THE NATIONAL CONFERENCE
ON
NEXT GENERATION ASSESSMENT SYSTEMS

An American Examination System

Lauren B. Resnick and Larry Berger

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.

An American Examination System[1]

Lauren B. Resnick

University of Pittsburgh

Larry Berger

Wireless Generation

Purpose and General Design

Advances in education research, statistics, technology, design, and policy have prepared the American education system for breakthroughs in standards, curriculum, assessment, and the relationships among them. This time, in ways that were not previously possible, we have the knowledge and the tools to keep ambitious teaching and learning at the center of the system, even as we sustain our commitment to accountability and equity. We also have new tools that will enable us to use the data that emerge from the system to continuously improve it and tune it to the needs of each child.

We propose a new vision for an aligned system of standards, assessment, and curriculum. Our vision comprises three main areas of innovation working in concert:

1. *Distributed Accountability Exams (DAEs),* given periodically during the school year rather than once near the end, with items designed to be not only valid and reliable but also *educative*. These would be administered after a unit of curriculum that is expressly designed to prepare students for the exam.

2. *A formative assessment system* that is designed from the ground up around how teachers will make use of assessment in the classroom. Because all of the assessment data, both summative and formative, feeds a comprehensive learning profile of each student, the technology of *adaptive mass personalization* can be applied to shrink the testing burden and target assessment to each child's current place on relevant learning trajectories.

3. *A technology platform* that makes it easy for schools and teachers to manage the assessment process and that puts at teachers' fingertips the insights and actions that should follow from assessment data. The platform facilitates scoring many item types instantly and, when human scoring is required, streamlines the workflow to provide feedback as swiftly as possible.

[1] The authors wish to acknowledge Brian Junker for his extensive contributions to this work. Junker's insights were essential to the design of the proposed system and he deserves the entirety of the credit for providing an analytical and statistical framework for explaining how a more educative assessment system could also be a more psychometrically rigorous one.

The Problem

Over the past two decades, our country has been trying to build a standards-based accountability system as a foundation for a more equitable and higher-achieving education system. In practice, however, we have created a *test-based* accountability system that does not reflect the standards we aimed for at the beginning of the 1990s, much less today's *fewer, clearer, higher* Common Core Standards.

Several studies, using several different methodologies, have shown that the state tests do not measure the higher-order thinking, problem-solving, and creativity needed for students to succeed in the 21st century. These tests, with only a few exceptions, systematically over-represent basic skills and knowledge and omit the complex knowledge and reasoning we are seeking for college and career readiness.[2]

The misrepresentation of standards by most current accountability tests has had negative effects on teaching and learning, especially for poor and minority students. The tests carry consequences, and many educators serving poor students aim to raise test scores in the most direct—in some cases, the only—way they know: They provide practice on exercises that substantially match the format and content of their state's end-of-year accountability tests. These exercises often depart substantially from best instructional practice. Some studies have documented a systematic *decline* from fall to spring in the quality of instruction. In reading, for example, the complexity of texts that students engage with is *lower*—in the same classrooms with the same children—in March than in October. And there is *less* discussion of text and word meaning as teachers direct children through workbook exercises that mimic state test items (Anagnostopoulos, 2003; Koretz & Hamilton, 2006; McNeill, 2002). Principals and district administrators encourage this practice. They introduce interim assessments that largely mirror the end-of-year tests rather than model the kinds of performance intended by the standards. They do this because the tests count, and they are afraid that without practice, students will not do well enough to meet adequate yearly progress (AYP) requirements.

Calls now abound for even more frequent testing and for focusing teachers' attention early and often on which items their students are having difficulty answering on the interim assessments. But unless the process is guided by a fundamental understanding of *what kind of teaching* helps children acquire robust competence, we should not be surprised when the most frequent response to weak early test scores is to practice the test. Though no one intended to do so, we have created a testing bind that, as it tightens, drives attention away from the intended standards. The effects are greatest in the poorest schools. The nation's current approach to raising achievement and increasing equity in the education system is having an effect opposite from the intended one. It is trapping poor children in a basic-skills teaching program that gives them little chance to acquire the deeper knowledge and abilities we seek for everyone. And it may be lowering the learning opportunities even for many more privileged children as schools turn their energies to the test-based basic skills program.

[2] The problem cannot be fixed by changing cut scores so that states no longer deem as being *proficient* test performances that barely meet NAEP standards for *basic* levels of achievement. The tests are fundamentally misaligned with 21st century expectations. For an analysis, see Resnick, Stein, and Coon (2008).

Many educators, parents, and citizens have responded by clamoring for an end to test-based accountability. Witness the one-sided reaction to a recent editorial in the *New York Times* written by Susan Engel (2010) calling for less testing and more play (and by implication, less direct instruction) for children. A stream of supportive commentary by readers ensued—but none expressing concern about how to educate poor children, minority children, or English language learners to college-ready levels of achievement. Most of the children of responders to Engel's article would not be harmed—might even benefit—by a weakened accountability system. But the others—the ones no one spoke for in the New York Times exchange—could lose even the slender chances we now offer them.

A Solution

Testing and accountability should remain at the heart of national education policy. Equity and national prosperity depend on a system that will stretch educators, the education system, and communities to work toward high achievement and that will enable clear accountability when achievement goals are missed. But there should be new forms of assessment, functioning in new ways within the education system, to meet the needs. As early as 1992, scholars showed how in many countries of the world, tightly linked examination and curriculum systems kept aspirations high, guided teachers in their work, and—sometimes—created pathways for young people who did not come from privileged families (Resnick & Resnick, 1992). The secret lay in charging teachers to *prepare their students for exams* and making sure that the exams were worth studying for. For the system to work, teachers and students needed to have a rough idea of the *kinds* of questions that would be posed on the exams—although not the specific questions that would appear. The systems also required trust that exam grades would be fair—that is, students would likely receive the same grade no matter who scored their written work (written essays predominated over short answer and multiple-choice items because the countries valued the kinds of thinking that were displayed in such essays). Systems for checking on grade fairness (and allowing challenges in a few cases) varied among the countries studied, but all found ways of maintaining public trust in the system.

In this paper, we outline an American Examination System, one that reflects key aspects of the substantive, cognitively demanding European systems, while maintaining standards of psychometric rigor necessary to support America's accountability, comparability, and equity agendas.

The American Examination System we have in mind:

- *models the kinds of instruction that are valued* so that preparing students for assessment works for —rather than against—high-cognitive demand instruction;
- *situates exams within the stream of ongoing instruction* so that assessments support teaching rather than distract from it;
- *ensures content and instructional validity of all assessments* so that the alignment problems that have plagued state testing systems can be resolved;
- *provides reliable and valid accountability measures* for student, school, and educator performance;
- *includes diagnostic tools for instruction* to meet individual student needs;

- *leverages advanced data collection and computational resources to mass personalize the formative assessments,* improving their precision and usefulness.

The American Examination System we outline would be *educative* for those who use it. It would not just tell us how well students, teachers, and schools are performing, but also teach *teachers* how to teach, teach s*tudents* how to learn, and teach education *organizations* how to develop teaching expertise. It would meet this educative goal through a system that combines *distributed accountability exams* linked to specific topics for instruction with *diagnostic, formative assessments* designed for teacher use *during* instruction.

An online platform will make it possible to deploy and manage all of these elements at scale in a cost-effective way while minimizing additional burdens for teachers, students, and administrators. This online platform would be much more than a system for administering, scoring, and reporting on assessments. It can surround the *what* of assessment outcomes with useful representations of *so what?* (professional development) and *now what?* (more targeted instructional resources) so that everyone focuses on the consequential and instructional validity of assessment results and not just the accountability pressure.

Distributed Accountability Exams (DAEs)

Accountability data in this system would be derived from exams that are administered at intervals throughout the school year, occurring after students have completed a unit of study on particular content and skills as identified in the Common Core Standards and state standards. Accountability data would be reported on the basis of individual student, subgroup, class, school, and district, as well as across classes, schools, and districts. The types of tasks on the exams would be largely familiar to students, who would have worked on similar tasks in the course of instruction. But neither teachers nor students would know prior to the DAE exactly what questions would appear. Based on what is required from the new Common Core Standards, we expect three to five DAEs per year in mathematics and literacy at each grade, with each exam assessing material covered through 3–7 weeks of instruction, but the specifics of number and timing would need to be worked out with states.

The DAEs would model the kind of high-cognitive demand performances intended by the Common Core Standards and rigorous state standards, as well as test basic procedural skills. In literacy, they would include extended written work and other open-ended expressions of student reasoning and thinking; in mathematics, they would include drawings, graphs, mathematical expressions, and explanations. They would assess basic knowledge both within these constructed performances and, where appropriate, in clusters of multiple-choice items. In addition to modeling high-cognitive demand instruction, the DAEs would reflect what should be taught (specific topics determined by state and Common Core Standards).

The Common Core Standards provide a foundation for a criterion-referenced examination system that is closely tied to instruction yet meets crucial criteria of technical quality of assessment. The core grade-

level standards are organized as a set of trajectories or sequences of learning goals.[3] They are specified at a granular size that can be used to organize meaningful units of instruction and correspondingly meaningful assessments.

Tasks or items for the DAEs would be pre-tested and calibrated using standard classical and multi-dimensional item response theory (IRT) frameworks. In addition, each DAE would undergo a rigorous process of establishing content validity and instructional validity—processes that test theory often calls for but are not part of standard procedure in most instances of education test design. As the project matures, tasks would be collected into item banks for use in future construction of DAEs. Information on student performance data, instructional targets, and the forms of instruction that result reliably in student learning would be shared with stakeholders including parents and students, teachers, schools, testing administrators, and those responsible for preparing and selecting teachers.

Ideally, every student would take each DAE when he or she is ready and not before. So the long-term goal should be to have sufficient alternate exams that students have more than one chance to take an exam (as they do for New York State Regents).

At the outset, a more limited set of equivalent exams—two versions of each DAE—would be developed. The two versions, one administered before instruction and one afterwards, would be used by the assessment developers to establish *instructional validity* of the exams. Availability of multiple forms of the DAEs would allow states and districts to use the content-based exams to plot *student growth*, along with *teacher and school effectiveness*. In addition, pre-instruction results could be used by teachers as part of the formative data they use to plan an instructional unit.

Figure 1, a diagram of how the DAEs might progress through the school year, shows how DAEs interact with formative assessments (described in subsequent sections) that are also integrated into the system.

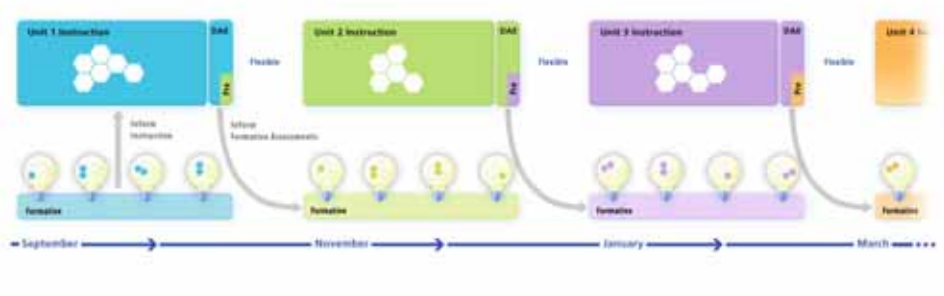


Figure 1. Example of How Distributed Accountability Exams (DAEs) Might Progress in a School Year.

[3] Some of the learning sequences in the standards are based on research conducted by multiple scholars over three decades. Others are based on well-honed intuitive judgments by expert scholars and practitioners. All will require further validation-in-use over the coming years. What is new and important in the current core standards effort is that the standards are organized into multi-dimensional sequences of learning that can inform both assessment and instruction.

The sixth and seventh grade Common Core Standards for mathematics specify five content areas:

- Ratios and Proportional Relationships
- The Number System
- Expressions and Equations
- Geometry (in sixth grade, Properties of Area, Surface Area, and Volume are explicitly named)
- Statistics and Probability

The Ratios and Proportional Relationships section for sixth grade mathematics (see Appendix A) includes two parallel sets of standards, one for Mathematical Understanding and one for Mathematical Skill. In addition, there is a set of standards for Mathematical Practice that the standards writers intend to apply at all grade levels, although it is understood that the student performances representing good mathematical practice will look substantially different at different age/grade levels. Our DAEs would provide a valid and reliable picture of how students are progressing on the Mathematical Practice standards as well as on the specific content standards.

