 2, below, for a more thorough discussion of accommodations for English language learners.)

» Consider carefully the use of pictures, graphics, and other nonverbal forms of communication in the test design. Such alternate modes of communication can help English language learners understand what is asked of them but can also create barriers to the effective assessment of students with visual impairments.

» Provide valid information about the performance of students whose skills and abilities may be some distance below what is called for in the standards. (As noted below, adaptive testing may help in this regard.)
Questions on the Assessment of English Language Learners

Use scales that allow one to track student progress over time. English language learners at lower levels of English language proficiency have a strong likelihood of performing below grade-level expectations, and it normally takes several years for English language learners to acquire enough English to successfully function in an English-medium classroom. As a result, English language learners need an assessment system that both accurately measures where they are and also tracks their performance over time, providing information about improved skills even while those skills remain below grade-level expectations. Consider a hypothetical example in which scores on a content assessment are reported on a scale of 0 to 100 and the “proficient” standard is set at 50 for grade 5, 60 for grade 6, and 70 for grade 7. A newcomer may arrive during the grade 5 year with little or no English language proficiency and score a 15 on the grade 5 summative assessment. That student then makes good progress over the following year and scores 35 on the grade 6 summative assessment. She continues to make good progress over the following year and scores 60 on the grade 7 summative assessment. An effective assessment system will show this student for what she is — a success story in the making — rather than simply labeling her as “below proficient” in grades 5, 6, and 7.

Consider the needs of English language learners in all research and validity efforts. Gather information on the performance of English language learners as a subgroup and consider these data in evaluating the performance and suitability of items and item types.

Finally, it should be recognized that any assessment system focused only on content assessments cannot meet all of the assessment needs of English language learners. To fully serve the needs of English language learners, an assessment of English language proficiency or an English language proficiency component to the general assessment system will be needed. Such a system should have a principled connection to the content-area assessments and should include a placement instrument that can be easily administered and locally scored; appropriate formative assessments; and a summative assessment of student progress in the acquisition of the academic English needed to learn in the content areas and succeed in English-medium classrooms. If thoughtfully designed, the system would collect and report valuable information about English language learner students’ skills and abilities while minimizing their need to be double-tested or to take tests not appropriate to their skill level.

Technology-enabled Assessments and English Language Learners

Innovative assessment designs and uses of technology have the potential to be inclusive of more students, including English language learners.

In considering the potential impact of technology-enabled assessments, it is worth noting that English language learners vary in their level of exposure to technology. Some English language learners may not have the technology skills that can be assumed in mainstream populations (e.g., the ability to use a computer mouse or basic skills in word processing). In addition, novel test formats (or ones that require linguistically complex directions) may disadvantage those English language learners who have limited experience with standardized testing. However, in the very likely circumstance that the standards call for working with digital and other types of technology (as do several standards cited in media...
Three areas in which technology-enabled assessments have the potential to improve the assessment of English language learners are discussed briefly below.

» Computer technology can make accommodations for English language learners considerably more efficient and user-friendly. For example, a computer-delivered test can be designed so that appropriate students have direct access to an online glossary (via “mouse over” or by clicking on hyperlinks). Also, computer-administration platforms can be modified to provide different lengths of time to students for whom this is an approved accommodation.

» In presenting items calling for a written response, computer technology can present students with a range of planning tools, such as graphic organizers, to help them construct their responses. Such tools can be particularly helpful to English language learners, but it is important that the students have an adequate opportunity to become familiar with the tools before being tested.

» An adaptive testing model can broaden the range of the scale covered in a single assessment, allowing accurate measurement of the skills of those English language learners who are currently some distance below grade-level expectations. Adaptive testing should be designed so that English language learners are not automatically shunted into a lower-level test, but have access to test items or sections that both (1) provide accurate information about their current skill level, as low as it might be, and (2) allow them to demonstrate proficient or advanced skill levels. In reading/language arts, for example, adaptive testing might allow a model in which the same passage could be used with different sets of test items: For students at or near the skill level specified by the standards, the test items could assess higher-level reading skills (e.g., challenging vocabulary, extended reasoning, subtle points of tone and voice) while students with lower skill levels could read the same passage but answer lower-level questions (e.g., literal comprehension and basic inferences).

2) In the context of reflecting student achievement, what are the relative merits of developing and administering content assessments in native languages? What are the technical, logistical, and financial requirements?

Content assessments administered in a language other than English (i.e., native language assessments) represent one type of testing accommodation that has been developed for use with English language learners. A testing accommodation is defined as “support provided to students for a given testing event either through modification of the test itself or through modification of the testing procedure to help students access the content in English and better demonstrate what they know” (Butler & Stevens, 1997, p. 5). The main purpose of providing students with a testing accommodation is to promote equity and validity in assessment. An interaction hypothesis has been proposed to justify the use of accommodations (Sireci, Li, & Scarpati, 2003). This hypothesis states that an accommodation will lead to
improved test scores for students who need the accommodation, but will not have an effect on the scores of students who do not need the accommodation.

Native language content assessments provide direct linguistic support for English language learners since the language used for the assessment has been altered (Rivera & Collum, 2008). Note that native language content assessments may not be valid for use in situations where English proficiency is integral to the construct being measured. To date, the number of research studies that have evaluated the effectiveness and validity of native language assessments is quite small. A recent meta-analysis by Kieffer, Lesaux, Rivera, & Francis (2009) reported that the only study to date that has rigorously investigated the use of Spanish versions of content assessments was conducted by Hofstetter in 2003. Her study used grade 8 NAEP mathematics items and involved two samples of students: one set of Hispanic students instructed in Spanish and a second set of Hispanic students instructed in English. A strong positive effect for the Spanish version of the test was found for the students instructed in Spanish, but a moderate negative effect was found for the students instructed in English. Based on this study, the use of native language assessments does not necessarily lead to improved performance on the part of English language learners. Most importantly, unless the language of instruction matches the language used for the assessment, there appears to be little to no gain in English language learner performance. It appears that unless students are familiar, through instruction, with the concepts and terms used in a given language, the use of an assessment in that language does not provide any discernible benefits.

In order to use native language assessments, there are a number of technical and policy considerations that must first be addressed:

- **Comparability/validity:** Can an assessment that has been translated, or developed in parallel in a different language, be assumed to measure the same underlying construct or set of skills? Can we assume that scores from an assessment in two different languages have comparable meaning? A recent paper by Young (2009) specifies a conceptual framework for test validity research on content assessments taken by English language learners, and identifies eight separate indicators of test comparability, including reliability, internal test structure, DIF, and predictive validity. Similarly, Sireci (2009) has identified several quantitative approaches for evaluating test comparability, including evidence based on internal test structure and differential item functioning.

