THE NATIONAL CONFERENCE
ON
NEXT GENERATION ASSESSMENT SYSTEMS

High-Level Model for an Assessment of Common Standards

Stephen Lazer

Center for
K–12 Assessment
& Performance Management

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

High-Level Model for an Assessment of Common Standards[1]

Stephen Lazer

Educational Testing Service

Part 1: General Design of a Next Generation K-12 Assessment System

Educators generally agree on the need to develop improved assessment systems, and also generally agree that a common set of fewer, clearer, and higher (that is, more rigorous) standards provides an opportunity to do that. Working together, states may be able to realize economies of scale and other efficiencies that allow the use of more innovative exercise types and the pursuit of a more robust research agenda. Furthermore, advances in cognitive science, task design, psychometrics, and natural language processing combined with the wide availability of technology make it possible to assess a more meaningful array of skills and knowledge than ever before.

While there is general agreement on the value and potential of an improved system, there is far less consensus on the priorities for uses of the new assessments. A number of priorities have been posited by various stakeholders. For example, many (including those who first advocated assessments of common core standards) see one of the main goals as providing individual scores that are comparable across the nation (or at least across members of a consortium of states). However, some see local choice of assessment activities as key goals, which would tend to limit data comparability. Many state officials see it as essential that any system be economically and practically sustainable, and not impose new financial or logistic burdens on state budgets and school schedules. For others the most important goal is to ensure that new assessment systems support positive instructional practices and encourage increased rigor, with the overall aim of improving education and US economic competitiveness. Such systems would tend to involve complex, human-scored tasks and would seem at first blush to run counter to both cost control and another perceived good: providing faster score turnaround. Still other policymakers want to be sure that assessments provide solid measurement of student growth, and not be limited to the status indicators provided by many current assessments. Others view it as essential that these growth indicators be used to evaluate teacher and school effectiveness.

It is impossible for one assessment or even one assessment system to meet each of these goals equally well. For example, assessments that best support positive instructional practices and encourage

[1] This paper is based on a white paper that was composed by Research and Development staff members at Educational Testing Service, Pearson, and the College Board. However, revisions were made for this version by Stephen Lazer, and any errors or omissions are therefore his responsibility. This paper focuses on the elementary and middle-grades assessment system. The high school system is discussed in general terms only.

increased rigor are likely to impose new financial or logistic burdens on state budgets and school schedules.

There are, however, some goals that are shared among most policymakers, and those goals may form the basis of an integrated assessment system. First, there is a great degree of agreement that these new assessments should measure a common set of standards that are fewer, clearer, and higher than those now used in many states. While there is less consensus on the specific nature and content of the standards that should be assessed, the clear implication is that the assessments should focus less on breadth of knowledge and more on depth of reasoning than do the current generation of state tests. Specifically, stakeholders believe assessments should tell us whether high school students are likely to be college and career ready (CCR), and whether younger students are on track to reach such a state of readiness by the time they leave high school. There is a near universal agreement that new technologies should be used to support the assessment system, though there is less agreement on the specific uses of the technologies and the pace with which the new technologies should be added. There is general agreement that the new system must continue to provide high-stakes summative data to be used for accountability purposes, as well as some instructionally actionable information (there is far less agreement on the components needed to meet these goals). Finally, there is also clear agreement that some of the data will need to support high-stakes accountability uses, and will need to have some comparable meaning over time and in different places.

In summary, there are some goals on which all stakeholders seem to agree. There are other goals among which designers will need to make choices: a system that tries to meet all goals equally well will meet none very well.

This paper represents one attempt to design an assessment system to measure emerging common standards. It is not the only possible system, and it makes certain choices (for example, it argues for use of performance exercises even if this prevents immediate score turn-around). But it is one system that could function, and could satisfy many of the goals of policymakers and educators.

General System Parameters and Assumptions

In designing any assessment system, one must set parameters and make assumptions. For example, one could design a system assuming that 10 years are available to develop psychometric and automated scoring techniques, and that each student will have his or her own computer. While entirely sensible as a long-term planning exercise, it takes the design effort well outside the horizons envisioned by policymakers. On the other hand, one would not want to assume that all elements of the design could be implemented today, as this would prevent meaningful innovation.

Therefore, the system described in this paper is based on the following assumptions:

- Common core standards will be adopted by large numbers of states within 1-2 years.
- The common core standards will cohere across grades so that assessments of the standards will support meaningful estimates of student growth.

- States and/or consortia may want to measure some attributes beyond those covered in the common core standards.
- Universal computer-based testing will be possible in 3–4 years (special-needs accommodations may still be given on paper). However, this paper does not assume enough technology availability to be able to test all students in mass administrations.
- Major elements of the new assessment system must be made operational within 3-4 years. However, this initial operational rollout does not need to represent the end state system; more elements may be added at later dates.
- Efficiencies from pooled test development and psychometric work will make possible modest (though not major) increases in the per pupil operational costs of assessment, allowing some use of human scoring in the system.
- The goals of the common core assessment can only be met by an assessment system, and not by a single test.

The argument in this paper is also based on the assumption that any new assessment system must meet two overarching goals: (a) new instructionally-relevant measurement based on common standards , and (b)sound measurement that meets professional technical standards for high-stakes use. Unless the first goal is met, the new assessment system will do little to provide more nuanced information than current systems or to enhance learning. The second goal is equally important: the professional standards are not simply arcane guidelines meant to discourage innovation; they are rather ways to ensure that the scores and other characterizations of students (e.g., proficient) are valid and fair, qualities that should not be taken for granted even if the tests are based on tasks that appear to be instructionally relevant. Comparisons between places and over time may become tenuous. This does not, of course, mean that all tests and test components must meet a single numerical standard, or can meet all goals equally well. It does mean that these standards must be considered alongside emerging construct demands. The new assessment system must find an appropriate balance among these imperatives.

Discussion of the Elementary and Middle-School System

In the sections that follow, this paper will discuss the comprehensive assessment system for elementary and middle schools. It begins with a discussion of the need for integrated assessment systems. This is followed by an extended section on the end-of-year tests. Finally, the paper discusses the possibilities of using components other than the end-of-year test as part of summative/accountability systems.

Designers should focus on an integrated system that involves formative components[2] as well as summative/accountability assessments. In addition, the summative system may itself be composed of various components and not solely of a single end-of-year test.

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

There are a number of reasons to favor an integrated system. First, formative and summative components will likely both function better if built to work together. Specifically, they should be built to measure the same standards, and reflect the same curricular elements and learning progressions. They should be constructed using open technology protocols so that material can flow from one set of instruments to others. Second, an integrated system should relieve pressure for the summative tests to serve a purpose for which they are not ideally suited: to provide in-depth, reliable, and valid instructionally actionable data. This is particularly true at the level of individual standards or learning objectives, where coverage on any summative test will be, by necessity, limited. Attempts to provide such data from a summative test will increase pressure to lengthen tests. This pressure will become especially important since we believe the system should exploit technology for delivery, and longer tests will necessitate longer administration windows. An integrated system should prove far more likely to meet the varied goals people have set for the assessment.