Figure 2 displays the sixth grade standards in a visualization we call the *honeycomb* that specifies our hypotheses about the interdependencies among them. The honeycomb, which we describe more fully below, serves as a visual representation (interactive map) of the instructional and assessment space that needs to be traversed in all grades, including the sixth, and also as a frame for assembling data on student performance in a manner that will support inferences about the progress of individual students, classes of students, schools, and school districts.

Taken together, the Mathematical Understanding, Mathematical Skill, and Mathematical Practice standards inform and constrain the assessments that would be built for the Distributed Accountability Exams. Assume, for purposes of developing an example, that the sixth and seventh grade mathematics teaching will be divided into five units of instruction, one unit for each of the five content areas. One would thus need five content-specific exams in mathematics each year for sixth and seventh grades. The exams (like the instructional units they reference) might not be of equal length, because some of the standards cover more material than others. But we envision exams of 40 to 75 minutes in length, each geared to a teaching unit of 3 to 7 weeks.

An example of an exam covering the sixth grade unit on Ratios and Proportional Relationships is included in Appendix A.

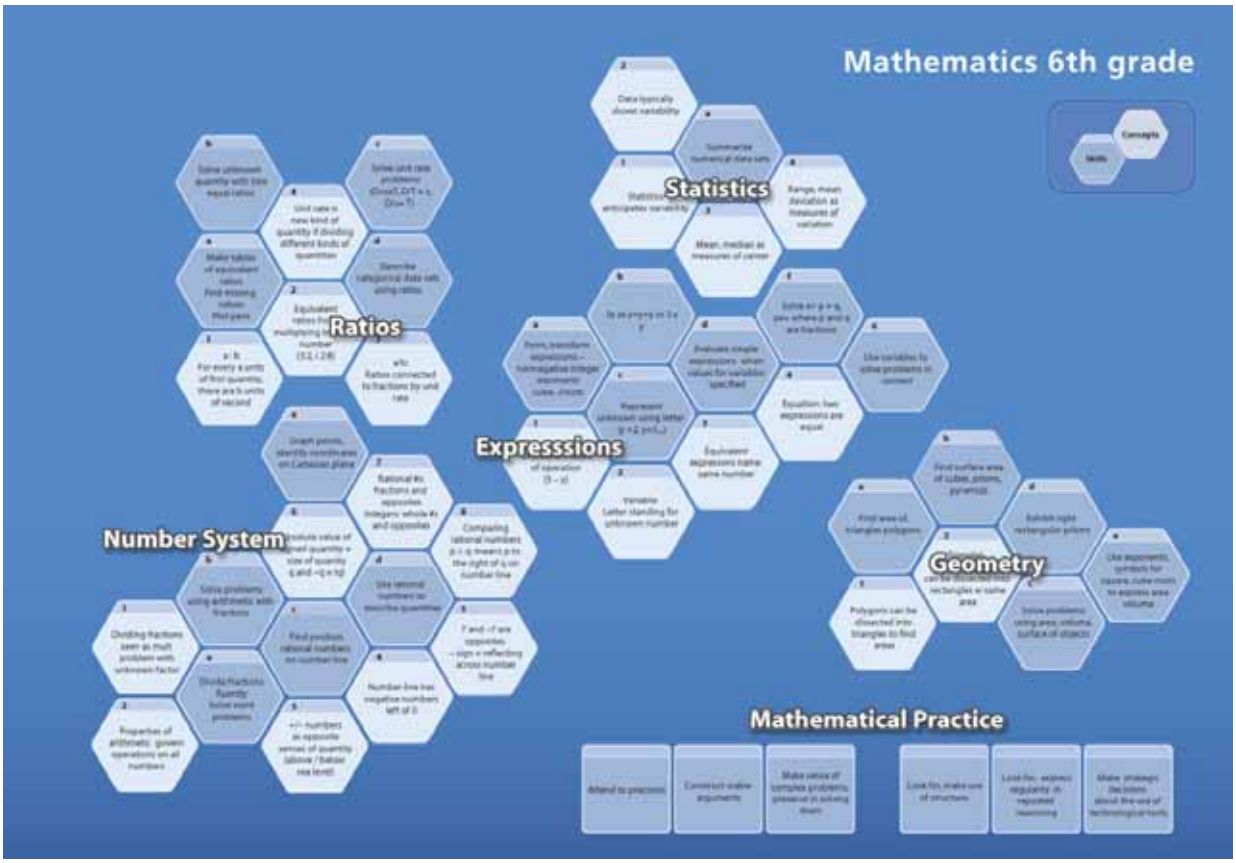


Figure 2. Visualization of Sixth Grade Standards as a Honeycomb.

An English Language Arts Example

The English language arts/literacy standards can similarly be used to specify sequences of instructional units and assessments. The core standards are organized in grade bands rather than grade by grade. As in mathematics, skills and understandings are expected to develop over multiple years. In addition, guidelines exist for choosing texts that use modern quantitative methods to characterize the cognitive and linguistic complexity of writing in several different genres.

Using all of these resources of the Common Core Standards, we have sketched distributed exams for English language arts; one example of an exam is in Appendix B.

Validity and Reliability in Distributed Examinations

The DAEs would be built to strong criteria of *content* and *instructional* validity. Each exam would provide a *reliable* estimate of student knowledge on the content of an instructional unit that is explicitly targeted to a standard, or set of standards, in the Core. The collection of exam scores for a year (e.g., five mathematics exams in each of Grades 6 and 7) would provide a valid estimate of the extent to which a student (class, school) has mastered the content specified by the standards for that year.

Content Validity

The exams would match closely, in both content and form, the content that is expected to be taught in each of the instructional units. New instructional units, explicitly linked to the Core standards, would be created to anchor the content validity of the units. Teams of independent content and instructional experts would review the model instructional units to ensure they match with the standards and are of high-instructional quality. The same teams would judge the alignment of exams to the model instructional units. This process would largely overcome the problem of weak alignment to standards that now troubles many state assessments. (States would not, however, be required to use the model instructional units in their actual classrooms.)

Instructional Validity

Assessments are considered instructionally valid when student performance improves after quality instruction on the content of the assessment. Although instructional validity is part of the gold standard for educational testing, it is almost never established in current assessment practice. We can do better. We will apply strategies of in vivo (live classroom) research developed by the Pittsburgh Science of Learning Center. These scientific (experiment-based) research strategies can be used to establish whether each particular DAE, in fact, responds to good teaching. States and school districts using the DAE system would be able to validate DAEs against best practice instruction developed by their most effective teachers.

Reliability

DAEs would contain a mix of short constructed-response items and more extended written responses, along with sets of multiple-choice items as appropriate to the standard being examined. Short and long constructed response components would require human scoring. Research has established that when constructed-response tasks are well-targeted, scoring rubrics are specific ,and graders are trained, a high level of inter-rater reliability can be attained (Mariano & Junker, 2007; Patz, Junker, Johnson, & Mariano, 2002; Rayn & Shepard, 2008).

Student responses on constructed-response items could be graded locally (within the same school but not by the student's own teacher) or by geographically and socially remote scorers (including teachers elsewhere in the district or state). These grades could be validated using one of a number of methods that have been used in European countries (e.g., cross-school or cross-state grading exercises; re-grading of a sample of student papers at the state level). Teacher participation in grading exams and the related validation exercises (some of which could be face-to-face) creates a good process for professional learning, one that many countries use.

DAEs open the possibility for increased use of constructed responses because they are distributed over the course of the year, yielding several times more opportunity to collect data than current end-of-year

tests. This also brings benefits in terms of increased test reliability.[4] Yet to obtain these more reliable results, students would not have to sit for a 5-hour exam or even take an end-of-year exam, depending on how a particular state system is designed. They just would have to take unit exams as they normally would in the course of teaching, but now with the unit exam contributing to an overall accountability score.

In addition, we propose to use earlier assessment data to help produce more precise proficiency estimates for each DAE. This approach, similar to what is used in some online tutoring systems and adaptive testing systems, could make it possible to shorten many of the assessments with no loss in measurement precision (see Appendix C).

Distributed content and instructionally validated exams are a next logical step in ending the testing bind and developing an assessment system that will detect and reward high-quality, effective teaching. Instead of supporting the use of practice materials that mimic the old end-of-year tests, states can provide high-quality instructional tools that help teachers prepare students for DAE examinations.[5] There will be no need for the current crop of interim tests that simply mirror the end of year test, since DAEs and related formative assessments will occur throughout the school year at times that make instructional sense. With this system, we gain ability to measure a set of higher-order skills that are not otherwise easily tested, including skills essential to college and career ready performance in reading, writing, and mathematics, without adding enormous burden of testing.

Educative Formative Assessments

The American Examination System will foster a rich environment of formative assessments that are educative in ways that directly resemble the summative system, but with more direct application to daily and weekly instruction.

- They would be aligned with the learning trajectories derived from the Common Core Standards, and thus aligned with *what* teachers need to teach.
- They would model approaches to *how* to teach, and would, at the request of educators, provide teachers structured opportunities for gaining experience in using those teaching methods.
- Teachers would make these assessments *part of their instructional routine*, rather than an addition to it. Data-entry/record-keeping burdens will be minimal, and teachers will have easy and quick access to student- and class-level reporting as well as tools to

[4] For instance, if the reliability of each single DAE hour-long exam were 0.7, the reliability of five DAEs taken together would be 5*(0.7)/(1 + 4*0.7) = 0.92. If, instead, half of each DAE's testing time were used for a pretest on the next instructional unit or simply for calibrating future test items, the improvement would be 2.5*(0.7)/(1 + 1.5*0.7) = 0.85—still a high rate of reliability.
[5] For a description of approaches to providing this kind of instructional guidance in forms that do not suppress teacher ingenuity and judgment, see Resnick (in press) and McConachie and Petrosky (2009).

understand the instructional significance of that data. By tracking fidelity in the use of these diagnostic tools, the system will help teachers to use them appropriately.

Formative assessment tasks that cannot be machine-scored will be accompanied by simple rubrics for quickly analyzing the student work. Teachers will be able to use digital devices to record these analyses. Through those devices, the teachers will also be provided with samples of answers that correspond to each level on the rubric, to help them calibrate their own analyses. As a form of professional development and to improve the reliability of analyses, teachers could also upload the student work into the system, along with their analyses, to get feedback from other teachers or subject matter experts.

A Mathematics Example

Educative formative assessments in mathematics will be designed recognizing that cognitively demanding tasks can typically be solved in many different ways. From a diagnostic perspective, it can be as important to know how a student is attempting to solve a problem as it is to know his or her answer. The problem-solving technique is, in many cases, part of what is specified in the standards. The sequence of how these techniques are used over time will often indicate a student's progress in understanding concepts and moving along a learning trajectory. So the Educative Formative Assessments would include items that capture this information and empower teachers to learn to recognize the different approaches that students take and their significance for differentiated instruction.

An example of this approach is the Ongoing Assessment Project (OGAP)[6] in mathematics, a framework and system for analyzing mathematical reasoning of elementary and middle school students as they solve problems. Teachers analyze written student work looking for evidence of mathematical reasoning and increasing levels of sophistication as students progress along learning trajectories. The diagnostic and instructional utility of the items are enhanced by examining the thinking and strategies that went into solving them. Fewer items can be used to produce far richer results because the underlying thinking is surfaced and made apparent to the teacher. Figure 3 illustrates how teacher-facing software enables quick analysis and recording of meaningful attributes of student work—correctness of response, sophistication of the reasoning (along a trajectory from additive-transitional-multiplicative strategies), and any errors or misconceptions that emerge; these tools and interfaces can also support remote analysis when student work is digitized and routed.

In general, it will be essential to ensure that formative assessment results are not included in accountability reporting to eliminate the incentives for misuse. We envision that the student-, class-, and school-level results would be available to teachers, coaches, and perhaps principals (to inform professional development as well as instruction), but not to district/state administrators.