- **Test translation/transadaptation:** The technical problem of translating assessments from English into other languages is a perplexing one that must be resolved. In many cases, it is necessary to go beyond directly translating a test from English into another language, by adapting the content of the assessment to account for the sociolinguistic and cultural differences between the two languages (a process known as transadaptation) (e.g., see Stansfield, 2003).

- **Native languages for assessments:** States have provided assessments in a number of languages including Arabic, Cambodian, Haitian Creole, Portuguese, Russian, Spanish, and Vietnamese, but only Spanish is widely used (Sireci, 2009). The criteria for choosing appropriate native language(s) for assessments are unclear. If one criterion is that a certain
percentage of students must have a given language as their native language before the assessment is created, should that percentage be based on students at the state, LEA, or school level? Determining which assessments should be provided in another language is another consideration. Clearly, one would expect that assessments in reading/language arts should be administered in English, but should native language assessments be made available for all other subjects?

» Fairness considerations: If content assessments are developed for some native languages, but not for others, questions concerning equity and fairness will naturally arise. In addition, testing accommodations may not be equally effective for all students, such that the use of native language assessments may benefit students with higher levels of native language literacy than other students. Furthermore, students may be proficient only in speaking their home language, not in reading it, and the language of instruction and the language of assessment must be aligned in order for native language testing to be useful (Abedi, Lord, Hofstetter, & Baker, 2000). Lastly, some states, such as Virginia, do not allow students to be tested in a language other than English. How would conflicts between the potential availability of native language assessments and state policies, such as Virginia’s, be resolved?

» Financial considerations: Developing, administering, and validating a native language assessment has been found to be as costly as producing a completely new assessment, even if the native language assessment was developed as a transadapted version of an existing assessment. Test translation is more expensive than translation of other types of documents due to the many additional steps and extensive reviews that must be built into the process (Stansfield, 2003).

In addition to native language assessments, states currently provide other testing accommodations to English language learners, including the use of dual language assessments, English dictionaries/glossaries, bilingual dictionaries/glossaries, native language instructions, response accommodations, scoring accommodations, and simplified language (sometimes referred to as linguistic modification of test items) (Rivera & Collum, 2008; Solano-Flores & Li, 2009; Young & King, 2008). The meta-analysis by Kieffer, Lesaux, Rivera, & Francis (2009) reported that the use of English dictionaries/glossaries was the only accommodation they investigated that showed a significant positive impact on the performance of English language learners. Studies of the use of bilingual dictionaries/glossaries and simplified language have found mixed results, with slightly positive to no impact on English language learners. However, further research on these and other testing accommodations for English language learners is clearly warranted before any definitive conclusions can be drawn about their effectiveness and the impact on the validity of test scores of English language learners. We suggest that consideration be given to these other accommodations for English language learners, in addition to the possible use of native language assessments.
Question on the Assessment of Students with Disabilities

1) Taking into account the diversity of students with disabilities who take the assessments, provide recommendations for the development and administration of assessments for each content area that are valid and reliable, and that enable students to demonstrate their knowledge and skills in core academic areas. Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

Researchers at ETS have analyzed the test data for several state assessments and have conducted experimentally designed research studies to examine the impact of testing accommodations on students with disabilities. Research results have shown that current state assessments are unreliable measures for a large portion of students with disabilities because they are too difficult relative to the students’ current achievement levels. In some state assessments, the proportion of students without disabilities responding at chance level (or below) is less than 3 percent, but this percentage jumps to 10 percent to 20 percent for students with learning disabilities. In addition to inducing reliability issues for this population, these sorts of tests may have a negative impact on students’ emotions and motivation, as well as the ability of the test to accurately measure student growth from year to year.

One possible solution to this mismatch between test difficulty and student achievement level is adaptive testing, which we have recommended as a component of the system for other reasons as well. There are a number of positive reasons for using adaptive testing models with students with disabilities. One of the most important reasons is that such tests provide a better match of the difficulty level of the test to the achievement level of the student. This is important because providing an assessment that is better matched to a student’s achievement level will not only result in a more precise estimate of the student’s skills, but it will also result in a less frustrating experience for the student. In addition, it may be possible for states to use an adaptive test design to objectively route some students with disabilities to a modified assessment.

One design includes a two-staged adaptive assessment which measures reading comprehension (using a read aloud accommodation) and reading fluency separately for students with reading-based learning disabilities who perform at (or below) chance level on a short routing test. This type of test design has the potential for allowing states to measure proficiency level, while also providing additional information to teachers (scores for two separate components of reading), providing students with test content that is closer to their current achievement level, and allowing a portion of students to use a read aloud accommodation.

Although adaptive testing models have the advantage of targeting the difficulty level of the assessment to the students’ current achievement level, there are several disadvantages that are nontrivial for students with disabilities.
A potential disadvantage of adaptive testing may be the impact of divergent knowledge patterns in students with specific disability subtypes. For example, many learning disability classifications are defined by divergent cognitive profiles or lower achievement levels in specific academic knowledge areas or subskills. The implication is that students with learning disabilities defined by a deficit in mathematics fluency, for example, may perform poorly on relatively easy test questions that measure calculation but perform well on relatively difficult questions that measure estimation. The use of computer-adaptive tests in the presence of idiosyncratic knowledge patterns has been studied, and results show that scoring of adaptive tests is problematic when a test taker responds to questions in an unexpected way. Additional research would be required to determine the impact of this for students with disabilities.

Another disadvantage of implementing an adaptive test is that providing some testing accommodations can be problematic. This is particularly challenging in developing alternate format tests (such as Braille) for item-level adaptive tests because the selection of questions in an item-level adaptive test is based on the specific performance of the test taker on the previous questions. Therefore, it is impossible to assemble a test prior to administration. In addition, many computerized testing platforms do not provide magnification or prerecorded audio, and none of the existing platforms currently provide refreshable Braille. For these reasons, individuals who require Braille test forms do not currently participate in item-level adaptive tests. Instead, these test takers typically take an alternate paper-based linear form of the assessment.

Ultimately, adaptive testing, particularly multistage adaptive testing, holds promise for students with disabilities; however, it is not a panacea. Below are several recommendations for how the Race to the Top assessment program funds could be used to develop the infrastructure for delivering accessible assessments that target test questions to student achievement. If these research studies are conducted during Generation 1 of the program, the issues could well be resolved by the time we move to Generation 2 with large-scale, adaptive testing in the summative component of the system.