While the ability of summative measures to provide formative data is limited, one could, in a carefully designed and integrated system, view summative assessments as providers of information to formative systems. This is easiest to consider in cases in which the summative system is not solely an end-of-year test. But even with single summative assessment events, there may be ways for those tests to inform formative systems. For example, even if small distinctions at the sub-score level lack instructional or statistical meaning for most students, it might be interesting and important to note those students who

[2] It is important that the use of the phrase *formative component* not be misleading. We do not envision formative assessments solely as fixed objects that are usable on their own. Rather we understand formative assessment as a process in which teachers employ assessments to make instructional decisions—the components described here would be libraries of tasks to be made available to teachers.

have outlier performance in some area (that is, students who do far better or worse than is expected in a specific area given their overall scores). In these cases, summative data might focus teachers on areas where more testing or diagnosis seems indicated (and thus call for the use of specific formative components). In the context of an end-of-year summative assessment, this could involve thinking across grades. For example, a summative result at Grade 5 could identify students who appear to be struggling in certain areas. Based on the specific nature of the results, the system might identify *diagnostic intake test* components that would be administered at the beginning of Grade 6 which teachers could use to tailor instruction. These intake tests would not go to all students but only to those whose Grade 5 results had indicated the need for further testing. This of course implies a summative system that allows some score disaggregation, and a quite different way about thinking about sub-scores (that is, that they are likely to be meaningful only for some students).

There are a number of different models for how an integrated assessment system might provide instructionally actionable information. An integrated system could include formal elements such as interim assessments, which are given throughout the year to get a snapshot of how students are doing in mastering the required skills. An integrated system could include diagnostic focused assessments, which provide more in-depth information on the gaps in student learning or performance. Probably more realistic are banks of tasks, assignments, and scoring rubrics coded to specific learning outcomes, available for teacher use. It is worth noting that any of these models assumes integrated delivery and management systems.

The formative assessment components of an integrated system allow for customization, differentiation, and local education agency involvement in development. While there are common standards, to the extent that districts and states use different curricula to address them it is possible that they will prefer to incorporate different formative systems within their instructional programs. Since the formative components will not be used to generate comparative data, such customization is possible.

One open question is whether accountability data will come solely from single summative tests, or whether data gathered over the course of the year can be part of a formalized accountability system. In the latter case, we can possibly increase the amount of instructionally actionable data that comes out of summative systems (although not to the point where it obviates the need for formative systems) and improve the quality of the summative data. This will be addressed below.

The remainder of this paper focuses on the summative/accountability system.

General Structure of the Summative System at Elementary and Middle School

This paper discusses a summative/accountability system rather than a single summative assessment. While the system proposed here includes an end-of-year test, we believe summative accountability data may be provided by both end-of-year tests and other instruments. Such instruments could include tests given over the course of the year and projects completed by students during the school year. We further believe that such periodic measures may or may not be part of the system at its initial roll-out, but may need to be added over time. The ease of adding these components depends largely on the degree of curricular consistency that comes with newly accepted standards.

End-of-Year Tests at Grades 3–8: General Comments

We assume that the summative/accountability assessment system will include (though not be limited to) end-of-year English language arts (ELA) and mathematics tests at Grades 3 through 8, all of which need to produce individual scores as well as aggregate scores. As we discuss below, these end-of year tests may not be the only components of the summative system.

In considering such a system, two seemingly opposite questions come to mind. The first is, Why test every year? The second is, Why give any special status to an end-of-year test. The answer to the first question is fairly straightforward, while the second is more complex. Annual testing between Grades 3 and 8 will be needed to support student growth modeling, which we believe to be a key goal of the new system. Skipping years in the sequence would make such growth projections more difficult. In addition, two years is a long time to go without having some formal information that students are off track. And various other system goals seem to heavily imply annual testing.

The second question is more complex. One could easily think of systems where assessments were distributed over the course of the year; in fact, this paper proposes such a system. As discussed below, in one version of this system albeit not a version we recommend) the end-of-year test is nothing more than the last quarter test, only measuring the content covered at the end of the year and is weighted no more heavily than any other periodic assessment. Alternatively, one could envision an assessment system being fully on-demand, in which students proved they had mastered the standards as soon as they were ready. Indeed, over time the system may evolve in this direction.

Then why do we argue for the need to maintain an end-of-year test as a key part of the system? If the year's curriculum is coherent, it is sensible to ask whether students can integrate different elements of that curriculum at a fixed point in time. While the appropriate point in time can be debated, schools are still organized by grade, and information about what students know and can do when moving from one grade to another is valuable. End-of-year tests also provide data at fixed points, which enhances comparability, a major goal of the system. Such testing should make the data easier to interpret for parents, teachers, and policymakers. For these reasons, we believe end-of-year testing should be kept, although we do not necessarily believe that these tests are the only possible sources of accountability data.

To get some sense of the scope of the project, this paper assumes these assessments will, at a minimum, replace the current generation of No Child Left Behind (NCLB) accountability assessments. Through use of technology, these assessments will provide a state-of-the-art range of accommodations to students who need them. Through use of computer adaptive administration, we should be able to tailor tests to individual students, which should allow us to reevaluate the need for modified (or 2%) assessments. While these new end-of-year tests will not obviate the need for either 1% or Title 3 tests, they should be designed to work in tandem with the latter.

We believe these end-of-year tests should have at least two major components, although it is likely federal funding will address only the initial one. Even though this design is based on the assumption that states adopt common-core standards, it does not assume that these are their *only* standards; the general guidance has indicated that states may augment the common-core standards with 15% of their own standards. Thus the common core assessment system must allow for two goals to be met: it 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 may be both common-core standards and state additions, the tests would likely have at least two major components. The first would be the test of common-core standards. This would be consistent across all participating states, districts, and schools, and would stand on its own as a data source. Note that we do not mean the same exact test form is required but rather the same assessment. The common components of the test will be designed to yield state, district, school, and individual results on the common-core standards and will not include state-specific augmentation. The second component could be composed of state-specific content or augmentations. These augmentations would be analyzed in tandem with common-core items to yield state-specific results.

Before moving forward, two notes of clarification are in order. First, while the paragraph above views the source of customization as the state, it could well be the consortium. In other words, if USED chose to fund multiple consortia, one model would be to ask consortia to work together on the core assessment of common standards, and allow them to customize the 15% components. The second point of clarification is that these variable sections could be used as places for state- or consortium-specific content designed to yield data at the individual student level, or possibly to provide useful state or district data in cases of hard-to-measure content.

Why do we believe that the common-standards components of the end-of-year 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. 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 delivery package and permitting states (or consortia of states) to add components as needed. Finally, this approach allows some states to decide they do not need state-specific content, without affecting the comparisons on the common components (which embedding items in the common core would risk).