[6] OGAP was developed as a part of the Vermont Mathematics Partnership funded by the U.S. Department of Education (Award number S366A020002) and the National Science Foundation (Award number HER-0227057).

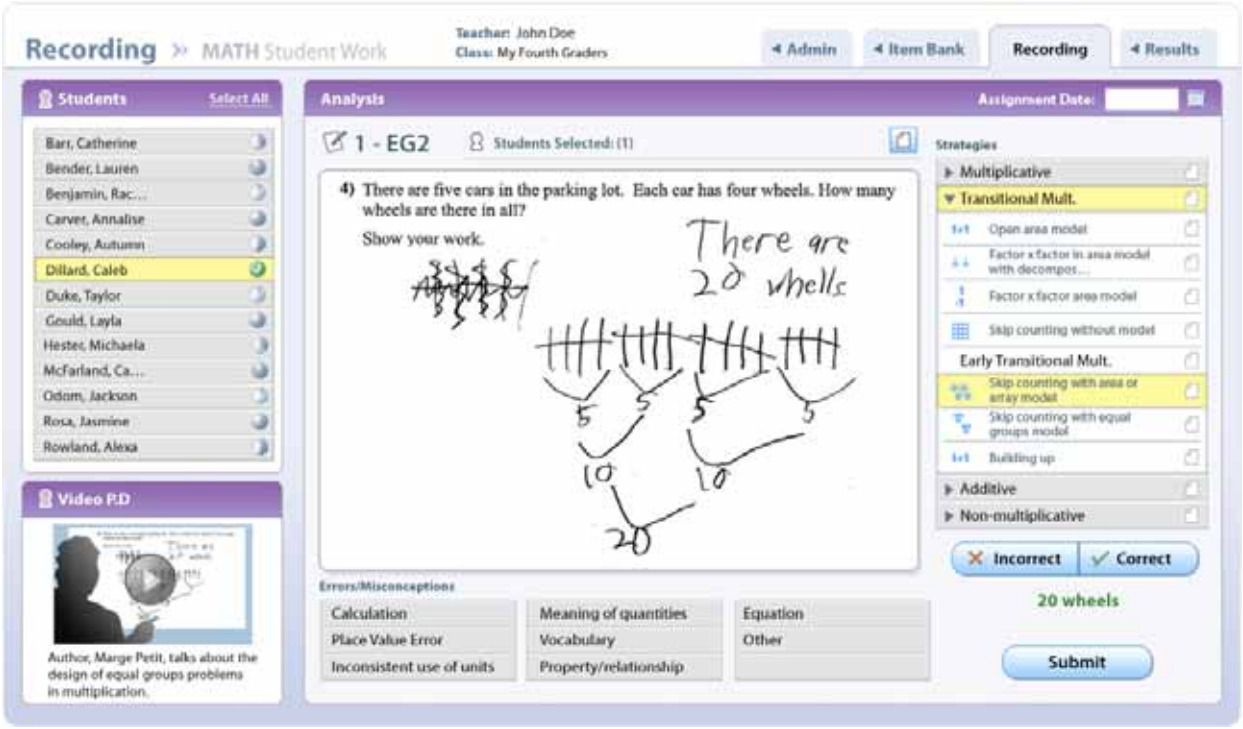


Figure 3. Teacher-Facing Software.

However, metrics of fidelity in implementing the formative assessments (and their associated instructional recommendations) could be used as part of teacher/school performance management/accountability. For instance, are teachers doing progress monitoring with the frequency appropriate for each student, given the longitudinal data about that student? Principals and district/state officials should have access to this type of information in real time, so they can spot where there may be weak instructional capacity and provide timely interventions (including targeted professional development). They will want to spot if teachers are using the formative system the way in which, and as often as, it should be used. (DC public schools is an example of a school system that is already using these types of formative assessment metrics as part of its SchoolStat approach to continuous, district-wide performance management.)

In addition, the American Examination System platform would provide to researchers longitudinal data including formative assessment data, organized by student/teacher/school/subgroup.[7] In particular, this data would be used as part of the research to support continuous improvement of the system: to fine-tune the learning trajectories, measures of proficiency for each standard, and algorithms for mass customization of assessments.

[7] All data would be anonymous to protect privacy (and prevent the formative data from being used for accountability). Researchers will be able to see that Student A had Teacher X in School Y and see data available for A, X, and Y, but not the identity of those individuals and institutions.

We expect that formative assessment fidelity data will be especially useful to researchers. Many instructional innovations, when tested under real-classroom circumstances, fail to show impact: researchers wonder whether the lack of results was because of poor design or simply because the teachers did not implement it correctly. In the field of learning research, scholars are pointing to the need for researchers to distinguish between poor design and poor implementation. They make the comparison with pharmaceutical trials, where a prerequisite for testing medical efficacy is knowing which of the trial patients took the correct dosage (Rowan, Correnti, Miller, & Camburn, 2009).

A New Paradigm for Educational Measurement: Adaptive Mass Personalization

We believe that an advanced model of educational measurement can be built on a foundation of gathering an order-of-magnitude more data—both informal and formal—about each student in the course of the year so that each test merely enhances the resolution of a picture that is substantially complete before each test begins. Moreover, by applying the tools of mass personalization already so prevalent in Internet-based commerce and social networking, we will eventually be able to personalize each assessment at the individual level so that the enhanced resolution it provides is targeted to an individual student's current learning level as well as to appropriate standards of reliability and validity. That is, the system can keep asking questions until it knows enough to be instructionally helpful to the student and the teacher and until it knows enough to support relevant policy and accountability decisions.

Standardization Versus Personalization

Standardization was the engine of the factory model that drove the economy of the 19th and 20th centuries (Resnick & Resnick, 1977, 1980). Now the powerful drivers of the economy are *personalization* and *customization*—often applied in direct contradiction to a previously valued standardized offering. Amazon.com, for example, learns what you like to read and offers an increasingly personalized bookstore just for you that becomes more precise over time. The video rental chain Netflix has now hosted several international competitions for improving their personalization engine.

The statistical engines underlying personalization on the World Wide Web are distinct from those underlying standardized testing, but they are now entirely robust and proven—indeed they are tested and refined on a daily basis in large-scale commerce, large-scale medical research, and financial market predictions.

It is time to bring these ideas to education in ways that will dramatically improve the precision with which our formerly standardized tests fulfilled their standard purposes, while simultaneously expanding their usefulness to inform daily instruction, to diagnose individual patterns in student learning, and to surround students with supports that are personalized to their needs.

Because the American Examination System aims to administer all types of assessments for a very large number of students over a period of multiple years, across multiple states, and can take account of various other education data, it should be able to serve as an engine for mass personalization of these

assessments. Attributes that could be the basis of personalization include past student performance on assessments, teacher and school characteristics, aggregated assessment performance of students in a school, previous effectiveness of teacher, which curriculum was used, and which assessments have been used. This technology is scalable—computing power is such that there is no practical limit on the amount of education data that could be included—so that as more states and more types of data are included, the more precise (and useful) the customization becomes.

This initial goal for mass personalization would be to apply it to formative assessment. There are, already in use, many modalities of formative assessment (diagnostic, progress monitoring, screening), each including a mix of assessment types (multiple choice, constructed response, observation). Some of these are best delivered as part of group activities and some one-on-one between a single student and teacher. Many teachers/districts use a blend of these formative assessments, which makes sense given the diverse needs of particular students at different moments of their academic development; but many other teachers—who are not themselves experts in formative assessment methodologies—struggle to decide how best to integrate all of these choices into their teaching routines for their particular students.

So, in addition to providing new educative formative assessments, the American Examination System would mass customize a much wider range of formative assessments at the student and class level. This is adaptive assessment at the level above individual items—it figures out *which* formative assessment to give and when—enabling teachers to get just the right next piece of information they need about their students, without wasting a lot of classroom or other school time. With this platform, teachers will be blending modes of assessment in individualized ways—varying what data they collect and how— based on what is known so far about each student. To support this, the system will host a bank of formative assessment materials, to cover the full range of diagnostic options a state or school district wishes to use, from open-source or commercial sources.

The mass personalization process can also add to the reliability and efficiency of DAEs. Appendix C shows how a standard statistical model can use data from previous DAEs to make the next DAE more efficient, as long as the student is behaving consistently from one unit to the next. If the student seems to be performing unusually well (or poorly), then the model can detect this and suggest a customization of the DAE to further explore what the student knows and can do.

The Assessment Platform

The assessment Platform manages both parts of the system—the DAEs and the educative formative assessments—to enable assessment delivery, scoring, reporting, and analysis. Based on widespread classroom experience with existing products and on current designs[8] (some of which have been funded by the Gates Foundation), it will be able handle all of these elements at scale in a cost-effective way, while minimizing additional burdens for teachers, students, and administrators.

[8] The authors wish to acknowledge the support of the Gates Foundation in conceptualizing a next-generation assessment platform and for more generally advancing the field of aligned units of curriculum and assessment.

Honeycomb

The American Examination System will provide a honeycomb—an interactive map of learning trajectories and our hypotheses about the dependencies among them. The honeycomb offers a visual representation of the instructional and assessment space that needs to be traversed in each grade as well as across grades, all the way from pre-K through Grade 12. It provides a frame for assembling data on student performance in a manner that will support inferences about the progress of individual students, classes of students, schools, and school districts. It will also support research to validate/refine the hypotheses about dependencies among the skills (within and across trajectories) in the Common Core Standards and similar state standards—for instance, identifying what level of which specific literacy skills are needed to achieve mastery of which mathematics skills.

The American Examination System would give educators summative and formative assessments for each skill step along each learning trajectory, starting with mathematics and literacy for Grades 3–10. Other assessment data—for instance, existing formative assessments for pre-K through Grade 3 students or high school exams—can also be mapped onto the learning trajectories. All of this data can be included in the honeycomb so that teachers, parents, and the students themselves can track individual student progress (and extent to which students are on track) toward college and career readiness.

The honeycomb builds on one of the intrinsic advantages of the American Examination System, which is that it offers a highly coherent and integrated package of summative and formative assessments. In particular, the system's rapid scoring workflow and reporting interface would enable educators to use the DAE results for diagnostic purposes at the individual student and class level. For example, where students have written an essay, teachers would be able to see whether students can write the sort of complex sentences and can make arguments out of ideas that are appropriate for the grade's learning trajectory. The pre-tests for each exam would be especially useful in this regard because the pre-tests assess the topics and standards that teacher is about to teach.

Each hexagon of the honeycomb could also link to instructional resources (including video exemplars and social networking/collaboration). See Figures 4 and 5.

This tool can be adapted for use in any state whose standards include learning trajectories comparable to those in the Common Core Standards. We envision that there would be two measures of proficiency indicated for each skill/hexagon: the first based on formative (no-stakes) data and the second based on summative (high-stakes) data.

Putting Power and Choice in the Hands of Teachers

The platform will include an assignment builder, so that educators can select formative assessment items as tasks for use by the students in the classroom or as homework. This allows the teachers to focus student work on the particular concepts and skills that they need to develop. So, for instance, a teacher could drill down from a specific honeycomb hexagon (Common Core Standard) to build an assignment for a subset of her students. See Figures 6 and 7.