Specify that some portion of the Race to the Top assessment program funds be devoted toward the development of an open-source computer-based testing platform that is fully accessible to students with disabilities. This is no easy task and will require the collaboration of individuals with experience in assistive technologies (both developers and teachers), universal design of assessment, and the development of existing computer-based testing platforms. In addition, the Department has already invested funds to develop accessible computer-based testing platforms (NAEP Writing and NimbleTools®) so we encourage consultation with colleagues in the National Center for Special Education Research (NCSER), the National Center for Education Statistics (NCES), and the Office of Special Education Programs (OSEP) to build upon their progress.

Conduct studies to determine if adaptive tests (particularly item-level adaptive tests) accurately measure the achievement levels of students with disabilities.
» Do not wait until the common assessment is developed to start planning for the development of alternate standards and alternate assessments, test forms, and test formats.

» As the Department considers innovative test items and design features, consider the role of graphical material, animations, and other media in tests, and have a plan in place to maximize the adaptability of such materials and/or devise strategies to develop alternative item types to replace graphical types when needed.

» Conduct research studies to document that the scores on all test forms and formats are comparable (or suggest ways to improve comparability). This is particularly true for any innovative technology-enabled assessments which may be proposed.

» Develop test content in a format that allows testing vendors to easily render test content in alternate formats such as audio, Braille, and large print. This may involve providing text descriptions of graphics and providing audio descriptions, captions, and text transcripts for movie clips and animations.

» Consider adaptive test designs which would allow the scores from the common assessments and alternate assessments based on modified achievement standards to be reported on the same scale.

» Build upon developments in universal design and accessible assessments that have been funded by NCSER, OSEP, and the National Science Foundation (NSF), such as the National Accessible Reading Assessment Projects (NARAP). The NARAP Accessibility Principles for Reading Assessments, which includes supporting research evidence, can be found at http://www.narap.info/publications/reports/NARAPprinciples.pdf.
Questions on Technology and Innovation in Assessment

1) Propose how you would recommend that different innovative technologies be deployed to create better assessments, and why. Please include illustrative examples in areas such as novel item types, constructed response scoring solutions, uses of mobile computing devices, and so on.

Our response to this question is formulated as a set of recommendations.

**Recommendation #1**

*Start with a long-term vision (5 to 10 years out) for a next generation assessment system and, only then, work backward to a set of steps to get there, including significant near-term ones.*

Throughout this document, we have referred to “Generation 1” and “Generation 2” as the two major phases of the implementation of the Race to the Top assessment program. If we start with the definition of Generation 2, we can work backwards into what Generation 1 looks like, and conduct the tryouts and innovation labs for new components in the system during Generation 1, using the interim and formative components of the system.

The reasoning behind this recommendation is that it takes 2 to 3 years to create, review, pilot test, calibrate, and administer a new parallel form of a *paper-and-pencil multiple-choice* test. If 3 to 4 years is the end-state time frame for creating a technology-based, next generation assessment system, then the likelihood of achieving fundamental change is not going to be high.

**Recommendation #2**

*In that long-term vision (and to the extent possible in the incremental steps), focus on such critical ideals as using technology to:*

1a. Measure important competencies that cannot be measured well in pencil-and-paper testing.
1b. Help teachers (and students) adjust instruction and learning.
1c. Model effective teaching and learning practice, so that the assessment becomes a worthwhile learning experience in and of itself.
1d. Make assessment fairer for all students, including those with disabilities and English language learners.

We suggest focusing on ideals because the danger is that, if we do not, worthwhile near-term efficiency targets (e.g., improving score turnaround) may dominate to the detriment of more fundamental goals (e.g., measuring what’s important).

Each of the above subrecommendations, 1a through 1d, deserves further elaboration.

**Recommendation #2a:** Use technology to measure important competencies that cannot be measured well in conventional form. There are many ways in which we can use technology to
Questions on Technology and Innovation in Assessment

measure important competencies. Examples include having students use simulations of dynamic systems to interpret evidence, discover relationships, infer causes, and pose solutions; mathematically model problem situations with a spreadsheet; write on computer and read (nonlinearly) on the Internet; search for, and critically evaluate, information on the Internet; respond to reading or writing problems that require the integration of many text sources and of various document types (including nontext types like video and animation); fluently execute basic procedures (which can offer information that is formatively useful); carry out complex extended projects; and assemble digital portfolios of their work. None of these uses of technology should be done for its own sake but only if it is used to measure important competencies that could not otherwise be assessed.

**Recommendation #2b:** Use technology to help teachers (and students) adjust instruction and learning. When a student’s summative test performance suggests the presence of either an overall proficiency deficit or of specific skill deficits, we should at the least provide “formative hypotheses” that point teachers toward students or skill areas of need, upon which teachers (and students) should follow up. As an alternative, we might route the student to a targeted diagnostic assessment.

**Recommendation #2c:** Use technology to model effective teaching and learning practice. We might use technology to model effective teaching and learning practice by building into test questions tools that practitioners use, and that students should be using routinely, in the course of their classroom work. Examples include making planning tools part of writing assessments, embedding into reading comprehension questions graphical organizers and tables for representing complex text (with appropriate alternatives for students with visual disabilities), and asking students to complete concept maps for representing physical or semantic relationships.

**Recommendation #2d:** Use technology to make assessment fairer for all students, including those with disabilities and English language learners. For example, embedding definitional links for difficult words into test questions (where vocabulary knowledge is not being tested) ought to lower irrelevant knowledge requirements for English language learners. A second instance is providing for students with print-related disabilities alternate representations of question components (e.g., translating stimulus text to speech, or describing orally the graphical components of an item). Finally, we can offer alternate questions measuring similar skills at similar difficulty levels, when a class of questions is important but not suitable for some students.

**Recommendation #3**

*Understand the benefits and limitations of each technology before deploying.*

All technologies have benefits and limitations. For example, automated scoring is operationally faster and cheaper than human scoring, and sometimes able to provide feedback on instructionally actionable performance components. But in many cases, automated scoring uses limited proxy measures, like
sentence and word length and sentence complexity, to predict a human score, and practicing the proxies may lead to higher machine scores but not necessarily to greater learning. A second example can be found in adaptive testing. Adaptive testing measures with precision throughout the skill range. However, in current implementations, it measures only a subset of what is important to test, potentially having the same (unwelcome) effects on instruction as current multiple-choice assessments are said to have. Therefore, it is important to recognize that there are tradeoffs associated with new technology that are best made by informed choice, rather than by accident.

**Recommendation #4**

*Manage risk.*

Most successful transitions from paper-and-pencil to computer delivery have put substantial time into planning and many have used a phased approach to implementation. Examples include Oregon and Virginia. Each state now delivers about 1.5 million summative tests annually on computer at the primary and secondary level, including for Adequate Yearly Progress (AYP) purposes. But it took those states the better part of a decade to achieve that level. The main point is that moving a large-scale testing program to computer is a very complex undertaking requiring, among other things, hardware and software availability and compatibility in all schools, extensive LEA training, and student familiarization. Getting the appropriate infrastructure and knowledge into place takes considerable time and effort. Our proposal is to use Generation 1 (up to four years) for tryout, experimentation, and building of infrastructure, and then launch Generation 2 in either the fifth or sixth year, with greater assurance of success. There is no need to repeat the entire decade of development it took the pioneering states, but there is also a danger in assuming that the other states can learn without their own trial and error on a small scale.