End-of-Year Tests at Grades 3-8: Nature of the Instruments

Item and exercise types. There is always a danger in talking about exercise types to be used before the final determination of the standards to be measured. Decisions about the sorts and arrays of tasks that ought to be included on these assessments should be the result of a careful evidence-centered design (ECD) process in which we gather expert groups, review research, and identify the sorts of behaviors that would convince us that students have reached the defined standards. Simply stated, we want to use the assessment tasks or items that most appropriately measure the construct desired. Focusing on using very specific items ahead of knowing the specific nature of the constructs to be measured risks ending up with an assessment of the wrong content and skills. There is a great danger that some who champion specific kinds of tasks will try to find ways to fit the standards to those tasks rather than letting specific exercise-design decisions issue from careful consideration of the evidence required to support certain claims about student mastery of standards and learning objectives.

However, there is enough evident in various discussions about the nature of emerging standards to make some general points clear. The construct and measurement needs of the system will certainly require a range of exercise types, ranging from selected response, to short-answer, to more extended tasks. These will likely include, though not be limited to, scenario-based tasks, long and short constructed responses, tasks that involve the exercise of technology skills, simulations, and tasks that measure listening and speaking skills. This mix of exercises is particularly likely given the general goals of providing college readiness information, measuring problem solving and critical analysis, and exploiting the benefits of technology.

The advantages of mixed instruments are that they allow measurement of an array of competencies, while still maintaining some of the technical characteristics needed to ensure solid measurement. However, use of what may appear to be traditional item types should not suggest that the questions themselves will look exactly as they always have. Selected-response questions, for example, can effectively be used to measure more than recall or simple skills. Short constructed responses can be used to measure reasoning. And more extended responses can measure the ability of students to perform complex, multistep tasks. Other tasks will, by necessity, appear to be quite different than traditional assessment tasks. This is because we assume that the standards will reflect skills related to the use of emerging technology, and that measurement of these skills will be major feature of the assessment.

As will be discussed in more depth below, we believe that to make this system affordable and sustainable we should attempt to use automated scoring where possible. Various item types can currently be scored by machine, and assess student abilities to create equations and graphic representations and solve mathematical problems, and use graphic organizers to show their understanding of complex reading tasks. However, while these sorts of items will form an important part of the assessment system, the main priority is solid measurement of the emerging constructs which will require human scoring of some tasks. We should use automated scoring where possible, but not when the use of such scoring would compromise measurement of the construct of interest.

If we to succeed in developing improved measured, it is desirable 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, and determination of college readiness and the role they may play in college placement decisions. Finding the appropriate balance will be key.

During the design effort, other questions will emerge about the sorts of items and tasks that can be used which are simply not possible to answer at this point. These will surround issues like use of audiovisual stimuli, as well as interactive tasks involving spreadsheets and databases. One interesting matter that will need to be resolved early in the process concerns the inclusion of tasks that measure ELA standards for speaking and listening (if these are in the final version of any set of standards). This is not uncommon in current state standards, but these skills are rarely if ever covered in assessments (which are normally limited to reading and writing). We will need to decide how to assess in these areas as this has broad implications for test design, length, and administration.

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

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 per subject, while tests at Grades 5 through 8 will take 60–120 minutes (for both common and state-specific components). High school tests could, conceivably, take between 2 and 3 hours. If extended tasks are used, assessment time may need to exceed these limits.

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

- Emerging standards in both mathematics and ELA will likely define constructs that can only be measured through the use of technology. This is likely to be true in subjects such as science as well. Maintaining parallel paper and computer systems on which results were supposed to be interchangeable would effectively prevent measurement of such skills. If these skills are not included in summative assessments, they will likely not be covered instructionally.

- 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 use skills that could not be invoked on a paper test.

- Testing some skills (such as writing) on paper may yield invalid results in the future because students will do almost all of their class and personal writing on computers.

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

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

This paper assumes that the summative-assessment system in general, and the end-of-year assessments in particular, should make use of adaptive administration. A variety of approaches may be used for this purpose (e.g., traditional computer-adaptive testing (where routing decisions are made item-by-item), multistage testing, variable, or fixed-length testing).[3] The appropriate adaptive testing solution will depend on the content and structure of the exams.

Adaptive testing is important to this model for the following reasons:

[3] In this paper, we do not use the phrases *adaptive testing* or *adaptive administration* solely to describe tests on which adaptation happens on an item-by-item basis (although that is often what CAT connotes). Multistage tests in which routing decisions are made after collections of items are given are also included under this general heading as well.

- 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 utilize assessments pools that cover more rigorous standards, while at the same time continuing to gain some meaningful information about what lower performers know and can do.

- It may allow us to identify standards on which students are struggling without unduly lengthening tests. Particularly in ELA with a heavy emphasis on authentic reading, we believe variations in traditional CAT approaches (e.g., section-based or passage-based adaptivity) can be implemented in an advantageous manner. Again, this will allow for far more personalization than traditional assessments.

- It will allow us to get better bang for the buck out of open-ended/performance-based testing; more extended items can be targeted at students for whom they will provide the most meaningful data.

- It can be implemented in ways that allow for high security in the context of the sorts of extended testing windows that are likely to be needed.

One possible challenge is the use of items that require human scoring in an adaptive system. There are in fact ways to use such items. In a multistage system, for example, routing decisions can be made based on a machine-scorable stage, with performance or open-ended exercises requiring human scoring administered during later stages.

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

Constructed-response scoring. The new end-of-year system faces a very basic design challenge: the new assessment needs to be innovative and forward-looking, but at the same time be affordable and sustainable, and provide rapid scores. Meeting the former goal involves using items beyond traditional selected-response. To optimize the speed and cost-effectiveness of scoring these items, this model calls for the adoption of a range of strategies.

First, we should push the limits of what can be scored electronically, and work to achieve constructed-response items that can be scored by computer. Given advances in automated scoring, *machine-scorable* no longer equals multiple-choice. Perhaps most familiar computer-scoring applications in use today are essay scoring systems, which are used at high stakes in various settings. Depending on the types of items used in the new assessment, such scoring systems may or may not be usable as the sole score. However, they can provide a valuable audit function, or provide second operational scores. In addition, content scoring systems are making real progress toward scoring short answer questions. Certain types of items in mathematics can already effectively be scored by machine: as early as 2000 Bennett and his colleagues wrote about three constructed-response mathematical item types that could be scored electronically. These exercises—mathematical expression items, generating examples items, or graphic modeling items—call for answers that are numerical (in cases where multiple correct answers are possible), graphical, or are equations. While the items do represent some delivery challenges (equation editors remain hard for students to use), scoring can be accomplished automatically. In ELA, certain emerging item types blur the boundaries between multiple-choice and constructed response. For example, in an ongoing research project, ETS has been working with items asking students to use graphic organizers to show relationships among elements of a reading passage.

Of course, we do need to understand the limitations of automated scoring systems, both in terms of their accuracy and the construct representation implied by how they work. But so long as we keep these factors in mind, they can and should be important parts of the assessment of common core standards.