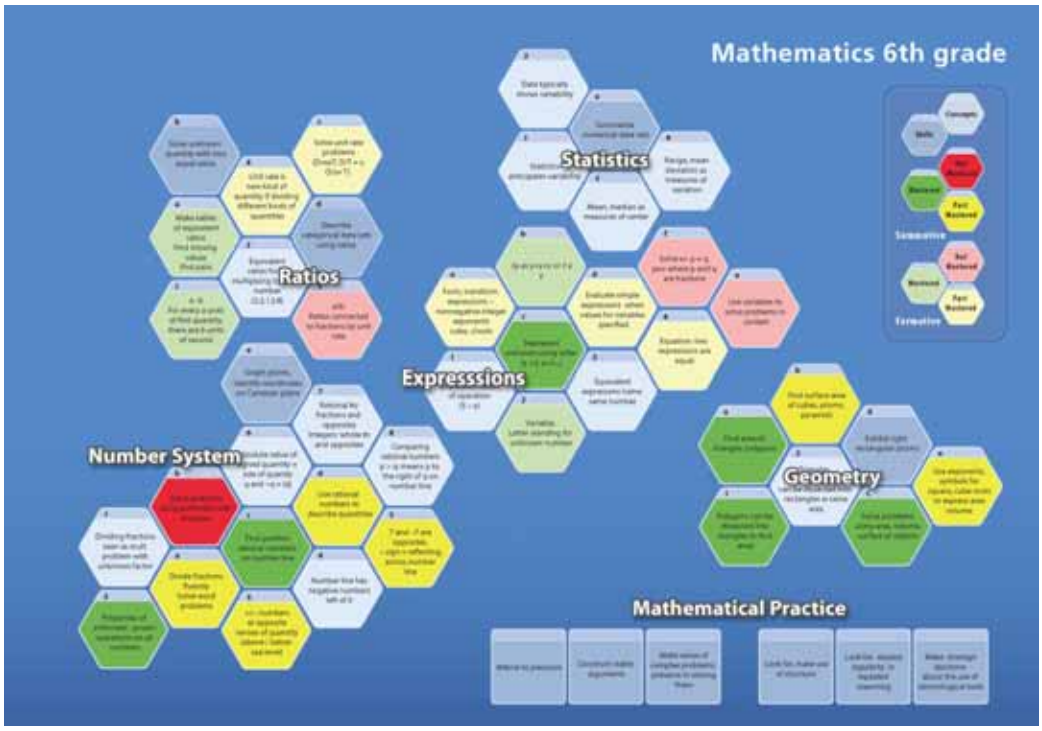


Figure 4. Honeycomb for Mathematics Sixth Grade.

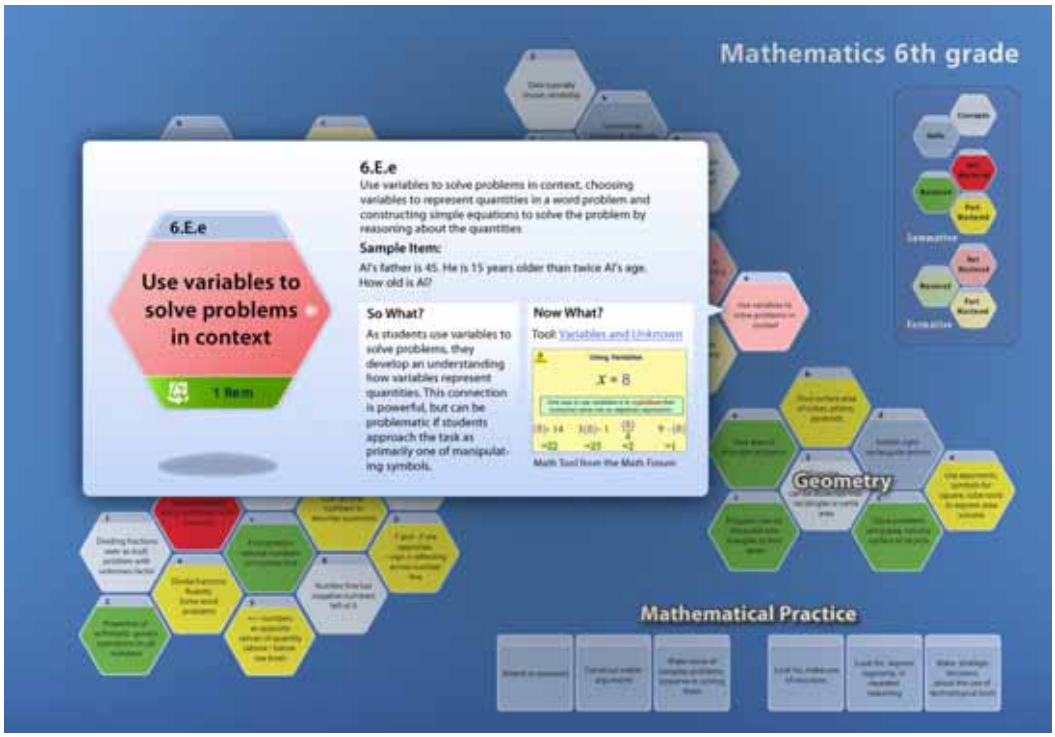


Figure 5. Each Hexagon of the Honeycomb Could Also Link to Instructional Resources.

Figure 6. Assignment builder.

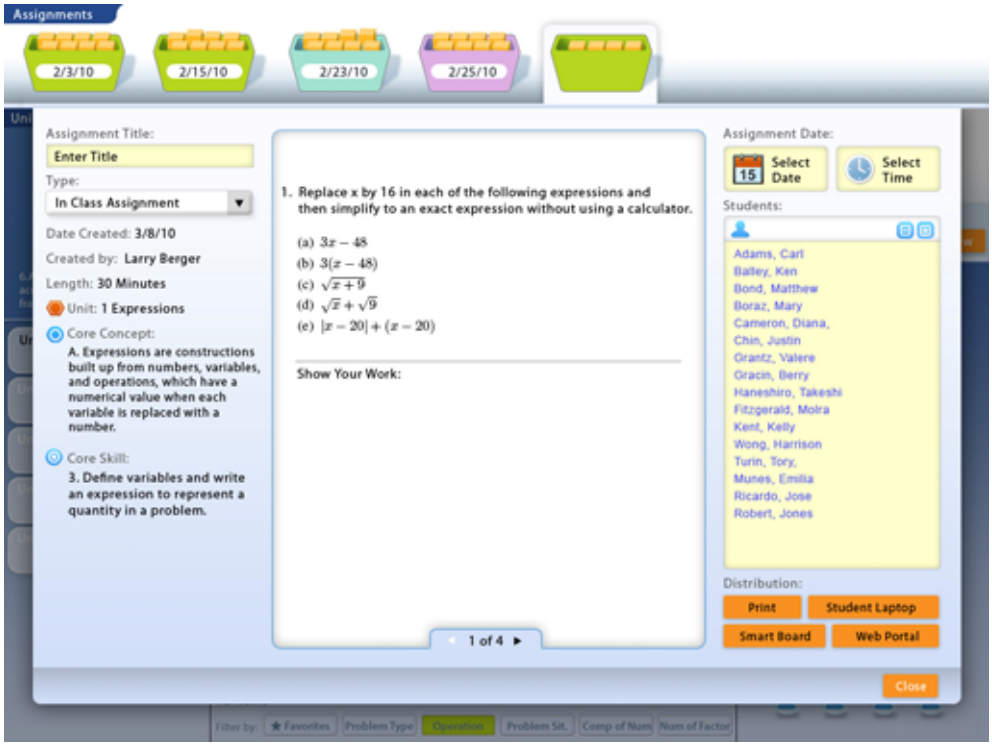


Figure 7. Individual assignment.

Other Platform Tools

In addition to providing the honeycomb and tools to support mass customization, the American Examination System platform will:

- Enable students to take the assessments online or on paper;
- Enable teachers/schools to scan and upload paper-based assessments and other student work;
- Manage remote scoring workflow and provide scoring interface for remote raters;
- Provide teachers with a scoring interface (could include ability to markup student work and record notes) and a reporting (gradebook) interface;
- Provide dashboard tools for tracking and analyzing the progress of particular students and groups and students;
- Provide principals and district/state administrators with a reporting interface that includes aggregate analysis (including cross-class, cross-teacher, cross-school, cross-district and cross-state and demographic comparisons, with the longitudinal dimensions—including value-added on end-of-year high-stakes—included in each);
- Allow users to generate custom reports in real time on demand with both teacher and principal/administrator interfaces;
- Allow teachers to share formative assessments with each other and experts to gain instructional advice and create opportunities for professional development; and
- Provide role-based access rights (including to protect student privacy).[9]

Thus, the system will gather and provide ready access to accountability information, and also help teachers and schools to improve learning measured by rigorous standards and good instructional practices. It would cover the full trajectory from Pre-K through Grade 12.

The American Examination System would not assume that all assessments will always be conducted with students sitting at computers. Given current school infrastructure, and given the challenge of showing mathematics work via keyboard, it may be more efficient to continue to rely to some extent on paper-and-pencil inputs to an otherwise digital system. The continued value of these "primitive" recording tools seems especially compelling when one considers that much of the value of the new generation of assessment tasks depends on soliciting open-ended expressions of student reasoning and thinking—and in the case of mathematics this includes drawings, graphs and explanations.

[9] To ensure protection of student privacy rights, the system has the capacity to make digitized student work anonymous before routing it to remote scorers.

So the American Examination System would include a process to enable scanning/digital photographing, uploading and archiving of very large volumes of paper-based student work, including for Distributed Accountability Exams, to enable remote scoring as well as online student portfolios. The scanning/photographing process, which has already been tested in North Carolina classrooms, puts minimal burdens on teachers or other school staff and does not require large per-school investments in hardware or network infrastructure.

For the foreseeable future, assessment of open-ended expressions of student reasoning and thinking will require at least some element of human scoring. Doing this rigorously and reliably, especially in a summative context where there are stakes for teachers and schools as well as for students, requires finding a cost-effective and time-effective workflow for directing the work to remote scorers (including cross-school or cross-state grading/validation exercises; re-grading of a sample of student papers at the state level).

The American Examination System platform enables this workflow. It automates delivery of digitized student work (including paper-and-pencil work) to raters and those validating the ratings. Student identity is kept private (the raters do not know whose work it is). The online interface for remote raters presents them with the student work alongside scoring forms based on the rubric appropriate for that type of work. See Figure 8.

The platform will allow teachers, principals, districts—and potentially parents and the students themselves—to generate custom reports in real-time on demand. These reports would aggregate longitudinal data from different Distributed Accountability Exams and formative assessments to provide a more complete picture of each student, class, and school.

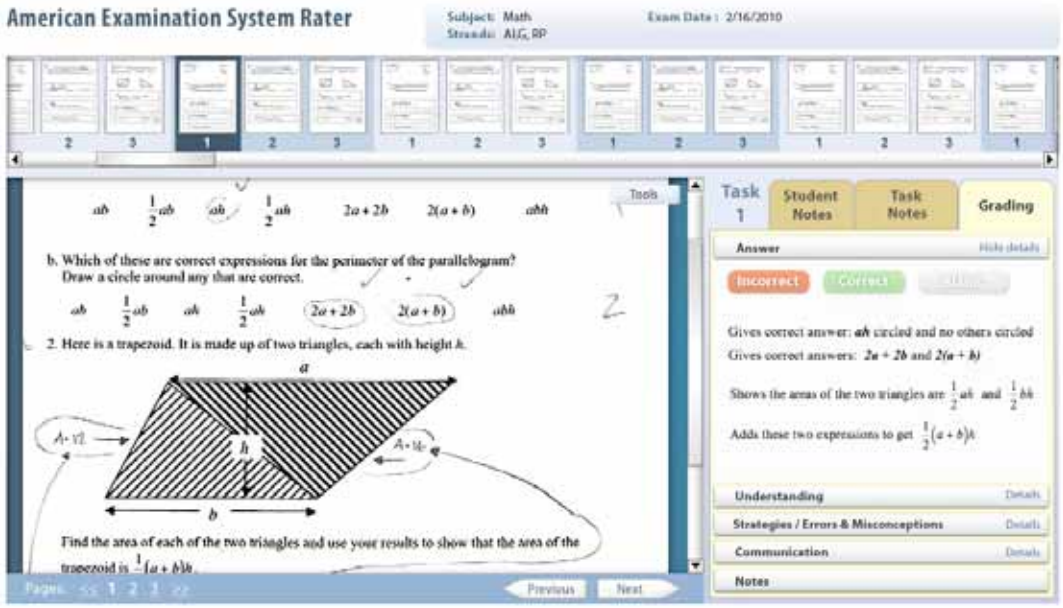


Figure 8. The Online Interface for Remote Raters for the American Examination System.

Development and Costs

Our vision for the American Examination System is ambitious. What makes it realistic is the substantial amount of work that has already been done in developing the content and tools needed to make it work.