**Recommendation #5**

*In the world of innovation, failure is a fact of life but one that can be put to beneficial use, so plan to fail but plan to fail early, often, small, and gracefully.*

The value of this type of controlled failure is that it will make clear relatively quickly that an approach is unworkable or, in the best case, help successively approximate over time a practical assessment system with the least cost and harm to all concerned.

**Recommendation #6**

*Fund multiple consortia so that significantly different assessment models (and uses of technology) can be explored and compared to one another, and consider giving preference to models that already have an existing theoretical base and that have been piloted.*

The assessment industry knows a lot about how to create innovative technology-based assessments, including ones that ought to have positive effects on learning, so we should build on that existing knowledge. However, we know a lot less about how to create innovative technology-based assessments that are affordable, practical, technically defensible, accessible, and fair to all students, so there is great
value in funding multiple approaches. Therefore, we recommend that the interim and formative systems be based on more than one approach and be tried out in innovation labs in a variety of participating states. The summative system is where we need the greatest level of comparability and standardization, so there would not be a variety of approaches on that system, although some experimentation can take place in the state-specific component of the summative assessment.

2) We envision the need for a technology platform for assessment development, administration, scoring, and reporting that increases the quality and cost-effectiveness of the assessments. Describe your recommendations for the functionality such a platform could and should offer.

We have also formulated our response to this question as a set of recommendations.

**Recommendation #1**

Our first recommendation with respect to question 2 comes in two parts.

*Recommendation #1a:* The platform should support the development, presentation, and scoring of assessments that represent as fully as possible not only the standards, but also the results of cognitive-scientific research because that research can help translate the standards to test specifications and to classroom practice. The research suggests the need to measure higher-order thinking skills (e.g., conceptual understanding, problem solving, reasoning, critical thinking, strategic thinking), lower-level components (e.g., declarative knowledge, automaticity), and problem-solving processes (which have value for formative purposes). Additionally, the research suggests the need for assessments to model the habits of the mind that are characteristic of proficient performers in the domain. These needs require that the platform have the capability to collect timing data (to measure automaticity), collect keystroke and mouse-click data (to measure problem-solving processes), and integrate tools and performance criteria into test questions so that students learn to use those tools and internalize those criteria in their work.

*Recommendation #1b:* The platform should support the development, presentation, and scoring of assessments that purposefully include dynamic stimuli (audio, video, animation), constructed responses of all types (written, spoken, digital representations of artifacts or of performances), simulations (e.g., of physical or social systems), information resources (e.g., Web sites, manuals), and scenario-based, extended exercises calling for the integration of multiple skills and knowledge components. The platform should also support the development, presentation, and scoring of traditional test questions. This array of competencies and tasks is necessary because the types of tasks encountered, and the competencies required, in workplace and advanced academic settings cannot be effectively represented through traditional testing approaches alone. However, traditional approaches do have value for efficiently measuring some types of competencies and should also be included.
Recommendation #2
The platform should support frequent measurement, with the capability to aggregate information over time to form a summative judgment.

Frequent measurement could include multiple summative tests distributed across the school year; one or more standardized projects; electronic portfolios of student work; or combinations of these elements. We recommend that the platform support frequent measurement because we should be able to make more meaningful (and fairer) decisions about students, teachers, schools, and education systems if we combine evidence from multiple time points and from multiple sources.

Recommendation #3
The platform should minimize the influence of irrelevant factors on performance.

To minimize the impact of such factors, the platform should include student tutorials, practice tests, formative assessments, and instructional exercises that use the same interfaces, representations, and tools that are found on the summative assessments. The intention behind this inclusion is to give students multiple opportunities to become familiar with the mechanics and tools used in the test. In addition, the design of the platform should account for the needs of students with disabilities and English language learners in formulating these mechanics and tools. The end goal is to help ensure that test performance depends, to the maximum degree possible, only upon those aspects of student competency that are the intended targets of measurement.

Recommendation #4
The platform should support an advanced type of adaptive testing.

Traditional item-level adaptive tests require short tasks that are machine-scorable in real time. As a consequence, we should look toward new approaches to adaptive testing. One example is a traditional adaptive test section that routes students to an appropriate extended constructed response (ECR) section that, itself, might not be machine-scorable. A second example is a multistage adaptive test in which each stage consists of an extended, scenario-based task including both machine-scorable and ECR items, with routing from one stage to the next based only on the machine-scorable items. The rationale for this recommendation is that adaptive tests can provide more precise measurement than traditional tests for low- and high-performing students but, in their current form, adaptive tests omit the measurement of key competencies.

Recommendation #5
The platform should support online human scoring and automated scoring, as well as their combination.

Online scoring allows for geographically distributed human rating, with real-time monitoring of rater performance. We have discussed these benefits at length in response to question 4 under “General Assessment Questions.” At the same time, significant advances have been made in the automated scoring of many types of constructed responses including essays; short text responses; mathematics
equations, and numerical and graphical responses; and some types of spoken responses. Automated scoring can often provide, in addition, detailed feedback on student task performance. These approaches can potentially make scoring cheaper, faster, and better, especially when online human scoring and automated scoring are employed in combination.

**Recommendation #6**
The platform should make it easy to switch testing vendors or use multiple vendors.

The platform should represent test questions and automated scoring models in common formats (e.g., SCORM, QTI) so that questions and scoring models can be moved from one vendor’s system to a subsequent vendor’s system without undue time, cost, and effort. The rationale for this recommendation is that states should be able to make vendor selections without having to bear the cost of repeatedly converting test content and scoring.

3) How would you create this technology platform for summative assessments such that it could be easily adapted to support practitioners and professionals in the development, administration, and/or scoring of high-quality interim assessments?

While we present more recommendations for the platform’s development here, we must point out that we do not recommend that a totally new platform be designed or built from scratch for administration of the summative assessment of common-core standards. This will be a high-stakes testing program in many respects, not only because of its accountability uses, but also because it represents the product of a tremendous amount of political will from the states and their stakeholders. It would be a tragedy for it to fail for reasons of technology and operational glitches. Version 1 of any software system usually has many bugs, and the kinds of systems we have described here as being necessary for the administration of the Race to the Top assessment program are very complex, with thousands of parts. Add to that the complexity of the technology infrastructure — computers, routers, servers, Internet connections — that must work together nearly flawlessly for a successful administration, and it becomes obvious that there are many places to fail in the rollout of the Race to the Top assessment program. We strongly recommend that the tests be administered using proven, existing systems, with desired enhancements being provided so that they are open-architecture applications that can be combined with existing platforms, provided those existing platforms are also QTI- and SCORM-compliant, open-architecture systems, as described in the answer to the previous question.