Second, while we will need to continue to assign numeric scores to student constructed responses, we must stop thinking about scoring as being limited to such assignment.. Technology can and should allow us to develop better ways to analyze data obtained from assessments and that augment simple scores on student responses. For example, the steps a student takes when writing an essay or engaging with a mathematical simulation can be captured by computer, and may be of great instructional use or interest. These sorts of data may be particularly valuable in a summative system in which components are given over the course of the year (see below).

Third, while some tasks can be machine scored, we must realize that emerging standards will likely necessitate the use of items that will require human scoring (at least for some number of years). If this is true, we will have to find ways to balance the need for these items with the other imperative to maintain the affordability of the system. 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.

Human scoring is, of course, useful in many ways. It allows use of items that are not constrained by limits of the current automated-scoring systems. Including teachers in the scoring process would also represent a powerful professional development activity, and serve as a way to demystify the assessments and give teachers ownership over the assessment system. Of course, teacher scoring in a system that may also be used for teacher evaluation will necessitate careful safeguards. Therefore, any final design will need to find ways to use human-scored items that optimize the instructional and professional development impact of those items, without placing undue burdens on the system. Without appropriate care and planning, professional development can easily and unfortunately devolve into drudgery. 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. Distributed scoring systems can allow mass involvement without the costs of face-to-face meetings, and can spread the burden over large numbers of participants.

One final note: It is important not to view teacher involvement in scoring as possible only if human scoring is used. In the case of items where automated scoring is used, teachers should be involved in the development of rubrics and the selection and verification of training papers (activities that are excellent professional development opportunities).

Analysis and reporting. Given the overall interest in student growth metrics (and the use of such metrics in teacher evaluation), the assessment should support cross-grade comparability of scores. This work will be greatly facilitated if the content standards and expectations are coherent across grades. In addition to supporting growth modeling, cross-grade comparability facilitates another element we view as desirable in the system: the ability of flexible administration engines to select out-of-grade content for either advanced or struggling students. We assume that this out-of-grade content will mirror the instruction the student has received regardless of his or her grade level or age. Note that use of out-of-grade content is forbidden under current rules of NCLB and a rule change would be necessary. In some cases, out-of-grade testing it is the best way to get meaningful information at the student level.

While the system will need cross-grade comparability of scaled scores, it will also need to support within-grade performance levels. In other words, the new system will report growth scores but will still need to support status indicators: Students, parents, and educators will still want to know if students are doing well enough, particularly if this suggests they are, or are not, on track to be college and career ready by the end of high school. This dual data need not pose a problem but simply must be considered as part of the work planning.

There are interesting questions that will need to be answered in this area. For example, while it is likely that some constituents will want to see tests at Grades 3 through 8 on a vertical scale (perhaps mistakenly thinking vertical scales are required for growth measures), it is not at all clear that high school tests can be placed on such a scale. It is also a bit more complicated to think about growth measures in general in high school. 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. Conversely, some may see a real need for data on whether or not high school students are proceeding as necessary.

It is worth mentioning that there are several ways to produce measures of growth and cross-grade comparability. How the requirements of specific growth models affect the system will need to be studied, and a specific plan put into effect.

Given the number of standards and the pressures on assessment time available, it would make the most sense from a measurement standpoint to establish any status scores (or proficiency scores) on the summative system as a whole and not at the level of specific standards. We will almost certainly need to produce sub-score and collateral information as well as disaggregated performance by standard; however attempts to set passing scores by standard may imply a different system than we currently imagine. In any case, reporting meaningful information at the standard level will become easier if new standards are fewer and more distinct.

Because this paper assumes that the new assessments will have performance standards, using appropriate methods and sources of information to set standards will be crucial. Standard setting is often not considered when designing an assessment, but the validity of claims made based on the assessment will be no stronger than the performance standards allow. Assessment designers should ensure that crucial evidence is brought to bear regarding topics such as what successful students around the world know and can do in different grades, and what sorts of texts 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. In other words, the content standards themselves must give us effective performance descriptors that developers can use, and then careful standard-setting procedures must be used to establish proficiency levels.

Stakeholders will likely require that the common-core standards assessments 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. The system could augment the judgmental linking with 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 they wish to make of the international benchmarks.

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

Given the desire to use adaptive testing, one might assume we would also recommend a pre-equating approach be taken to the analysis. It will certainly be necessary to calibrate the items to allow routing decisions, and we believe the system will eventually need to be geared toward pre-equating as allowed. One complexity associated with pre-equating, however, is the use of human-scored items. Pre-equating will only work if we can ensure that the scoring of the operational responses is of the same effective rigor as that used to calibrate the items; this will require very careful control over the human-scoring process.

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

Periodic Assessments and Project-Based Scores as Elements of the Summative/Accountability System

There are several reasons to consider including in the summative/accountability system components other than the end-of-year tests. First, covering a full year of curriculum will tend to make the end-of-year tests rather long, even if adaptive administration is used. Shorter tests, while still capable of producing overall scores, will by necessity give little information at the level of individual standards or learning objectives. Conversely, the need to shorten tests while maintaining technical/psychometric quality will tend to mitigate against use of the sorts of open-ended performance items so valued in assessment reform discussions. Periodic testing, used in addition to an end-of-year test, should provide us with much richer information on what students know and can do, particularly if we find meaningful ways of aggregating skills information across testing events. Second, tests at the end of year, by definition, provide relatively little data of instructional value. Data collected over the course of the year ought to do a better job (even though the summative components will never replace formative assessments).

There are several ways in which one could consider other assessment events or data sources as formalized parts of the summative/accountability system. In one family of approaches, there would be multiple standardized 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 nonoverlapping, sets of content and skills. For example, a three-hour test might be broken into three one-hour tests that would be given over the course of the year. In this conception, the end-of-year test would essentially cover the last third of the year. A similar possibility is to build assessments around discrete instructional units (even if those were not equally spaced over the course of the year). Because we believe there is a role for end-of-year testing, this paper does not champion this precise approach, although it could be adopted.

A variant on this approach is a system in which the end-of-year test did cover the entire year’s worth of content, but that earlier standardized tests covered content from the first part of the school year in more depth. This is the model we recommend, and 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.

These periodic assessments could take several forms. They could look very much like the end-of-year measure described in this paper, in that they would be mixed instruments and adaptive. They might use the promise of extra testing time to include more performance tasks, and would certainly be more focused from a curricular perspective than the end-of-year test.

There are obvious advantages to such approaches and real challenges as well. On the plus side, one would get some early-warning data on students from the summative system itself; students might be able to retake modules they have failed over the course of the year. Because such systems would allow more aggregate data, they might give more stable results. Such testing also necessarily forges a stronger link between assessment and instruction.

On the other hand, the challenges are real. Such a system almost certainly involves making decisions about the ways content and skills are to be ordered (or at least combined) in the curriculum, and this may be beyond what is currently possible. Indeed, whether or not a system such as this can be implemented depends largely on the degree of curricular uniformity that comes with common-core standards. If a grade's work is really organized around a small family of learning objectives that can be combined into logical aggregations, then this system becomes conceivable to implement. If not, there are real problems.