For instance, IFL has extensive experience in developing model instructional units such as the ones that will be part of the system and in working with school systems to tailor the units to local needs and preferences (McConachie & Petrosky, 2010). IFL units (and accompanying assessments) are built into the curriculum guidance system of several urban school districts and have been shown to produce high levels of teacher engagement and improved instruction when accompanied by appropriate forms of professional training (David & Greene, 2008; Resnick, in press; Talbert & David, 2008).

Many of the required technologies are already in use in existing assessment and data-management applications or are now being developed through Wireless Generation and through various initiatives of the Bill and Melinda Gates Foundation to create aligned systems of curriculum and formative assessment. Thus, for instance, much of the platform for authoring and administering cognitively demanding assessment items at scale will be available for online use in December 2010.

The system we describe is one that will operate fully about 3 years from the beginning of the process, with mass personalization of summative assessment playing a larger role at the end of that timeframe. Much of the system, including the Distributed Accountability Exams, Educative Formative Assessments, and other aspects of the technology platform, will be operational in 2 years.

Platform

Based on the direct experience of Wireless Generation in building a system of comparable complexity (ARIS, the education information system for the country's largest public school system), we estimate that a secure and scalable version of the initial platform can be available for use in 6 months after work on the project formally begins; additional functionality would be available after 12 months; and a comprehensive system in use at scale in 18 months. Additional development, related to the research and roll-out of the mass personalized aspect of the assessments, would take place within a longer time frame (36 months).

Assessments

The platform could be used to develop and roll out assessments according to the following three phases: establishment of content validity (for both the DAEs and Educative Formative Assessments); establishment of instructional validity (for the DAEs); and then the use of the system for summative accountability purposes. States should be consulted to determine which grades and what subjects to prioritize.

We anticipate that, *after 12 months*, the Distributed Accountability Exams, including the units for model instruction, for use for Grades 3–10, will receive sign-off on content validity by State Departments. The Educative Formative Assessments could begin to be used during this first year after the content is validated.

During months 13–24, or sooner if possible, we will do the experiments to capture instructional validity, beginning as soon as content validity is established.

After this 24–month period, the Distributed Accountability Exams would be used for summative accountability purposes.

Operational Costs

We estimate, for a typical state, the ongoing costs of the system will be about the same as for current NCLB tests. Current expenditures are typically $20–$30 per student, and in some cases higher than $80, to cover reading and mathematics (U.S. Department of Education, 2010).

Administering the Distributed Exams (including the pre-instruction version) will cost more to develop and score than the current high-stakes tests, if for no other reason than their frequency. But the current interim exams, and expenses associated with those (typically $15–$20 or more per student per year), could be eliminated.

Teachers within the school district could score the exams from each other's students, but a significant portion of the ongoing cost would be from validation of samples of teacher scoring.

Apart from provision of tablet/handheld and scanning devices for teachers (off-the-shelf, industry-standard technologies that are coming down in price each year), costs associated with the maintenance of the technology platform would be minimal when considered on a per-student basis.

Key System Characteristics

Rigorous Standards and Good Instructional Practices

The new Common Core Standards provide a foundation for a criterion-referenced examination system that is closely tied to instruction yet meets crucial criteria of technical quality of assessment. The core grade-level standards are organized as a set of trajectories or sequences of learning goals. They are specified at a grain size that can be used to organize meaningful units of instruction and correspondingly meaningful assessments.

The American Examination System includes Distributed Accountability Exams, for use over the course of the school year, which measure the specific higher-order skills that are articulated in the Common Core Standards and state standards, as well as basic knowledge. The Distributed Accountability Exams will include extended written work and other open-ended expressions of student reasoning and thinking; in mathematics, these would include drawings, graphs, and explanations. They will assess basic knowledge both within these constructed performances and, where appropriate, in clusters of multiple choice items. After 24 months, these tests will begin to replace current summative tests for accountability purposes.

The DAEs will reflect what should be taught (specific topics determined by state and Common Core Standards). Distributed Accountability Exams will address each of the skills/topics articulated for each

year of the state and common standards. In the first wave, there will be Distributed Accountability Exams for mathematics and literacy for Grades 3-10 in literacy and mathematics. After that, sample items would be published and invitations extended for participatory authorship of assessment items that relate to the standards that are being tested and the particular item and assessment types.

The DAEs would be built to strong criteria of *content* and *instructional* validity. Each exam would provide a *reliable* estimate of student knowledge on the content of an instructional unit that is explicitly targeted to a standard, or set of standards, in the Core. The collection of exam scores for a year (e.g., five mathematics exams in each of Grades 6 and 7) would provide a valid estimate of the extent to which a student (class, school) has mastered the content specified by the standards for that year.

Content Validity

The exams would match closely in both content and form the content that is expected to be taught in each of the instructional units. New instructional units, explicitly linked to the Core standards would be created to anchor the content validity of the units. Teams of independent content and instructional experts would review the model instructional units to ensure they match with the standards and are of high-instructional quality. The same teams would judge the alignment of exams to the model instructional units. This process would largely overcome the problem of weak alignment to standards that now troubles many state assessments.

Instructional Validity

Assessments are considered instructionally valid when student performance improves after quality instruction on the content of the assessment. Our development process would include tests of instructional validity, similar to the experiment-based ones used by the Pittsburgh Science of Learning Center. These tests would involve panels of teachers with good knowledge of an instructional unit's content as well as demonstrably good pedagogical skills (as judged by an expert panel). These teachers would be put into four groups. Two of the groups would teach the instructional unit that corresponds to the Distributed Accountability Exam. In one of these groups, they would get Pre-test A for their students before the unit is taught and then the students would take Test B. In the second of these groups, the tests are flipped: Test B is the pre-test and Test A is given to students after the unit is taught. In the third and fourth groups, students would not be taught the particular instructional unit at that time, but would still be given the pre-tests and post-tests (one group with A as the pre-test and B as the post-test, the other with B as the pre-test and A as the post-test). Only tests that, through these experiments, systematically register improvements in student performance as a result of corresponding instruction (and demonstrate equivalence through the pre- and post-test swaps) will be included in our Distributed Accountability Exams. Both pre-test and end of unit exams in any given year will be drawn from a bank of tasks that will be developed as part of this validation process.

Both pre-test and end of unit exams in any given year will be drawn from a bank of tasks that will be developed as part of this validation process. Items or tasks for the DAE's will also be pre-tested and calibrated using standard classical and multidimensional IRT frameworks. Availability of multiple forms of the DAEs will allow states and districts to use the content-based exams to plot *student growth*, along

with *teacher and school effectiveness*.[10] In addition, pre-instruction results can be used by teachers as part of the formative data they use to plan an instructional unit.

Reliability

Distributed Accountability Exams would contain a mix of short constructed response items, and more extended written responses, along with sets of multiple-choice items as appropriate to the standard being examined. Short and long constructed response components would require human scoring. Research has established that when constructed response tasks are well-targeted, scoring rubrics are specific and graders are trained, a high level of inter-rater reliability can be attained (Mariano & Junker, 2007; Patz et al., 2002; Rayn & Shepard, 2008).

Student responses on constructed response items could be graded locally (within the same school but not by the students' own teacher), or by geographically and socially remote scorers (including teachers elsewhere in the district or state). These grades could be validated using one of a number of methods that have been used in European countries (e.g., cross-school or cross-state grading exercises; re-grading of a sample of student papers at the state level). Teacher participation in the grading exams and the related validation exercises (some of which could be face-to-face) is a good process for professional learning and is used in most countries. Though the process is more costly in dollars than machine scoring, it is an educative process worth building into our Examination System. Grade validation at scale would be supported by the American Examination System platform, which can enable rapid, cost-effective remote scanning, transmission, grading, validation, and reporting. To ensure protection of student privacy rights, the system has the capacity to anonymize the digitized student work before routing it to the remote scorers and validators, as well as, for limited purposes, automatic essay scoring technologies.

The Distributed Accountability Exams open the possibility for increased use of constructed responses because they are distributed over the course of the year, yielding several times more opportunity to collect data than current end-of-year tests. This also brings benefits in terms of increased test reliability. For instance, if the reliability of each single DAE hour-long exam were 0.7, the reliability of five DAE's taken together would be 5*(0.7)/(1 + 4*0.7) = 0.92. If instead, half of each DAE's testing time were used for a pretest on the next instructional unit or simply for calibrating future test items, the improvement would be 2.5*(0.7)/(1 + 1.5*0.7) = 0.85—still a high rate of reliability.

Yet to obtain these more reliable results, students would not have to sit for a 5-hour exam, or even take an end-of-year exam. They just would have to take unit exams as they normally would in the course of teaching, but now with the unit exam contributing to an overall accountability score. Another advantage is that students would be tested on recently-learned material at all times, so that nuisance effects of delayed recall would not influence measures of how well students were learning what the teachers taught; this would probably increase reliability even more.

[10] If the pre-instruction versions are not long enough to be reliable to estimate instructional effects on individual students, then those effects will be estimated on some aggregate level.

Although the illustration above is useful, in reality it likely will not be possible to string together DAEs into a single unidimensional measurement to which classical reliability calculations apply. Instead we believe the DAEs within a subject will be at least mildly multidimensional; if we consider each DAE within a subject within a year as a measure of one proficiency, five DAEs would be measuring five different but substantively related proficiencies. These proficiencies are likely to be statistically related as well. For example in NAEP, proficiency subscales within the same subject area are typically correlated 0.8 or higher, and seldom lower than 0.5-0.6. We can exploit these correlations by building a multidimensional Bayesian latent variable model to take advantage of proficiency estimates from old DAEs to help produce more precise proficiency estimates for the next DAE, or indeed to shorten the next DAE with no loss in measurement precision.

For example, suppose[11] we wish to estimate a student's proficiency with a margin of error of 0.2 (SEM = 0.1), and each item contributes roughly one unit of Fisher information to proficiency estimation (here we are borrowing an IRT formulation for specificity), then the student would need to answer roughly 100 items. However, if we could already predict the proficiency on this DAE with a margin of error of 0.4 using past DAE performance, we would need only roughly 20 more items to obtain a margin of error of 0.2 on this DAE.

This calculation depends on the student's performance on the new DAE being consistent, in a way that can be made precise using Bayesian modeling, with his/her performance on past DAE's. If the student's responses on the next DAE are inconsistent with his or her older DAE results, we would need to do follow-up testing to get a more precise estimate of the student's proficiency. Thus for students who learn consistently from one unit to the next, we can exploit past performance to help estimate proficiency on the current unit of instruction. However, for example, for the student who performs unusually well (or poorly) on the current unit, we can use the Bayesian machinery to see the inconsistency, and offer another block of items in order to more precisely assess that student's learning. A similar process is used in online tutoring systems and adaptive testing systems, and is an illustration of the kind of useful customization that is discussed below.

Distributed content and instructionally validated exams are a next logical step in ending the testing bind and developing an assessment system that will detect and reward high-quality, effective teaching. Instead of supporting the use of practice materials that mimic the old end-of-year tests, states can provide high-quality instructional tools that help teachers prepare students for DAE examinations.[12] There will be no need for interim tests, since DAEs and related formative assessments will occur throughout the school year at times that make instructional sense. With this system, we gain ability to measure a set of higher-order skills that are not easily otherwise tested, including ones essential to college and career ready performance in reading, writing and mathematics, without adding enormous burden of testing.

[11] The numbers are chosen here mostly for computational convenience, and may not reflect the actual values obtained from item precalibration, etc.

[12] For a description of approaches to providing this kind of instructional guidance in forms that do not suppress teacher ingenuity and judgment, see Resnick (in press).

Another way of validating the Distributed Accountability Exams scores would be to compare them to NAEP scores. States might expand the use of the NAEP test (every year and/or increase the percentage of students).

The American Examination System will also foster a rich environment of formative assessments that are educative in ways that directly resemble the summative system, but with more direct application to daily and weekly instruction.