We treat this issue further in the answer to question 4, below.

**Recommendation #1**
The technology platform should allow for the possibility of making the interim-assessment part of the summative system, as well as for incorporating other sources of evidence like projects and portfolios.

The interim assessments, themselves, should be composed of an even greater mix and variety of innovative and traditional tasks as found on the end-of-year assessment, especially in Generation 1. Even in Generation 2, the need for faster score turnaround, greater comparability, and lower cost will
Questions on Technology and Innovation in Assessment

put more constraints on the summative component of the system than on the interim and formative components. In addition, the interim assessments should incorporate learning progressions, where such progressions are available. Finally, the interim assessments should be constructed such that they are learning experiences in and of themselves, not just tests.

The intention of this recommendation is to distribute the evidence used for summative judgments over additional sources, reducing the influence of a single, end-of-year assessment and employing the same model teachers routinely use to award course grades (i.e., take an average across quizzes, a midterm, final, and other sources). One should note that, in such models, the more interims (or other sources of evidence) there are, the less each counts individually. Interim assessment that is part of the summative system would more frequently model for teachers and students the competencies and tasks that are critical to proficient domain performance, and the learning progressions that are likely to lead there. Finally, such a model would give timely (but preliminary) formative feedback, pointing teachers to students and areas of need on which teachers (and students) should follow up.

**Recommendation #2**

*The technology platform should have the capacity to offer a variety of teacher-optional, curriculum-embedded, formative-assessment materials linked to the standards and to the summative assessments (as embodiments of the standards).*

In making this recommendation, it is important to note that interim assessment and formative assessment are distinctly different entities. The interim assessments primarily serve to help identify and document overall student status, whereas the formative measures are intended to provide more specific and targeted information for day-to-day classroom learning needs.

For the formative-assessment materials, the platform should have the capacity to include traditional items targeted at specific component skills; innovative tasks, projects, and portfolios targeted at skill integration, problem solving, reasoning, critical thinking, and conceptual understanding, among others; scoring rubrics to identify characteristics of good performance to teachers and students; exemplar student responses illustrating different score levels; pointers to additional, relevant instructional resources; learning progressions linking items, tasks, and instructional resources to the standards; and, finally, guidelines for teachers on a suggested process for using traditional items and innovative tasks for formative-assessment and instructional practice.

The rationale for this recommendation is that, on its own, interim assessment is insufficient for supporting classroom assessment needs. Teachers (and students) need curriculum-relevant items, integrated tasks, rubrics, interpretive materials, and instructional resources that they can use on a daily basis if they are to focus on, and make progress toward, achieving the standards.
**Recommendation #3**

The technology platform should have the capacity for teachers to add, modify, and share formative materials.

Teaching contexts and student populations vary, so the ability to customize is important. But many contexts and populations are similar enough that contributions by one teacher may be useful to other teachers, so mechanisms for sharing are critical.

**Recommendation #4**

The technology platform should have the capacity for teachers (and students) to formatively score constructed responses of all types.

The platform should be able to present rubrics, exemplar responses illustrating score levels, qualification sets (so that teachers and students know how well they are judging responses), and tools for annotating responses and recording scores. The intention behind this recommendation is that teachers and students can develop a shared understanding of what makes for good performance in a domain through scoring, particularly through identifying the features in responses that make those responses of higher or lower quality.

For the technology “platform” vision you have proposed, provide estimates of the associated development and ongoing maintenance costs, including your calculations and assumptions behind them.

As mentioned in the answer to the previous question, we do not recommend a new platform be developed for the administration of the Race to the Top assessment program, particularly for the summative assessment of common-core standards. Using existing platforms would not only save on the cost of developing a totally new system, which is not needed, but would also help ensure that the administration will work with less chance of failure. Of course, these systems may need to be updated to meet the needs of the new assessments.

The cost of developing the system would therefore be the cost of developing new enhancements or applications that would bring existing systems into compliance with what is needed to administer the new tests. These costs cannot be estimated until an inventory, or gap analysis, is done between the features and functionality of the desired system and existing test delivery systems. When that gap analysis is done, it will likely be different for each of the different existing systems. For example, Delivery System A from Company A might have gaps that would cost $2 million to close, while Delivery System B from Company B might have gaps that would cost $10 million to close. We would have to determine how much of that expense is reasonable to undertake as part of this development project, and how much should be left to the testing providers to close using their own sources of funding.

This issue interacts with the issue discussed in the section below “Project Management Questions,” and question 1 in particular. Our recommendation is for the Department to use the Race to the Top assessment program funds to design, build, research, and test the assessment system, but not to
Questions on Technology and Innovation in Assessment

operationally administer it consortium-wide in the states. The actual administration of the assessment in “live” census administrations should be done by the states using their normal procurement policies in a competitive marketplace. Because we are recommending an entirely computer-administered assessment in Generation 2, it means that states would procure services for the computer administration of the tests by qualified vendors of their choice. Of course, there would have to be some central repository and “keeper” of the assessment content and data, and some process for approving the vendors who are qualified to administer the Race to the Top assessment program. This is not an unusual need, and could be handled in a manner similar to how the GED® Tests are administered by providers authorized by the American Council on Education. In this model, the central repository contractor would be responsible for sending assessment components to authorized vendors in QTI-compliant “ready to use” format.

Assuming that there is a distributed delivery system to allow competition for administration and other services in the states, the Department would have to decide what percentage of the Race to the Top assessment program funds would go toward build-out of the system to facilitate the proper administration of the assessment with all of its attendant educational advantages. The modules developed by the consortium of states using the Race to the Top funds could be designated as open-source, open-architecture modules available to any providers of online test administration.

So, as far as ongoing maintenance costs are concerned, those would be the responsibility of the testing companies providing the services to the states, or the states themselves. Those dollars could come from the Elementary and Secondary Education Act (ESEA) funds, enhanced with states funds, that are normally used for administration of statewide tests required under the ESEA.
Project Management Questions

1) Provide estimates of the development, maintenance, and administration costs of the assessment system you propose, and your calculations and assumptions behind them.

The costs of a common-assessment system are very hard to predict in the abstract, as long as certain decisions remain unmade. We recognize that getting approximate estimates of cost are important to policy making and other decisions, but it would not be very helpful to say that the development costs could range from $10 million to $400 million, and the operational costs for administration and maintenance could range from $10 per student to $200 per student. However, the options we have heard considered and the range of possible decisions about the structure and size of the assessment could actually produce costs in that range. We would be pleased to provide more specific cost information once some decisions have been made.