There are analytic issues as well. One would need to decide on appropriate weightings for the different components. Some have discussed more complex ideas about how to combine components (that is, beyond simple or weighted averages); however, such approaches would require the invention of new technical methods. Finally, the system would need to be prepared to deal with a possible conundrum. Imagine two districts obtained the same average scores on the end-of-year test. This result would normally be interpreted to mean that the two districts ended the school year, educationally, in the same place. Rating one district higher because of higher performance on one of the intermediate tests might be problematic.

Other issues may arise regarding comparability of results over time and place. If all participants in educational systems take the periodic assessments at the same point in their school years, the results ought to be comparable for accountability purposes. However, differences in the schedule or order of administration of periodic components might introduce comparability issues—at a minimum they introduce equating issues that go beyond those that currently face most assessment programs.

An alternative model, used in some other countries, is described below. There would still be an end-of-year test, but accountability scores would also use data from standardized projects conducted over the period of the course of study (for example, research papers, laboratory reports, or book summaries). Scores from these projects would represent a fixed percentage of the final summative score. These projects could be used instead of or in addition to the periodic standardized assessment.

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

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

The use of assessments or projects conducted over the course of the year as part of a formal summative/accountability system is an important idea. There are challenges to be met before such a system could be implemented, and the existence of such a system presupposes infrastructures for data maintenance and transfer that are currently beyond the scope of many states. Thus it is possible that states will want to first focus on the end-of-year tests, and add these periodic measures at lower stakes until we are ready to add them to the accountability system. We believe that strong end-of-year assessments will be part of the system. We also believe that they may not be the only elements and that the system available on day one may not be the final system. Periodic assessment and adding project-based assessment to the summative/accountability system will do much to improve the instructional value of assessment.

In closing, let us restate one basic fact. Periodic assessment and standardized project approaches will be easiest to implement if common standards lead to a common sequence of learning objectives and clusters of those objectives. Giving a midterm presupposes knowing what is taught in the first half of the semester.

Thoughts on High School Assessment

For the most part, this paper envisions a high school system that shares most elements with the system for Grades 3–8 described above. For example, we would imagine that formal instruments would use a similar array of item types to those described above (selected-response, short constructed-response, and extended-constructed response). The paper envisions a combination of automated and human scoring. We also are certain that high school assessments will need to be technologically delivered. Any high school assessments will almost certainly be longer than those used at elementary and middle school.

However, while some elements of the system are predictable, there are two possible models of high school assessment, and the decision about which to adopt will be a policy or political decision rather than a measurement decision, per se. These two models are the *end-of-domain assessment* and the *end-of-course assessment*.

Let us consider these options. An end-of-domain assessment would be a direct assessment of college and career readiness (CCR) standards. It would be designed to be given at the point in the curriculum at which students are expected to have mastered CCR skills (for example, Grade 11 or 12). There would be two assessments, one in ELA and one in mathematics. They would be designed to measure the full array of ELA and mathematics skills needed for CCR.

Given the likely challenging nature of the CCR standards, these tests would surely need to be adaptive, so that the system would both measure rigorous skills and still give us some meaningful ability to report on all students. Depending on the stakes of the test, it would likely need to be offered periodically, so that students who had failed to show readiness could try again.

There are advantages and disadvantages of an end-of-domain system. On the advantage side, one needs only develop and maintain two tests (ELA and mathematics). Assuming all students take these tests at fixed points in their high school lives, the results should allow for easy comparisons among students, schools, and states. If a fairly broad range of difficulties was built into the items in the adaptive pool, one might well be able to use the system in a manner that allowed for some growth measurement. For example, students might first take the assessment as incoming freshmen, and then again at a later point in their high school lives (in principle, adaptivity would prevent students from receiving content that they had not yet been taught). Because these tests are not linked to specific courses and teachers, immediate score turn-around may not be a key need. Finally, an end-of-domain test gives individual teachers, schools, districts, and states maximum freedom in making curricular decisions.

On the negative side, if one of the goals of the assessment system is to forge an improved link between assessment and instruction, end-of-domain tests have some real disadvantages. Because they cover content and skills taught in several courses, they will not be optimal to provide useful feedback to individual teachers. The same factor makes their use in teacher accountability systems problematic. In addition, models like the one we recommend above—in which periodic assessments are part of summative/accountability systems—are not possible in the case of end-of-domain tests (there is not specific course into which to embed the assessment components).

The alternative model—the end-of-course test—has much to recommend it intuitively. CCR can, of course, be shown by having taken a series of rigorous courses. Assessments can show that students have in fact mastered course content. If the assessments are developed properly, they can drive instruction in positive ways; most would agree that Advanced Placement examinations have had this sort of effect on high school education.

As with the end-of-domain assessments, we would recommend that these end-of-course tests be computer delivered. There is less curricular argument in favor of adaptive testing (those with no chance of getting the content are likely not in the course and therefore not taking the assessment), but it is still possible that this form of testing would give us some advantages (shorter testing times, greater information about failing students, etc.). If the end-of-course paradigm is chosen, we believe that the system should include periodic and project-based elements as described above, as well as the end-of-course test itself.

The end-of-course paradigm has advantages and disadvantages. This is clearly the way to forge the most direct link between assessment and instruction. If one values periodic assessment and project-based assessment approaches such as those recommended for Grades 3-8, then these make perfect sense in the context of an end-of-course test (and not in the context of an end-of-domain test). These periodic elements might provide useful information for making instructional decisions. While there may be more tests to develop and maintain in this paradigm, they only need to be made available at fixed points in the school year (at the end of the year for the end-of-course component, and in defined windows for any periodic components). If scores can come quickly enough, assessment scores might in this case contribute to course grades, possibly obviating the need for the teacher to give and score a separate final.

The end-of-course system has weaknesses as well. Data are comparable for all students within a course, but aggregation to the school level and higher becomes problematic. For example, if Algebra II is a key course for showing CCR, it is an unfortunate fact that not all students make it to that level. Clearly, one could compare schools on the percentages of students taking Algebra II and passing the test, but this is not quite the same as comparing aggregate scores for all students on a common test. Related to this, it may be challenging for end-of-course assessments to provide growth information (and this is especially true if the content of the courses differs). Some form of before and after testing is probably possible, although this seems, for a variety of reasons, to be a difficult route to go. One would need to decide on the array of courses to be covered through this form of end-of-course testing. Would it just be the courses that showed college readiness, or the array of courses offered in high schools or that represented the capstones of individual course taking? If just the former, we limit teacher accountability uses to only those teachers who teach the right courses. Finally, this model works well if we assume a high level of curricular uniformity—to support common assessments, courses of the same name must cover the same skills and lead to the same outcomes. This may be easy to conceive of in courses like geometry or chemistry, but Grade 11 English would pose a different family of problems in gaining consensus.