- They would be aligned with the learning trajectories derived from the Common Core Standards, and thus aligned with *what* teachers need to teach.
- They would model approaches to *how* to teach, and would, at the request of educators, provide teachers structured opportunities for gaining experience in using those teaching methods.
- Teachers would make these assessments *part of their instructional routine*, rather than an addition to it. Data-entry/record-keeping burdens will be minimal, and teachers will have easy and quick access to student- and class-level reporting as well as tools to understand the instructional significance of that data. By tracking fidelity in the use of these diagnostic tools, the system will help teachers to use them appropriately.

Formative assessment tasks that cannot be machine scored will be accompanied by simple rubrics for quickly analyzing the student work. Teachers will be able to use digital devices to record these analyses. Through those devices, the teachers will also be provided with samples of answers that correspond to each level on the rubric, to help them calibrate their own analyses. As a form of professional development and to improve the reliability of analyses, teachers could also upload the student work into the system, along with their analyses, to get feedback from other teachers or subject matter experts. The formative assessments would in general not be used for summative purposes, but *metrics of teacher fidelity* in implementing the formative assessments (and their associated instructional recommendations) could be used as part of teacher/school performance management/accountability.

To enable teachers to make best use of all of these, the system will provide an online platform which includes: the honeycomb (to track student progress on learning trajectories towards college and career readiness, and to access diagnostic and instructional support for each stage of each trajectory); other dashboard tools for tracking and analyzing the progress of particular students and groups and students; and interfaces for uploading, sharing, scoring, reporting and analyzing student work.

Because the system will administer both types of assessments (Distributed Accountability Exams and formative), for a very large number of students over a period of multiple years and potentially across multiple states, and can take account of various other student, teacher and school data, it would also eventually be able to serve as an engine for the mass personalization of assessments. Mass personalization for formative assessment could be done across many dimensions to include: past student performance on assessments; teacher and school characteristics including aggregated assessment performance of students and other measures of previous effectiveness; and which

curriculum was used. This adaptive or personalized approach to assessment will enable greater precision in the data; closer alignment to the taught curriculum; and less testing.

This initial goal for mass personalization would be to apply it to formative assessment. There are, already in use, many modalities of formative assessment (diagnostic, progress monitoring, screening), each including a mix of assessment types (multiple choice, constructed response, observation). Some of these are best delivered as part of group activities and some one-on-one between a single student and teacher. Many teachers/districts use a blend of these formative assessments, which makes sense given the diverse needs of particular students at different moments of their academic development; but some teachers—who are not themselves experts in formative assessment methodologies—struggle to decide how best to integrate all of these choices into their teaching routines for their particular students.

So, in addition to providing new Educative Formative Assessments, the American Examination System would mass customize a much wider range of formative assessments at the student and class level. This is adaptive assessment at the level above individual items – it figures out *which* formative assessment to give and when – enabling teachers to get just the right next piece of information they need about their students, without wasting a lot of classroom or other school time. With this platform, teachers will be blending modes of assessment in individualized ways – varying what data they collect and how – based on what is known so far about each student. To support this, the system will host a bank of formative assessment materials, to cover the full range of diagnostic options a state or school district wishes to use, from open-source or commercial sources.

The mass personalization process can also add to the reliability and efficiency of the Distributed Accountability Exams. Above, we showed how a Bayesian model can use data from previous DAE's to make the next DAE more efficient, as long as the student is behaving consistently from one unit to the next. If the student seems to be performing unusually well (or poorly) then the Bayesian machinery can detect this and suggest a customization of the DAE to further explore what the student knows and can do.

Technology

Delivery

Integrated online delivery of all assessments. Both summative (Distributed Accountability) and formative assessments delivered to teachers and/or students across and within states through a single software platform. The system enables a coherent use of multiple types of assessments (including types that will be administered on paper and then scanned) as part of efforts to have students meet the standards and move along the skill trajectories towards college readiness and career readiness.

The honeycomb offers an interactive online map of learning trajectories based on the Common Core Standards. It provides an intuitive and accessible way for educators to understand and make use of these trajectories all the way from Pre-K through 12. It will also enable them to grasp the dependencies among and within the trajectories—for instance, identifying what level of which specific literacy skills are needed to achieve mastery of which mathematics skills. This tool can adapted for use in any state whose standards include learning trajectories comparable to those that will be in the Common Core.

The American Examination System will deliver Distributed Accountability Exams, formative assessments and available instructional options for each step along each learning trajectory, starting with mathematics and literacy for Grades 3-10. The honeycomb allows educators to visualize the sequence of assessments and instructional options aligned with the learning trajectories; they will be displayed for educators at intervals along scales that include the entire range of skills to be taught in PreK-12. Other (non-DAE) formative assessments and instructional options, including for PreK-2 and 11-12, can also be aligned and delivered through the same interface to help educators use them in a coherent way to identify and address the particular learning needs of each student as they move on the paths towards college and career readiness.

Mass customization of assessments. Because the System will administer all types of assessments for a very large number of students over a period of multiple years and potentially across multiple states, and can take account of take account of various other education data, it will be able to serve as an engine for the mass personalization of assessments. (Dimensions and benefits of mass customization discussed in Rigorous Standards and Good Instructional Practices section.) This technology is scalable—computing power is such that there is no practical limit on the amount of education data that could be included—so that as more states and more types of data are included, the more precise (and useful) the customization becomes.

Scoring

Enable teachers/schools to scan and upload student work. The American Examination System does not assume that all assessments will always be conducted with students sitting at computers. Given current school infrastructure, and given the challenge of showing mathematics work via a keyboard, it may be more efficient to continue to rely to some extent on paper-and-pencil inputs to an otherwise digital system. The continued value of these "primitive" recording tools seems especially compelling when one considers that much of the value of the new generation of assessment tasks depends on soliciting open-ended expressions of student reasoning and thinking – and in the case of mathematics this includes drawings, graphs, and explanations.

So the American Examination System includes a process to enable scanning/digital photographing, uploading, and archiving of very large volumes of paper-based student work, including for Distributed Accountability Exams, to enable remote scoring as well as online student portfolios. The scanning/photographing process, which has already been tested in North Carolina classrooms, puts minimal burdens on teachers or other school staff and does not require large per-school investments in hardware or network infrastructure.

Remote scoring workflow and interface. For the foreseeable future, assessment of open-ended expressions of student reasoning and thinking will require at least some element of human scoring. Doing this rigorously and reliably, especially in a summative context where there are stakes for teachers and schools as well as for students, requires finding a cost-effective and time-effective workflows for directing the work to remote scorers (including cross-school or cross-state grading/validation exercises; re-grading of a sample of student papers at the state level).

The American Examination System platform enables this workflow. It automates delivery of digitized student work (including paper-and-pencil work) to raters and those validating the ratings. Student identity is kept private (the raters don't know whose work it is). The online interface for remote raters presents them with the student work alongside scoring forms based on the rubric appropriate for that type of work.

Formative assessment interface. For formative assessment, the platform provides a scoring interface for teachers similar to the one for remote scoring of Distributed Accountability Exams. This interface includes tools to mark-up student work and record notes. Teachers can also easily e-mail the marked-up work to students and their parents (so they get feedback on the same day that the assessment was delivered). When electronic essay scoring technologies will be used to add precision and/or can help teachers manage the triage associated with knowing which papers might require special attention. Similar to Wireless Generation's mClass platform, the American Examination System platform could also include mobile tools that enable teachers to digitally record what they are observing while they are actively involved with the class. Because formative assessment is part of each teacher's day-to-day instruction, capturing the resulting data provides a way to track instructional fidelity (whether the teachers are using the recommended good instructional practices).

Reporting

Platform provides reports and reporting interfaces described in the *Reporting* section below.

Summative Assessments That Measure Growth and That Project Readiness

The Common Core provides a foundation for a criterion referenced examination system that is closely tied to instruction yet meets crucial criteria of technical quality of assessment. The core grade-level standards are organized as a set of trajectories or sequences of learning goals.[13] They are specified at a grain size that can be used to organize meaningful units of instruction and correspondingly meaningful assessments to judge progress toward college and career readiness.

Tasks or items for the DAEs would be pre-tested and calibrated using standard classical and multi-dimensional IRT frameworks. At the outset, two versions of each DAE would be developed. The two versions, one administered before instruction and one afterwards, would be used by the assessment developers to establish *instructional validity* of the exams. Availability of multiple forms of the DAEs would allow states and districts to use the content-based exams to plot *student growth*, along with *teacher and school effectiveness*.[14]

[13] Some of the learning sequences in the standards are based on research conducted by multiple scholars over three decades. Others are based on well-honed intuitive judgments by expert scholars and practitioners. All will require further validation-in-use over the coming years. What is new and important in the current core standards effort is that the standards are organized into multi-dimensional sequences of learning that can inform both assessment and instruction.

[14] If the pre-instruction versions are not long enough to be reliable to estimate instructional effects on individual students, then those effects would be estimated on some aggregate level.

Student growth for purposes of assessing progress toward college and career readiness can be defined as progress along the Common Core learning trajectories. In this way, the American Examination System measures the extent to which students are on track (and student growth) all the way from Pre-K through 12.

This approach allows measurement not just of whether students are on track, but also identifies which specific skill deficits are holding each of them back. It allows teachers to answer the question: what should the instructional focus be right now, to move this particular student or groups of students forward towards college and career readiness? It also identifies where instructional practices and/or curriculum may need to be reworked (where the measures show that the majority of students have not gained a skill).

The honeycomb serves as a valuable way to display these measures of student growth for the students and their parents, because it offers an easily-comprehensible map of that student's progress, relative to time, and to the standards for each grade as well as to the ultimate goals of college and career readiness.

Accessibility

All parts of our system incorporate the principles of Universal Design for Learning.

The exams can and should remove barriers for non-native English speakers and for students with special learning needs. For non-native English students, the tests should be designed so that language will not unnecessarily make the meaning of the questions unclear—so that these students will understand the exams so that they can be measured fairly.

The DAEs would mirror the instruction that students will receive in the classroom; we would carefully design and validate accessibility for students with low-incidence disabilities. Some students may deviate from the learning trajectories, but they should remain focused on academic content. The system should maintain expectations for all students and guide teachers on how all students can master concepts and skills. Assessments would be designed for all students, modifications would allow as many students as possible to be validly assessed within the system, and there would be flexibility in terms of modality of test administration and item type.

Technical Quality

The new Common Core Standards provide a foundation for a criterion-referenced examination system that is closely tied to instruction yet meets crucial criteria of technical quality of assessment. The core grade-level standards are organized as a set of trajectories or sequences of learning goals. They are specified at a grain size that can be used to organize meaningful units of instruction and correspondingly meaningful assessments.

The American Examination System includes Distributed Accountability Exams, for use over the course of the school year, which measure the specific higher-order skills that are articulated in the Common Core Standards and state standards, as well as basic knowledge. The Distributed Accountability Exams will

include extended written work and other open-ended expressions of student reasoning and thinking; in mathematics, these would include drawings, graphs, and explanations. They will assess basic knowledge both within these constructed performances and, where appropriate, in clusters of multiple-choice items.

Distributed Accountability Exams will address each of the skills/topics articulated for each year of the state and common standards. They would be built to strong criteria of content and instructional validity. Each exam would provide a reliable estimate of student knowledge on the content of an instructional unit that is explicitly targeted to a standard, or set of standards, in the Core. The collection of exam scores for a year) would provide a valid estimate of the extent to which a student (class, school) has mastered the content specified by the standards for that year.

Content Validity

The exams would match closely in both content and form the content that is expected to be taught in each of the instructional units. New instructional units, explicitly linked to the Core standards would be created to anchor the content validity of the units. Teams of independent content and instructional experts would review the model instructional units to ensure they match with the standards and instructional quality. The same teams would judge the alignment of exams to the model instructional units. This process would largely overcome the problem of weak alignment to standards that now troubles many state assessments.