What we have provided here is a description of how various characteristics of the program would affect cost.

First of all, it is important to separate development costs from maintenance costs from administration costs. While they are strongly related, there are some very important differences. For the purposes of this discussion, we assume that “development costs” include the design and initial creation of the testing system, up to but excluding its first “live” administration (when it is administered statewide in each participating state in the consortium, and the scores count for accountability). “Administration costs” include the live administration and the costs of delivering the assessments to the computers at which the students will be tested, as well as all the support services required, including constructed response scoring, customer service, and psychometric work and analysis to produce scores and summaries that take place each year the test is administered. “Maintenance costs” include maintaining and refreshing the item bank and keeping the software systems up to date.

**Development Costs**

Development costs depend on the type of assessment, number of grades and subjects, and number of forms to be developed. The number of students to whom the assessment will be delivered is not a major factor in developing paper-and-pencil tests, but it would influence the number of items needed in a computer-administered test if the ratio of students to computers is so high that it calls for an extended testing window. We must also consider the number of students needed for pilot/field testing, which depends on the pilot/field test design, assessment design, and type of administration is a factor. Other significant factors in development costs are associated with meetings, honoraria, and travel for teacher committees, administrators, and test developers. We have seen very large differences in the development costs for state assessments based on the state’s preference for how many meetings of in-state professionals are required. Can these meetings be conducted using technology, or do they have to be face-to-face? Are there meetings of special groups like bias committees and other stakeholders?
While the testing company can accomplish many of these activities highly effectively outside the state, frequent local meetings provide the great benefit of getting buy-in from in-state stakeholders.

Item tryouts can be very costly if they require a special testing occasion, which is likely during the initial development phase of the Race to the Top assessment. In order to get the best data, the students should be tested as close to the time of year as possible as when the assessment will be administered in a live program. That process could be quite intrusive in the Spring of 2011 and 2012, when states are still administering their existing assessment programs. Once the test is in operation, new items can be tried out by embedding them in the live assessments as they are administered online, but it is likely that the first phases of this new assessment would require a special administration and accompanying costs. Another factor to keep in mind is the difference between the new common standards and the standards previously in place in the states. If they are different enough, there will be an effect on student performance coming from the teachers’ ability to teach the new standards and the availability of instructional materials and resources. This might cause the need for recalibration of the item bank over time.

A significant driver of development and maintenance costs is the size of the item bank. The larger the item bank, the more items there are that need to be developed, which increases cost. Factors that determine the size of the item bank include obvious ones, such as the number of standards that are measured and the number of items per standard that are needed for reliable scores (which in turn depends on the level of detail desired in score reports) — but also less obvious ones, such as the length of the test administration window and the item release strategy. The longer the testing window (because states test in different weeks or have to spread out testing over many days because of lack of access to computers), the more items are needed to protect security. The larger the portion of the test item bank that is released each year after testing, the more items needed to be developed to keep a live item pool for building the next year’s pool.

Other factors that impact cost are the type of stimuli used and whether they are proprietary (i.e., requiring payment of fees for use). Many state assessments demand the use of published literature for reading comprehension passages, for example, and this demand involves a permissions process and payments of sometimes large fees for the use of that intellectual property.

**Administration Costs**

For administration costs, the number of students being assessed is a primary driver, as is the type of assessment (e.g., scorable by computer or requiring human scoring), and method of administration (computer-administered versus paper-and-pencil). Because we are proposing that the entire Race to the Top assessment be computer-delivered in Generation 2, except for special accommodations for certain students, we will not treat paper-and-pencil costs here, but would like to point out that they can be significant, especially for special forms like Braille and large print.

During Generation 1, we will likely have to allow for some states and even some LEAs to administer the test via paper and pencil. This introduces another whole set of costs to the system: composition,
printing, packaging, distribution, retrieval, document staging, scanning, and editing. The cost of a dual paper- and computer-delivered system is quite high, but might be unavoidable for at least a few years.

It costs more to have people read and score test results than to use a computer algorithm to score the tests. While automated scoring of constructed response items is possible and is widely used for certain item types, it is still not sufficiently developed to score all kinds of items and tasks likely to be required on the type of assessment we propose. We assume that at least part of the assessment would need to be scored by trained human raters, and that some of those raters be teachers. The size of the labor effort and the level of expertise needed by the raters is a significant factor in the cost. If the test is given online, much expense is saved in presenting the responses to the raters for scoring in an automated, or even distributed, system, but this is still a major cost.

**Maintenance Costs**

In many ways, the maintenance costs are driven by the ongoing decisions about item pools and assessment administration modes. The costs that drive those factors will obviously affect the ongoing maintenance costs of the system.

Estimating the cost for state-administered summative assessments is a complex task. We would recommend that the Department, in the models it promotes and funds, take full advantage of the competitive free-market system that exists today to spur innovation and drive down costs to states for test administration services.

Many think that a common assessment across, say, 30 states would greatly reduce the testing costs for those states. While this might be true for some, but not all, development costs, it is not true at all for administration costs. As a matter of fact, a common assessment might result in increased administration costs under certain conditions. If, for example, having a common assessment forced a narrower testing and scoring window across the participating states, this might force testing companies to increase their capacity for scanning documents in a system that relied on paper as well as computers during Generation 1, because more documents would be scanned in a shorter time. These increased costs for infrastructure would have to be passed on to states. If states wanted to keep their existing, spread-out testing windows, the current infrastructure is adequate, but that might require the development of more alternate forms to protect test security, which would increase development costs.

We are recommending that the entire program be computer-administered by Generation 2 and as soon as possible in Generation 1, except for some students with special needs. This creates different cost variables. Printing, shipping of paper, and scanning and paper handling are all but eliminated, which saves money, but there are additional infrastructural costs associated with the computer administration that can be significant. Again, we recommend that the states doing the common assessment create a structure that allows for multiple companies to provide test administration services for online administration, competing in the market to lower costs and improve quality and service. If this were the model, states can be sure to get the best service at the lowest cost, and also take advantage of innovations in delivery methods over time.
2) Describe the range of development and implementation timelines for your proposed assessment system, from the most aggressive to more conservative, and describe the actions that would be required to achieve each option.

A program as critical and complex as the Race to the Top assessment program must be designed and developed on a timeline that balances the need for an aggressive implementation plan and the necessity for a resulting assessment that is both valid and reliable. In addition, the timeline must also support input and involvement from a variety of stakeholders both from within the consortium of states and from the Department.