As has been argued above, both of these systems have distinct advantages and drawbacks. The decision on which goals to optimize will help in choosing a final system. Alternatively, one could make a strong case for developing both end-of-domain and end-of-course assessments. This might give one the strength of both systems, and would free the end-of-domain assessment to function solely as a summative assessment. Of course, this approach is expensive, and could lead to situations in which the results from the two systems gave conflicting information about individuals.

Whichever system is chosen, it is especially important at high school that we consider the need for provision of research evidence that supports intended uses of scores from the assessment system. Even if we start with internationally benchmarked standards, we will need an ongoing method for checking and updating these standards, and for making attendant changes to test specifications. We may also not be able to simply rely on those standards: if the high school tests purport to measure college readiness, we should plan to have some data validating that claim. There are various ways to obtain these data; the key point is that some plan to gather validity data should be part of the design from the beginning. Therefore, we propose that this system should include a plan for establishing and enhancing a validity argument.

Part 2: Evidence of Meeting Race to the Top Expectations

Note: This part contains information largely redundant with Part 1, but broken into components that address the categories provided in the template provided to authors by Educational Testing Service for the National Conference on Next Generation Assessment Systems. Rather than repeating all the text, this part at times refers to appropriate components of Part 1.

Rigorous Standards and Good Instructional Practices

There are several features of the proposed assessment system that will promote rigorous standards and good instructional practices. Since the promotion of rigorous standards and the encouragement of good instructional practices are somewhat different issues, we address them separately.

Rigorous Standards

Several features of this system will promote and measure rigorous standards. However, before discussing those features, it is worth mentioning an important if somewhat obvious fact: The promotion of rigorous standards presumes that the standards adopted are, in fact, rigorous. Evaluation of the common standards is beyond the scope of this paper, and for purposes of discussion we simply assume that the standards will be rigorous and internationally benchmarked.

What about the proposed system will promote these rigorous standards? First, systems value what is measured in summative/accountability assessments, and this system will take careful steps to ensure that the new assessments really measure the new standards. This will be accomplished through use of an evidence-centered design (ECD) process where educators will gather to determine what the standards mean in terms of claims we want to make about students at different grades. Claims are based not only on the standards but on a review of the results of learning-sciences research, which can help in identifying the processes, strategies, knowledge, and habits of mind associated with meeting those standards. Identification of these claims will lead to discussion of the specific behaviors that would provide evidence to convince us that we can make these claims about students. To use a concrete example, let us assume a standard called for students to solve multistep algebra problems involving two variables. The (ECD) discussion would then ask experts to determine what specific behaviors would provide evidence to allow us to believe that students had (likely) met the standard. This discussion of evidence then leads to a discussion of what tasks would provide such evidence (yielding task models, which are used to generate items).

Use of the ECD process helps in creating the validity argument for an assessment, precisely because it helps ensure that tasks and items are closely tied to standards. It is thus one of the first ways this system ensures that rigorous standards will be measured.

Second, the use of mixed instruments made up of selected response, short answer, and extended answer questions will also promote rigor. The choice of which question types to use is dictated by the ECD process because that process helps identify what behaviors constitute sufficient evidence to support a claim and what types of tasks are appropriate to providing that evidence. The use of open-ended questions will ensure that a range of higher-level competencies will be measured through the system. Of course, there can be bad open-ended questions, so steps will be taken to ensure that these questions are used to measure complex skills.

Third, we believe that the use of adaptive administration is a major way in which the system will promote rigor. This educational reform comes with a built-in contradiction: It champions more rigorous standards while we have evidence that many students are today struggling to meet the lower standards currently in force. Therefore, an assessment aimed solely at the higher standards could, unless designed appropriately, provide very little useful data about large percentages of students. This could, in turn, place pressure on assessment designers to lower the difficulty of their instruments. Computer-adaptive administration should provide a way out of this conundrum. Assessments can measure whether students have met rigorous standards, while at the same time continuing to provide accurate information on students who are performing at lower levels. In addition, adaptive administration allows us to get solid information about high performers who are exceeding the standards.

Finally, we believe that level setting and ongoing international benchmarking efforts are ways to ensure that the scores have a meaning consistent with the rigorous standards. This is discussed in *Analysis and Reporting* section in Part 1.

Promoting Good Instructional Practices

Several elements of this system will promote good instructional practices. First, one of the goals of the ECD process will be to create tasks that are instructionally valuable in that they not only measure important processes, strategies, knowledge, and habits of mind but also model good instructional and learning practice. Assuming we are successful, we would recommend a regular release plan in which assessment tasks (and information needed for delivery and scoring) are made broadly available. Second, the use of periodic assessment would further promote a strong tie between assessment and instruction. Third, teacher involvement in scoring should have positive instructional impact. Finally, the building of formative and summative components from a common assessment framework may well be the most effective way of encouraging effective instructional practices.

Technology

The system outlined in this paper is truly a technology enabled system: the paper proposes aggressive use of technology in a number of areas. Because these are discussed in depth above, we repeat only highlights here.

Computer-Based Testing

Because we believe that emerging skills will involve the use of technology, we believe that assessments must be conducted on computers (or similar technology). In other words, it is not simply that a pencil-and-paper test be delivered via technology; it is rather that the items will be written, in some cases, to take account of technology as part of the emerging constructs. This will in all likelihood, be consistent with the standards. For details, see the *Item and Exercise Types* and the *Computer-Based Assessments* sections in Part 1 of this paper.

Computer-Adaptive Administration

For reasons discussed in depth in the *Computer-Based Assessments* section in Part 1 of this paper, we recommend that at a minimum the end-of-year components of this system make use of computer adaptive administration (although not necessarily a system in which routing decisions are made item-by-item). This will allow for good measurement across the performance range, which is essential in the light of more rigorous standards. While a specific adaptive model cannot yet be identified, it will need to be one that allows for human-scored items so we are likely to use a multistage model in which such items are administered after routing decisions have been made.

Automated Scoring

The *Constructed-Response Scoring* section in Part 1 of this paper argues for aggressive use of automated scoring (which, of course, assumes computer delivery). This allows for an expansion of the types of items that can be used while ensuring that the system is sustainable from a cost perspective. This section also discusses the valuable information that can be gained if automated scoring systems capture and analyze information beyond a simple determination of the correctness of a response. For example, scored process information might become an important part of the system.

Distributed Scoring

While automated scoring will be important, in all likelihood not all questions will be machine-scored. There will, therefore, be a family of questions scored by teachers and/or other human scorers. To facilitate such scoring, to provide appropriate quality checks, and to reduce costs, the *Constructed-Response Scoring* section in Part 1 argues for the use of distributed scoring systems in which training, scoring, and monitoring are conducted over the internet. This should allow for the professional development opportunities inherent in scoring while at the same time restraining cost and allowing for rapid turn-around.

Open Architecture and Standards

This paper assumes that all assessment materials will adhere to open standards for such material (consider, for example, the QTI standards). This will facilitate transfer of materials, and easy state or consortium customization (see the *Item and Exercise types* section in Part 1).