Instructional Validity

Assessments are considered instructionally valid when student performance improves after quality instruction on the content of the assessment. Our development process would include tests of instructional validity, similar to the experiment-based ones used by the Pittsburgh Science of Learning Center. These tests would involve panels of teachers with good knowledge of an instructional unit's content as well as demonstrably good pedagogical skills (as judged by an expert panel). These teachers would be put into four groups. Two of the groups would teach the instructional unit that corresponds to the Distributed Accountability Exam. In one of these groups, they would get pre-test A for their students before the unit is taught and then the students would take Test B. In the second of these groups, the tests are flipped: Test B is the pre-test and Test A is given to students after the unit is taught. In the third and fourth groups, students would not be taught the particular instructional unit at that time, but would still be given the pre-tests and post-tests (one group with A as the pre-test and B as the post-test, the other with B as the pre-test and A as the post-test). Only tests that, through these experiments, systematically register improvements in student performance as a result of corresponding instruction (and demonstrate equivalence through the pre- and post-test swaps) will be included in our Distributed Accountability Exams.

Both pre-test and end of unit exams in any given year will be drawn from a bank of tasks that will be developed as part of this validation process. Items or tasks for the DAEs will also be pre-tested and calibrated using standard classical and multidimensional IRT frameworks.

Reliability

Distributed Accountability Exams would contain a mix of short constructed response items, and more extended written responses, along with sets of multiple-choice items as appropriate to the standard being examined. Short and long constructed response components would require human scoring. Research has established that when constructed response tasks are well-targeted, scoring rubrics are specific and graders are trained, a high level of inter-rater reliability can be attained (Mariano & Junker, 2007; Patz et al., 2002; Rayn & Shepard, 2008).

Student responses on constructed response items could be graded locally (within the same school but not by the students' own teacher), or by geographically and socially remote scorers (including teachers elsewhere in the district or state). These grades could be validated using one of a number of methods that have been used in European countries (e.g., cross-school or cross-state grading exercises; re-grading of a sample of student papers at the state level). Teacher participation in the grading exams and the related validation exercises (some of which could be face-to-face) is a good process for professional learning and is used in most countries. Though the process is more costly in dollars than machine scoring, it is an educative process worth building into our Examination System. Grade validation at scale would be supported by the American Examination System platform, which can enable rapid, cost-effective remote scanning, transmission, grading, validation, and reporting. To ensure protection of student privacy rights, the system has the capacity to anonymize the digitized student work before routing it to the remote scorers and validators, as well as, for limited purposes, automatic essay scoring technologies.

The Distributed Accountability Exams open the possibility for increased use of constructed responses because they are distributed over the course of the year, yielding several times more opportunity to collect data than current end-of-year tests. This also brings benefits in terms of increased test reliability. For instance, if the reliability of each single DAE hour-long exam were 0.7, the reliability of five DAE's taken together would be 5*(0.7)/(1 + 4*0.7) = 0.92. If instead, half of each DAE's testing time were used for a pretest on the next instructional unit or simply for calibrating future test items, the improvement would be 2.5*(0.7)/(1 + 1.5*0.7) = 0.85—still a high rate of reliability.

Yet to obtain these more reliable results, students would not have to sit for a 5-hour exam, or even take an end-of-year exam. They just would have to take unit exams as they normally would in the course of teaching, but now with the unit exam contributing to an overall accountability score. Another advantage is that students would be tested on recently-learned material at all times, so that nuisance effects of delayed recall would not influence measures of how well students were learning what the teachers taught; this would probably increase reliability even more.

Although the illustration above is useful, in reality it likely will not be possible to string together DAEs into a single unidimensional measurement to which classical reliability calculations apply. Instead we believe the DAEs within a subject will be at least mildly multidimensional; if we consider each DAE within a subject within a year as a measure of one proficiency, five DAEs would be measuring five different but substantively related proficiencies. These proficiencies are likely to be statistically related as well. For example in NAEP, proficiency subscales within the same subject area are typically correlated 0.8 or

higher, and seldom lower than 0.5-0.6. We can exploit these correlations by building a multidimensional Bayesian latent variable model to take advantage of proficiency estimates from old DAEs to help produce more precise proficiency estimates for the next DAE, or indeed to shorten the next DAE with no loss in measurement precision.

For example, suppose[15] we wish to estimate a student's proficiency with a margin of error of 0.2 (SEM = 0.1), and each item contributes roughly one unit of Fisher information to proficiency estimation (here we are borrowing an IRT formulation for specificity), then the student would need to answer roughly 100 items. However, if we could already predict the proficiency on this DAE with a margin of error of 0.4 using past DAE performance, we would need only roughly 20 more items to obtain a margin of error of 0.2 on this DAE.

This calculation depends on the student's performance on the new DAE being consistent, in a way that can be made precise using Bayesian modeling, with his/her performance on past DAEs. If the student's responses on the next DAE are inconsistent with their older DAE results, we would need to do follow-up testing to get a more precise estimate of the student's proficiency. Thus for students who learn consistently from one unit to the next, we can exploit past performance to help estimate proficiency on the current unit of instruction. However, for example, for the student who performs unusually well (or poorly) on the current unit, we can use the Bayesian machinery to see the inconsistency, and offer another block of items in order to more precisely assess that student's learning. A similar process is used in online tutoring systems and adaptive testing systems.

Reporting

Distributed Accountability Exams will be given throughout each year (along with corresponding pre-tests); we expect three to five per year in mathematics and literacy at each grade, although the specifics of number and timing would be worked out with states. The American Examination System platform will enable pre-tests and exams to be delivered to remote scorers and scoring validators within seconds. The scorers and score validators will have interfaces that speed their work and enables them to report back the scores with the speed of a mouse click. If the remote scorers are dedicated/hired for the task, then scoring can be done with a 24-hour turnaround time. If other teachers are asked to do the scoring, then the turnaround time would depend on state/district expectations of their own schedules.

Teachers determine when to give formative assessments. Guidance to them as to when to use which formative assessments, and for which students, comes from the honeycomb and the mass personalization engine. Where automated scoring of formative assessment is possible, scoring is done overnight; with other assessments teachers will do their own scoring on their own schedule; with some types of formative assessment (some forms of questioning or observing students' writing in notebooks), teachers can simply make online notations about what they are learning about each student or group of students.

[15] The numbers are chosen here mostly for computational convenience, and may not reflect the actual values obtained from item precalibration, etc.

Produce Results That Can Be Aggregated at the Classroom, School, District, and State Levels

Yes.

Produce Reports That Are Relevant, Actionable, Timely, and Accurate for Various Audiences

The System will enable assessment reporting to be relevant and actionable as well as timely and accurate. For each Distributed Accountability Exam (and the corresponding pre-tests), reports to teachers, districts and state systems will be populated and distributed within a few days of scoring is completed. Student and parent reports will be ready for teachers and/or other school officials to review within those same few days.

Equally important, the System will allow teachers, principals and districts—and potentially parents and students themselves—to generate custom reports in real-time on demand. These reports aggregate longitudinal data from different Distributed Accountability Exams and formative assessments to provide a more complete picture of each student, class, and school.

Reporting interfaces and reports will be based on role (teacher, student or parent, principal, district or state administrator). The System will enable role-based access rights.

Specific reporting components include:

- Comprehensive interface for teachers, which includes:
    - gradebook functionality to allow them to enter and organize their own approaches to tracking student progress and achievement;
    - dashboard reporting of Distributed Accountability Exams and formative assessments (quick reporting enabled by System workflow allows this data to be instructionally useful);
    - other student data (including demographic, high-stakes and other longitudinal assessment data);
    - tools to enable educators to share their formative assessment work with other teachers (as well as experts) to gain instructional advice and create opportunities for professional development.
- Reports to parents and students.
- Dashboard tools for tracking and analyzing the progress of particular students and groups towards college and career readiness:

- The honeycomb will enable teachers, parents and the students themselves to track individual student progress (and extent to which students are on track) from Pre-K through 12.

- Provide principals and district/state administrators with a reporting interface that includes aggregate analysis (including cross-class, cross-teacher, cross-school, cross-district and cross-state and demographic comparisons, with the longitudinal dimensions—including value-added on the high-stakes scores—ncluded in each)

Student engagement will be increased with the American Examination System, as compared to existing NCLB assessments, because the exams are directly related to what they are trying to learn, and because these assessments will be mass personalized: each student will get fewer questions that are too easy or too hard for them (with greater precision than typical adaptive tests).

Informing Instruction and Leadership

Reports will provide information about how students are progressing to meet the common standards and state standards and the learning trajectories that are captured within the standards. Our whole system is geared at the rigorous standards and good instructional practice. Every part of the series of Distributed Accountability Exams is directly designed to address a standard in each grade, with a one-to-one mapping that has not been done in the past. The new formative assessments are mapped the same way. Both types of assessments model how teachers could teach.

Because Distributed Accountability Exams look at higher-thinking ability and are tied to the specific standards for the grades that we are testing, the information will be far more useful that having information only from the summative tests that are now being used. The reports that we produce can point quickly to those specific standards and show how students have achieved, so that teachers will have diagnostic information about their students that can help craft later teaching. For example, our data will show where an individual student is with reading comprehension. Reports for teachers can also include summaries of pre-tests and formative assessment data, for diagnostic purposes and for seeing how successful is re-teaching or customized teaching to specific students.

Reports will also track students by class, by teacher, and by other subgroups that are important for policy makers to follow, so that they will be useable in informing teaching and learning.

What types of policy questions can be appropriately addressed with data from our system model?

- which teachers and schools are doing well

- efficacy of curriculum materials or professional development systems, for all students and by subgroup

- equity issues (performance by subgroup)

Leveraging Common Standards and Assessments

The American Examination System and its Distributed Accountability Exams will be directly based on and designed to follow the actual sequence of standards found in the common standards and in the standards of the states that will use this system.

Our design promotes the identification of effective practices through the honeycomb. If we are using a common system across states, that provides more data into the System to identify what instruction—including curriculum and pedagogy—is the most successful, overall and with specific subgroups of students.

Please see the *Rigorous Standards and Good Instructional Practices* section as well.

Implementation Timeline

Our vision for the American Examination System is ambitious. What makes it realistic is the substantial amount of work that has already been done in developing the content and tools needed to make it work.

For instance, IFL has extensive experience in developing model instructional units such as the ones that will be part of the system and in working with school systems to tailor the units to local needs and preferences (McConachie & Petrosky, 2010). IFL units (and accompanying assessments) are built into the curriculum guidance system of several urban school districts and have been shown to produce high levels of teacher engagement and improved instruction when accompanied by appropriate forms of professional training (David & Green, 2008; Resnick, in press; Talbert & David, 2008).

Many of the required technologies are already in use in existing assessment and data-management applications, or are now being developed through Wireless Generation and/or initiatives of the Bill and Melinda Gates Foundation. Thus, for instance, much of the platform for authoring and administering cognitively-demanding assessment items at scale will be available for online use in December 2010.

The System we describe is one that will operate fully about 3 years from the beginning of the process, with mass personalization of summative assessment playing a larger role at the end of that timeframe. Much of the system, including the Distributed Accountability Exams, Educative Formative Assessments, and other aspects of the technology platform will be operational in 2 years.

Platform

Based on the direct experience of Wireless Generation in building a system of comparable complexity (ARIS, the education information system for the country's largest public school system), we estimate that a secure and scalable version of the initial platform can be available for use in six months after work on the project formally begins; additional functionality would be available after twelve months; and a comprehensive system in use at scale in eighteen months. Additional development, related to the research and roll-out of mass customized assessments, would take place within a longer time frame (36 months).

Starting with the initial roll-out, there would be a full-time Help Desk to support teachers and schools in using the American Examination System platform.

Assessments

The platform could be used to develop and roll out assessments according to the following three phases: establishment of content validity (for both the DAEs and Educative Formative Assessments); establishment of instructional validity (for the DAEs); and then the use of the system for summative accountability purposes. States should be consulted to determine which grades and what subjects to prioritize.

- We anticipate that, after 12 months, the Distributed Accountability Exams, including the units for model instruction, for use for Grades 3-10 will receive sign-off on content validity by State Departments. The Educative Formative Assessments could begin to be used during this first year after the content is validated.
- During months 13-24, we will do the experiments to capture instructional validity, or sooner if possible, beginning as soon as content validity is established.
- After this 24 month period, the Distributed Accountability Exams would be used for summative Accountability purposes.

Training will occur in the first year for state level professionals. Starting in the second year there will be extensive teacher training.