We are proposing that the system take place in two “Generations,” as follows:

**Generation 1**

The end-of-year assessment consists only of machine-scorable items, or very limited use of human-scorable items. The end-of-year assessment should be administered on computer where LEAs are ready to do so, or on paper otherwise. Note that this is a decision that needs to be made after careful deliberation. On the one hand, allowing for dual (paper-based and computer-based) administration would increase the participation rate in the new assessment system by states not ready to switch to computer-based testing, but it would also have negative consequences. The use of any paper-based testing in the end-of-year component in Generation 1 would have three limiting effects: (1) it would eliminate the possibility of using non-multiple-choice item types that could still be scored using automated scoring technology; (2) it would slow turnaround time of results because of the need for shipping, scanning, and scoring, and (3) it would cost extra money — both for the extra costs associated with producing, shipping, retrieving, and scanning a paper test and the extra costs of double production, the creation of extra forms, and the need to equate the two versions — that could be better spent on advancing other parts of the system. Because the first live administration of the Race to the Top assessment would not be until 2013 under the most aggressive timelines, our recommendations would be for states and the Department to make computer-based testing a condition for participation, and accept the consequences of some states deciding to wait it out. However, we are aware that many states feel strongly about having a paper-based option for Generation 1.

The interim-assessment system should be completely computer-administered, so that a variety of item types are possible. The interim system would also be computer-adaptive in some components that lend themselves best to that mode of administration. Data from the interim administrations is accumulated and added to the summative system for accountability purposes.

The formative-assessment system is computer-administered or paper-based, at the discretion and convenience of the teacher. The formative system is entirely for the benefit of the teacher and the learner, and data are not collected for any accountability purpose. Innovation labs are encouraged and funded throughout the participating states to try out new technologies and item types for eventual promotion into the interim system and then, perhaps, the end-of-year assessment for Generation 2.
**Generation 2**

The end-of-year assessment is computer-adaptive in all cases and consists of items that are scored by computer for the most part, including constructed response items. The interim system is the same as in Generation 1, and the formative system continues to be independent from restrictions, other than good assessment practice and alignment with the standards.

**Timelines**

We are assuming that the key part of the summative-assessment system would be administered in the Spring of the year, and that participating states would need some flexibility in the timing of administration that might result in a testing window from March through May. As pointed out elsewhere, the longer the testing window, the more items need to be developed to protect security, which increases cost. If the states could agree to a narrower testing window (e.g., April to May), we would have some relief in the development schedule in the more aggressive timelines. Also, if there would be an option of paper-based testing, all of these timelines would have to be pushed back by a year, because of an additional 4- to 6-month window required for preparation, printing, shipping, and other efforts. While the paper-based scenarios could be absorbed in the 2014, 2015, and 2016 timelines, those scenarios would be impossible in the 2013 timeline.

Also note that the 2013 timeline could not accommodate any interim assessments as part of the summative system in the first year of the administration of the Race to the Top assessment.

In Figure 1, we offer four potential timelines leading up to the first administration of the Spring assessment in 2013, 2014, 2015, or 2016. We strongly recommend that the 2013 timeline be considered unrealistic and that the earliest possible administration would be 2014 — and even that schedule is extremely aggressive. The reason we are presenting a timeline for a 2013 administration is to simply point out the steps that are necessary and show how much those steps have to be curtailed to achieve an administration date that soon.

Note that there are parallel tracks in each timeline for the development of content and the development of the technology systems to accommodate the new assessment. We have pointed out elsewhere that we recommend a system for administration that uses existing technologies for computer administration of the assessments, but there would still need to be some customization and additions to these systems. We believe that customization can be done on a parallel timeline with the content development. We should consider the content development path the critical path in this project, and it would dictate the amount of time that would be available for the technology work.

Below are descriptions of the steps we outlined in Figure 1:

- **Formation of assessment design and technology design teams** — include meetings with representatives from all states within the consortium to evaluate designs and needs
- **Assessment design** — include meetings with representatives from all states within the consortium
» **Creation of test blueprints and test item specifications** — include review and input from stakeholder groups representing each state within the consortium

» **Development of items** — include review and input from stakeholder groups representing each state within the consortium

» **Assessment of technology infrastructure** — work with each state to determine technology gaps for online delivery of assessments

» **Design and implementation of a plan to close technology gaps** — include a plan for schools across all states in the consortium and determine phase-in plan for online testing

» **Design and customization of the online delivery platform** — work with the consortium to develop the online delivery system

» **Provision of training to staff and students throughout the consortium states** — include training on the assessment design and content and on the online platform

» **Integration of online testing platform** with other data systems within consortium states

» **Small-scale pilot test of items and delivery platform** — administer items via the online system to small numbers of classrooms throughout the consortium and analyze

» **Field testing of forms** — include creation, distribution and retrieval of documents for paper-and-pencil field-tests and/or distribution via the online platform for computer-delivered tests, scoring, and analysis of results

» **Creation of operational assessments** — include review, reconciliation, and approval from stakeholder groups representing each state within the consortium
### Timelines for implementation of RTTT Assessment

#### 2013 Administration

<table>
<thead>
<tr>
<th>Content</th>
<th>Delivery System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form assessment and technology design team</td>
<td></td>
</tr>
<tr>
<td>Assessment Design</td>
<td></td>
</tr>
<tr>
<td>Design and Implement Technology Gap Closing</td>
<td></td>
</tr>
<tr>
<td>Creation of Blueprints and Test Design Specs</td>
<td></td>
</tr>
<tr>
<td>Development of Items (including reviews)</td>
<td></td>
</tr>
<tr>
<td>Train Staff and Students</td>
<td>Integrate Online Testing Platform with Other Systems</td>
</tr>
<tr>
<td>Small Scale Pilot</td>
<td></td>
</tr>
<tr>
<td>Field Test Forms</td>
<td></td>
</tr>
<tr>
<td>Create Operational Assessments</td>
<td></td>
</tr>
<tr>
<td><strong>LIVE ADMINISTRATION</strong></td>
<td></td>
</tr>
</tbody>
</table>

#### 2014 Administration

<table>
<thead>
<tr>
<th>Content</th>
<th>Delivery System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form assessment and technology design team</td>
<td></td>
</tr>
<tr>
<td>Assessment Design</td>
<td></td>
</tr>
<tr>
<td>Assessment of Technology Infrastructures</td>
<td></td>
</tr>
<tr>
<td>Design and Customize the Online Delivery Platforms</td>
<td></td>
</tr>
<tr>
<td>Development of Items (including reviews)</td>
<td></td>
</tr>
<tr>
<td>Train Staff and Students</td>
<td>Integrate Online Testing Platform with Other Systems</td>
</tr>
<tr>
<td>Small Scale Pilot</td>
<td></td>
</tr>
<tr>
<td>Field Test Forms</td>
<td></td>
</tr>
<tr>
<td>Create Operational Assessments</td>
<td></td>
</tr>
<tr>
<td><strong>LIVE ADMINISTRATION</strong></td>
<td></td>
</tr>
</tbody>
</table>