Integrated Data Management Systems

The paper argues for integrated assessment systems and periodic assessments(see the *End-of-Year Tests at Grades 3–9* section in Part 1) and project-based components as elements (see the *Periodic Assessments and Project-Based Scores* section in Part 1) of the summative/accountability system. Both of these elements presume the need for an integrated data management system that makes tracking and entry of student data manageable.

Summative Assessments That Measure Growth and Project Readiness

As indicated in the *Analysis and Reporting* section in Part 1, a major feature of the system is the support of measures of student growth. Therefore, cross-grade comparability of scores (at Grades 3–8) will be a major feature of the system if, and only if, the standards themselves cohere across grades. At the high school level the system is more complex: Measures of growth will be possible if an end-of-domain assessment system is adopted. Growth measurement is far more complicated in an end-of-course system. This is not a reason to reject end-of-course approaches, in our opinion, but is a feature to be kept in mind when making a choice.

Projecting whether someone is on track to be CCR at various points in their school career is not the same as having growth data emerge from the assessments. One can have one without the other, although growth modeling may make it easier to determine how far behind a given student may be.

There are three ways this system will ensure this sort of projection ability, the first judgmental and the second two empirical. First, we presume each grade will have an associated set of performance levels on the ELA and math assessments, and that these levels will be set based on a priori available descriptors of what students of that age should be able to do if they are to be on track to be CCR (see the *Analysis and Reporting* section). Second, we assume that we will gather validity data from students in college to ensure that the levels set on high school tests do, in fact, identify CCR students (that is, neither showing too many students requiring remediation as CCR, nor too many students not requiring remediation as not CCR). Finally, over time we recommend that the system track students, so that we can establish empirical relationships between scores obtained at earlier grades and final attainment of CCR.

Accessibility

The system proposed in this paper will allow a full range of accommodations for students with disabilities. The use of technology should facilitate certain accommodations such as voiced administration. One set of complexities to be solved surround the use of adaptive testing, which is impossible in the case of certain accommodations such as Braille. In all likelihood, there will be linear forms needed for these accommodations.

The use of adaptive administration should allow for the new assessments to replace both the current main NCLB tests and the 2% tests. We do not believe that the 1% or Title 3 tests will be replaced by these instruments.

In the case of English language learners, final determination of accommodations available cannot be made absent policy decisions regarding the specific nature of the constructs to be assessed. At a minimum, we would ensure that all questions are reviewed to ensure there is no unnecessary linguistic complexity. It is certainly possible to provide bilingual or translated versions of mathematics assessments; doing so would depend on decisions made during the definition of the standards. Such approaches seem less appropriate in the case of the ELA assessment, where the construct is by definition defined as existing in English.

Technical Quality

We feel strongly that the model proposed here must meet the highest standards for validity, reliability, and fairness; in fact, we believe any solution must meet the standards set forth by the American Educational Research Association, American Psychological Association, and National Council for Measurement in Education for technical quality. This is particularly true of the end-of-year components, in which a use of a variety of item types and adaptive administration should allow for high-quality measurement using known techniques for evaluating validity and fairness. In addition, our proposal to have an ongoing validity research program strengthens this approach yet further.

However, there are challenges in implementing some of the more forward-looking elements of this design. The use of periodic and project-based elements *could* introduce the need for new statistical techniques to combine scores. We use the term *could* because some approaches could be as simple as averaging scores on a number of assessments that were taken in the same sequence in all places. However, even in these fairly simple approaches, the need for equating of these periodic elements will introduce some operational complexities. But if we choose to try to combine the results of periodic assessments by using approaches beyond weighted averaging, or provide choice in the order or selection of periodic assessment events, then the development of new psychometric approaches will clearly be needed.

One related issue is how much user choice will be allowed regarding periodic and project-based elements. Choice is obviously a good thing from a local perspective, but would almost certainly limit the comparability of the data across different places. When such comparability problems would become serious enough that we would effectively lose a common scale cannot yet be determined; if this occurred we might need to rely on the end-of-year components to make comparisons. Stated differently, loss of a common scale is not a trivial matter. Common scales are essential if we want to aggregate scores across students, and compare students, schools, and districts to one another. To the degree to which we lose the common scale, these comparisons and aggregations become, at a minimum, more tenuous, and at worst, meaningless.

Because of these challenges, the use of these periodic elements as part of the summative/accountability system may or may not be an element of the new system at its initial inception.

There are other technical issues that are present but we believe manageable. For example, human scoring will need to be moderated, but there are available techniques for that. Security will need to be sufficient to ensure that results are valid, but we believe our approach allows for that.

Reporting

The systems proposed here should allow rapid score turn-around, although we are not proposing a system that would allow immediate score turn-around, as we believe such an approach would undercut certain other system goals. The fastest score turn-sound would come from a system made up entirely of machine-scored items in which all tests were pre-equated. In such a system, students can get final scores immediately upon finishing a test. We do not propose such a system because it limits us to machine-scored items, which do not seem likely to be sufficient to us. They will not cover the entire construct of interest, and may not provide solid instructional models. Additionally, for reasons discussed in the *Analysis and Reporting* section above, pre-equating will almost certainly not be possible in the early years of this program. Finally, pre-equating and human scored items may not exist in easy harmony—at a minimum it requires some assurance that the operational scoring is occurring with the same degree of rigor as when an item was calibrated. This assurance involves extra scoring, and takes some time. Specifically, in the case of items requiring judgmental scoring of any type, it becomes necessary to have some papers from a calibration sampler rescored during operational scoring, to make sure that the scoring has not become either easier or harder over time. Without such assurance, it is possible that a change in scoring rigor would be misinterpreted as a change in average student performance.

Therefore this is not the fastest possible system. But there is much we propose that will optimize score turnaround. Technology delivery removes the need for shipping and scanning of books and answer documents. Distributed scoring will speed the process of dealing with constructed-response items. And the fact that there will be one system across a number of states will allow for the focused use of technical resources to speed reporting. So while this is not an immediate feedback system, it will be far faster than is now the norm.

In response to the template questions, it is worth mentioning that the system will allow for all appropriate levels of data aggregation. The possible problem here would be the provision of very open choice options for periodic or project-based elements of a summative/accountability system. This could limit the comparability of data, and thus create aggregation issues.

Informing Instruction and Leadership

For the reasons argued above, this system will serve the needs of informing instruction and leadership better than current systems. However, it is also essential to remember that, as discussed in Part 1, providing instructionally actionable information requires an integrated assessment system and not just a summative assessment or series of summative assessments.

Why do we believe that the system proposed here represents such an improvement? First, the system we envision is an integrated system with formative and summative elements built from a common framework and designed to work together. Second, if we are able to adopt periodic and project-based elements, the data that can help inform instruction at the student and classroom level should be available during the course of the school year. Again, the data may not originate simply from the summative components, but the summative components may indicate where deeper diagnostic formative testing is needed. Finally, we believe that some of the exercises themselves can be valuable from an instructional perspective, thus providing positive instructional impact.

The template raises further questions about other sorts of policy questions that can be addressed. As discussed above, the end-of-year elements of the system should allow for data aggregation to support subgroup comparisons. Under the new system, these comparisons can be made on growth as well as status indicators, which should be a marked improvement.

A far more complex set of questions surrounds using results from such a system to help measure teacher and school effectiveness. The provision of growth data should clearly help from this perspective, at least in the case of teachers who are responsible for ELA or mathematics instruction. However, creating statistical indices of teacher or school effectiveness involves many other issues beyond a student assessment system that can measure student growth. It is beyond the scope of this paper—or in fact the student assessment system itself—to resolve all the many and varied issues associated with measuring teacher and school effectiveness. Suffice it to say here that there are myriad complexities, some of which will be quite difficult to address.

Leveraging Common Standards and Assessments

The adoption of common standards and assessments will have a number of positive impacts on the system. First and foremost, test development and psychometric analysis costs should be substantially reduced, at least in aggregate. Let us say 30 states choose to use a single assessment system. This means that instead of 30 assessment development efforts, there will be one. To be sure, this will not reduce costs by a factor of 30, as the needs of the system proposed here will be greater than any one of the 30 it replaces. But they will not be 30 times as great, so money should be saved.

This saving could be used to support an improved system. For example, doing a single development effort would allow focused research on tasks that measure skills involving technology and tasks we believe are needed to measure emerging standards. In addition, research in support of psychometric models related to periodic or project-based elements would only need to be done once. In general, a common assessment would allow for pooled research and development efforts, systems, and management in ways that introduce substantial efficiencies.

Pooled development should also help save some funds to pay for ongoing operational costs, which with human-scored items could become a factor. However, since development and psychometric costs are usually a modest portion of total program expenses, they will be unlikely to defray all increased operational costs. Further ways to control costs are discussed below.

Implementation Timeline

As discussed in the *General System Parameters and Assumptions* section in Part 1, we believe that the basic operational system can and must be implemented within 3–4 years. By *basic system*, we mean the end-of-year computer-based tests, including innovative exercise types, a full range of accommodations, automated scoring where appropriate, and distributed scoring where not. We would also have growth measures, performance levels, and libraries of formative materials in place. Finally we will have periodic assessments and project-based measures ready for the initial roll-out. What is not clear is whether we will be ready to make those elements parts of the summative/accountability system at the time of the initial roll-out, or whether these elements will need to be added to the summative/accountability system over time. The answer will depend largely on the amount of curricular uniformity that comes with adoption of the standards.

Cost

We have not at this point projected either initial development cost or ongoing system maintenance costs. It is important to note that the general design we suggest can be adjusted to minimize ongoing costs, if necessary. However, such actions might limit the skills that could be measured. In addition, the system we propose has substantial start-up costs, but with careful planning these can lead to efficiencies later.

While there is much that cannot yet be determined, some basic facts are clear:

- Development efficiencies will be real, and will be greater depending on the number of states in a consortium. All individual states now face development efforts. This system would replace the development efforts of each of those states with a single effort, albeit a larger one than any single state now undertakes. The savings should be substantial in aggregate, and especially noteworthy for small states.

- Scoring costs are likely to be higher in this assessment than in many current state assessments, due to the use of human scored items. Costs should not be meaningfully higher than state assessments that use multiple human scored items. However, these costs will be somewhat lower than one might expect because certain fixed expenses (development of rubrics, calibration papers, and training materials, for example) will need to be accomplished only once to serve the entire consortium. Development and psychometric efficiencies will offset much of this cost, though perhaps not all of it. Of course, a design parameter could be set to limit ongoing total costs to no more than the aggregate now paid by consortium states, but depending on the number of states involved this might limit design options.

- Many of the costs associated with the program are start-up rather than ongoing costs, including development and initial calibration of exercise pools and obtaining equipment needed to enable computer delivery. Delivery and scoring systems are available and need not be developed. There may be other ways to be able to reduce the cost of scoring. For example, scoring costs may be somewhat limited if the scoring effort can be viewed as professional development for teachers. However, doing this must be a decision reached by teachers and their employers, and we cannot presume as a precondition of design such negotiation will occur, or if it occurs, be successful.

- Aggressive use of automated scoring will further limit ongoing operational costs.

In summary, it is likely that the system we propose is likely to have higher ongoing scoring costs than current systems. These will be at least partly offset by efficiencies in development and psychometrics to be obtained from pooled development efforts. Over time, efficiencies from technology delivery (no printing, shipping, and scanning) will also help, although in the short term, acquisition of technology will require certain start-up costs.

Limitations and Need for Research and Development

Any assessment system is based on choices of which priorities to give precedence. This system is no different: It places a high value on the comparability (both over time and in different places) of data obtained from formal testing events, on measuring a broad range of skills including those that involve new technologies, and on using technology aggressively in an integrated assessment solution. It similarly places a high value on being able to measure growth.

There are, of course goals this system is not optimized to realize. While the system supports rapid score turnaround, it is not the fastest possible system. Using only machine-scored items and pre-equating would yield faster scores, but we do not believe we can reasonably commit to such instrumentation without hindering the system's ability to meet other more central goals.

Furthermore, data comparability over time and across different users of the system is a key system goal. Therefore this system puts some real limits on user customization of common core assessment components. In fact, there is substantial allowance for customization, but customization that is limited to elements not supposed to contribute directly to common core measurement of comparable data (see the *End-of-Year Tests at Grades 3–8* section). In limited cases (for example, possible project-based elements whose scores may be included in the accountability system) we have suggested the possibility of some customization in the summative/accountability system. However, we believe this customization should not be an option in the common core elements of the end-of-year assessments or periodic standardized tests. Even in the case of project-based elements to be included in accountability results, the degree of customization and choice does have implications for the comparability of data that will need to be carefully considered.

Value Versus Burden

We believe that this system adds extraordinary value. Assessment results will be more meaningful and will provide better instructional models. The new system will provide growth data, and will allow for better comparisons across states and over time. Assessments will cover more rigorous content, and will test emerging skills including those that involve the uses of technology.

Of course, none of this improvement comes without some burden. Ongoing systems costs are discussed under point 10 above—we believe they can be controlled. Teacher involvement in scoring can either be seen as a burden or a value; it will be seen as a value if it is viewed as effective professional development. In terms of teaching burden, the real upwards pressure should come from having to adjust to new standards, and not from the assessment. However, if the new standards really represent fewer learning objectives, then the additional short term burden will be mitigated. And the benefit is freedom from the pressure for curricular coverage currently plaguing many teachers.

Perhaps the most interesting point relates to the possible use of periodic standardized tests and project-based elements in addition to an end-of-year test. This clearly increases testing time. The only way to balance this burden is to ensure that the testing events are valuable instructional experiences, and we believe we can ensure that they are indeed so. This, combined with the improved data and more actionable feedback, would seem to make the burden a worthwhile investment.