#### 2015 Administration

<table>
<thead>
<tr>
<th>Content</th>
<th>Delivery System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form assessment and technology design team</td>
<td></td>
</tr>
<tr>
<td>Assessment Design</td>
<td></td>
</tr>
<tr>
<td>Design and Implement Technology Gap Closing</td>
<td></td>
</tr>
<tr>
<td>Creation of Blueprints and Test Design Specs</td>
<td></td>
</tr>
<tr>
<td>Development of Items (including reviews)</td>
<td></td>
</tr>
<tr>
<td>Train Staff and Students</td>
<td>Integrate Online Testing Platform with Other Systems</td>
</tr>
<tr>
<td>Small Scale Pilot</td>
<td></td>
</tr>
<tr>
<td>Field Test Forms</td>
<td></td>
</tr>
<tr>
<td>Create Operational Assessments</td>
<td></td>
</tr>
<tr>
<td><strong>LIVE ADMINISTRATION</strong></td>
<td></td>
</tr>
</tbody>
</table>

#### 2016 Administration

<table>
<thead>
<tr>
<th>Content</th>
<th>Delivery System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form assessment and technology design team</td>
<td></td>
</tr>
<tr>
<td>Assessment Design</td>
<td></td>
</tr>
<tr>
<td>Assessment of Technology Infrastructures</td>
<td></td>
</tr>
<tr>
<td>Design and Implement Technology Gap Closing</td>
<td></td>
</tr>
<tr>
<td>Creation of Blueprints and Test Design Specs</td>
<td></td>
</tr>
<tr>
<td>Development of Items (including reviews)</td>
<td></td>
</tr>
<tr>
<td>Train Staff and Students</td>
<td>Integrate Online Testing Platform with Other Systems</td>
</tr>
<tr>
<td>Small Scale Pilot</td>
<td></td>
</tr>
<tr>
<td>Field Test Forms</td>
<td></td>
</tr>
<tr>
<td>Create Operational Assessments</td>
<td></td>
</tr>
<tr>
<td><strong>LIVE ADMINISTRATION</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1:** Potential Timelines for administration of the initial Race to the Top assessment in the Spring of 2013, 2014, 2015 or 2016.
3) How would you recommend organizing a consortium to achieve success in developing and implementing the proposed assessment system? What role(s) do you recommend for third parties (e.g., conveners, project managers, assessment developers/partners, intermediaries)? What would you recommend that a consortium demonstrate to show that it has the capacity to implement the proposed plan?

A successful consortium of states needs a detailed, well-defined organizing structure with a strong management plan. States should enter into an agreement or memorandum of understanding that explicitly lays out the consortium’s governance. We recommend that one state serve as the “lead state” to act as the fiscal agent to manage the grant money, and help ensure that the consortium meets all legal and regulatory requirements based on the states’ statutes and regulations. Only a state that has met all of the Race to the Top requirements should act as the lead state.

The consortium should establish a board with an internal organizing structure, clear voting rules, and defined operational procedures to manage their activities. The leadership should be collaborative and emphasize consensus building among the peer states. Regular timelines and meeting dates should be established for each phase of the effort with an annual evaluation of progress, with results shared among the state membership.

The board would select a project management team to work with the states and third party providers to manage the system components, such as grades 3 through 8 assessment development, high school assessment development, delivery and scoring/reporting, and outreach and professional development. The management team would also work with LEAs on a quality management and implementation plan, as well as a plan for how to engage the public and address concerns related to the new assessment system. Management team staff members should have a record of success in their area of expertise.

Third parties who support the consortium (technical consultants, assessment developers and psychometricians, operations contractors, and professional development support staff) need to have proven themselves in the marketplace as competent and successful. Each one should have a well-defined role by the consortium, and each should have the responsibility of sharing their activities with the entire third party team as well as the states so that all are aware of the totality of the activities.

As we mentioned previously, we believe the Race to the Top assessment program should be created in a way that encourages innovation, quality and service, and low cost from a competitive marketplace that is consistent with the procurement policies of the states. There is more than one way to accomplish this, but we would like to propose the following method:

1. The consortium is organized with a Lead State and Member States.
2. The Lead State, with the approval of the Member States, does a procurement to choose an Organizing Entity for the consortium. The Organizing Entity would be responsible for managing the business of the consortium, scheduling meetings, disseminating information, and other major duties.
3. The states in the consortium, with the help of the Organizing Entity, would write and submit the application for funding under the Race to the Top assessment program, following the Department's process.

4. Simultaneous with the submission of the application to the Department, the Lead State would initiate another procurement (e.g., Request for Proposal or RFP) for contractors needed to perform the work the consortium proposed in the Department application. The process would be managed by the Organizing Entity, which would be prohibited from bidding on any work in the RFP, so that the Member States had appropriate input into the choice of the contractors. We would recommend that contractors be permitted to bid on parts of the RFP, or the whole RFP, so that the consortium would eventually get a set of contractors who are the best at the work required, whether it is development, administration, psychometrics, technology, special forms, or other specialties. The award of contracts that results from this process would be contingent upon the approval of the application for Race to the Top assessment program funding by the Department. Note: It would make more sense if consortia bid with groups of contractors preselected, so that the Department could get a complete technical proposal with the application. However, we assume that many states’ procurement laws would not allow this.

5. Upon award of Race to the Top assessment program funding by the Department to the consortium, contracts would be finalized with the successful contractors and the work would begin and be conducted according to the process described in the application and approved by the Department.

6. At the appropriate time in the development process, the consortium would do another competitive procurement, again under the auspices of the Lead State, to choose a Maintenance Contractor. The role of the Maintenance Contractor would be to keep the item bank and test bank secure, refresh the item bank and test bank, develop and maintain a process for authorizing Approved Assessment Providers, and interact with Approved Assessment Providers to help make sure they have what they need to deliver the assessments to the states in the consortium and to new states joining the consortium over time.

7. When the system goes “live,” states would use their normal procurement processes to choose Assessment Providers. The states would add to their programs additional, state-specific elements, which would be provided by the Approved Assessment Providers. The Approved Assessment Providers would conduct all the activities for which the state needs a contractor, and would have access to the current year’s forms of the Race to the Top assessment from the Maintenance Contractor.

We believe that a process like we have described here provides the best quality, consistency, and fidelity to the purpose of a common assessment, while permitting healthy and robust competition and innovation in the field.
References