Cost

We estimate for a typical state the ongoing costs of the System will be about the same as for current NCLB tests. Current expenditures are typically $20-$30 per student, and in some cases higher than $80, to cover reading and mathematics (U.S. Department of Education, 2010).

Administering the Distributed Exams (including the pre-instruction version) will cost more to develop and score than the current high-stakes tests, if for no other reason than their frequency. But the current interim exams, and expenses associated with those (typically $15-$20 or more per student), could be eliminated.

Teachers within the school district could score the exams from each other's students, but a significant portion of the ongoing cost would be from validation of samples of teacher scoring.

Apart from provision of tablet/handheld and scanning devices for teachers (off-the-shelf, industry-standard technologies that are coming down in price each year), costs associated with the maintenance of the technology platform would be minimal when considered on a per-student basis.

Limitations

We have learned to live with standardized tests as accountability measures for decades. Now, some important things are changing: setting new accountability goals of college-and career-readiness; measuring student growth over time and establishing a meaning of *teacher effectiveness* that relies heavily on student learning. So, the creators of the American Examination System are suggesting that if we are to have our student population college-ready, we need to move the accountability system directly into good teaching and learning. That is, create and use distributed exams that model the kinds of student performance that we are aiming for, and provide customized diagnostic tools for instruction that meet individual student needs. That sounds very different from what we have. One limitation might be the reluctance of some from the public, policy makers, educators or statisticians to make such a large shift.

We hope to overcome such reluctance, since the new common standards and the goals of Race to the Top create an opportunity to embed accountability with guidance to teachers to ensure that students can achieve to the newer, higher standards.

Value Versus Burden

The overarching value of our system is that it moves accountability toward improving instruction. The Distributed Accountability Exams measure higher thinking skills and processes, as well as basic knowledge and will show the extent to which students are meeting the higher, common standards. This system will have great content relevance and offer instructional feedback that is far greater that any system that is out there right now. Strict standardization, that has governed testing for decades, is replaced by greater precision and greater usefulness.

For some states, using this System would increase testing time, with multiple assessments every year. But most states already have interim tests along with the summative high-stakes test, so in these cases the increase would be less. In all cases, increased testing time should be seen in the context of a system in which all assessments support the instructional process (as opposed to the current high-stakes system which impedes it). In particular, a benefit of our formative assessments over the current interim exams is that ours will not look like, nor be, an added burden because they are based exactly on the designed curriculum.

Remote scoring by teachers within a district does create a burden for those teachers (as compared to having an assessment vendor do the scoring), but there is considerable evidence that this activity can be one of the most valuable forms of professional development.

The American Examination System platform offers teachers, principals and district/state offices online interfaces, including the honeycomb, that bring all the assessment data together along with other student, class and school information, in formats that make the data actionable to an unprecedented extent. This also reduces the potential burden that might arise from a distributed assessment system.

References

Anagnostopoulos, D. (2003). Testing and student engagement with literature in urban classrooms: A multi-layered perspective. *Research in the teaching of English*, *38*(2), 177-212.

David, J. L., & Greene, D. (2008). *Improving mathematics instruction in Los Angeles High Schools: Follow up to the evaluation of the PRISMA Pilot Program.* Palo Alto, CA: Bay Area Research Group.

Engel, S. (2010, February 1). Learning to play. *The New York Times.* Retrieved from http://www.nytimes.com/2010/02/02/opinion/02engel.html

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. L. Brennen (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics, 32,* 287–314.

McConachie, S. M., & Petrosky, A. R. (Eds.). (2009). *Content matters: A disciplinary literacy approach to improving student learning.* San Francisco, CA: Jossey-Bass.

McNeill, L. (2002). *Contradictions of school reform: Educational costs of educational testing.* New York, NY: Routledge.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341–384;

Phillips, V., & Wong, C. (2010, February). Tying together the Common Core of Standards, instruction, and assessments. *Phi Delta Kappan, 91*(5),. 37-42.

Rayn, K. E., & Shepard L. A. (Eds). (2008). *The future of test-based educational accountability.* Mahwah, NJ: Lawrence Erlbaum Associates.

Resnick, L. B. (in press) Nested learning systems for the thinking curriculum. *Educational Researcher.*

Resnick, D. P., & Resnick, L. B. (1977). The nature of literacy: An historical exploration. *Harvard Educational Review, 47*(3), 370-385.

Resnick, D. P., & Resnick, L. B. (1980). The nature of literacy: An historical exploration. In M. Wolf, M. K. McQuillan, & E. Radwin (Eds.), *Thought and language/language and reading* (pp. 396-411). Cambridge, MA: Harvard University Press.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.

Resnick, L. B., Stein, M. K., & Coon, S. (2008). Standards-based reform: A powerful idea unmoored. In R. D. Kahlenberg (Ed.), *Improving on No Child Left Behind: Getting education reform back on track*. New York, NY: The Century Foundation Press.

Rowan, B., Correnti, R., Miller, R., & Camburn. E. (2009). School improvement by design: Lessons from a study of comprehensive school reform programs. In B. Schneider, D. Sykes, & D. N. Plank (Eds.), *AERA handbook on education policy research*. New York, NY: Routledge.

Talbert, J.E., & David, J.L., with Lin, W. (2008, September). *Final report. Evaluation of the Disciplinary Literacy-Professional Learning Community (DL-PLC) Initiative in Austin Independent School District.* Palo Alto, CA: Stanford University, Center for Research on the Context of Teaching.

U.S. Department of Education. (2010). *State and local implementation of the No Child Left Behind (NCLB) Act: Volume IX—Accountability under NCLB. Final report.* (2010). Washington, DC: Author.

Appendix A

Common Core State Standards—Mathematics[16]

	Standards for Mathematical Understanding *Students understand that:*	Standards for Mathematical Skill *Students:*
Sixth Grade **Ratios and Proportional Relationships**	1. Two quantities are said to be in a ratio of a to b when for every a units of the first quantity there are b units of the second. For example, in a flock of birds, the ratio of wings to beaks might be 2 to 1; this ratio is also written 2:1.[6.1] 2. Multiplying both terms in a ratio by the same nonzero whole number produces an equivalent ratio, e.g., 3:2 and 12:8 are equivalent ratios. 3. Ratios are connected to fractions by the unit rate a/b. If there are a units of the first quantity for every b units of the second, where $b \neq 0$, then there are a/b units of the first quantity for 1 unit of the second. For example, if a recipe has a ratio of 3 cups of flour to 4 cups of sugar, then there is $3/4$ cup of flour for each cup of sugar. 4. Motion with constant speed means that distances covered and elapsed times are in equivalent ratios; the unit rate is distance divided by time, or speed. In a ratio of quantities of different kinds, dividing to obtain the unit rate produces a new kind of quantity.	a. Make tables of equivalent ratios relating quantities with whole-number measurements, find missing values in the tables, and plot the pairs of values on the coordinate plane. b. Solve for an unknown quantity in a problem involving two equal ratios. c. Solve unit rate problems including unit pricing and constant speed, including reasoning with equations such as $d = r \times t$, $r = d/t$, $t = d \div r$, etc. d. *Describe categorical data sets using ratios (e.g., for every vote candidate A received, candidate C received nearly three votes; the number of people with blood type O was four and a half times as large as the number of people with blood type B).[6.3]

[16] Common Core State Standards Initiative MATHEMATICS

Standards for Mathematical Practice

Proficient students of all ages expect mathematics to make sense. They take an active stance in solving mathematical problems. When faced with a non-routine problem, they have the courage to plunge in and try something, and they have the procedural and conceptual tools to carry through. They are experimenters and inventors, and can adapt known strategies to new problems. They think strategically.

Students who engage in these practices individually and with their classmates discover ideas and gain insights that spur them to pursue mathematics beyond the classroom walls. They learn that effort counts in mathematical achievement. The practices described below are those that expert mathematical thinkers encourage in apprentices. Encouraging these practices in students of all ages should be as much a goal of the mathematics curriculum as is teaching specific content topics and procedures.

Taken together with the Standards for Mathematical Understanding and the Standards for Mathematical Skill, the Standards for Mathematical Practice provide the infrastructure for mathematical teaching and learning that will help ensure students are prepared for entry into college courses or career pathways when they graduate from high school.

1 Attend to precision.

Mathematically proficient students organize their own ideas in a way that can be communicated precisely to others, and they analyze and evaluate others' mathematical thinking and strategies noting the assumptions made. They clarify definitions in discussion with others and in their own reasoning. They state the meaning of the symbols they choose, are careful about specifying units of measure and labeling axes to clarify the correspondence with quantities in a problem, and express their answers with a degree of precision appropriate for the problem context. In the elementary grades, students explain precise concepts to each other, such as why 1 is not a prime number. By the time they reach high school they have learned to examine the generality of claims and definitions and to make explicit the domain of a statement.

2 Construct viable arguments and critique the reasoning of others.

Mathematically proficient students understand and use stated assumptions, definitions and previously established results in constructing arguments. They make conjectures and build a logical progression of statements to explore the truth of their conjectures. They break things into cases and can recognize and use counterexamples. They use logic to justify their conclusions, communicate them to others and respond to the arguments of others. They reason inductively about data, making plausible arguments that take into account the context from which the data arose. Mathematically proficient students are also able to compare the effectiveness of two plausible arguments, distinguish correct logic or reasoning from that which is flawed, and—if there is a flaw in an argument—explain what it is. Elementary students can construct arguments using concrete referents such as objects, drawings, diagrams and actions. The arguments can make sense, even though they are not generalized until later grades. Later they can learn to analyze the scope of the domain to which an argument applies. Students at all grades can listen or read the arguments of others, decide whether they make sense and ask useful questions to clarify or improve the argument.

3 Make sense of problems and persevere in solving them.

Mathematically proficient students start by explaining to themselves the meaning of a problem and looking for entry points to its solution. They analyze givens, constraints, relationships and goals. They make conjectures about the form and meaning of the answer and plan a solution pathway rather than simply jumping into a solution attempt. They consider analogous problems, try special cases and work on simpler forms or easy numbers. They monitor and evaluate their progress and change course if necessary. Older students might, depending on the context of the problem, put algebraic expressions into different forms or change the viewing window on their graphing calculator to get the information they need. Mathematically proficient students can explain correspondences between equations, verbal descriptions, tables, and graphs or draw diagrams of important features and relationships, graph data, and search for regularity or trends. Younger students might rely on using concrete objects or pictures to help conceptualize and simply a problem. Mathematically proficient students verify their answers to problems using a different method, and they continually ask themselves, "Does this make sense?" They can understand the approaches of others to solving complex problems and identify the correspondences between approaches and their shared mathematics.

4 Look for and make use of structure.

Mathematically proficient students look closely to discern a pattern or structure. Young students, for example, might notice that seven and three more must be the same amount as three and seven more, or they may sort a collection of shapes according to how many sides the shapes have. Later students will see that the distributive property means that 7 x 8 equals $7(5 + 3) =$ the well remembered $7 \times 5 + 7 \times 3$. In the expression $x^2 + 9x + 14$, older mathematically proficient students can see the 14 as 2×7 and the 9 as $2 + 7$. They recognize the significance of an existing line in a geometric figure and can add an auxiliary line to make the solution of a problem clear. They also can step back for an overview and shift perspective. They can see complicated things, such as some algebraic expressions, as single objects. At the same time, they can also see $5 - 3(x - y)^2$ as 5 minus a positive number times a square and realize that it cannot be more than 5 for any real numbers x and y.

5 Look for and express regularity in repeated reasoning.

Mathematically proficient students pay attention to repeated calculations as they carry them out, and look both for general algorithms and for shortcuts. Upper elementary students might notice that dividing 25 by 11 that they are repeating the same calculations over and over again and conclude they have a repeating decimal. For example, by paying attention to the calculation of slope as they repeatedly check whether points are on the line through (1, 2) with slope 3, middle or high school students might abstract the equation $(y - 2)/(x - 1) = 3$. Noticing the regularity in the way terms cancel in the expansions of $(x - 1)(x + 1)$, $(x - 1)(x^2 + x + 1)$, and $(x - 1)(x^3 + x^2 + x + 1)$ might lead them to the general formula for the sum of a geometric series. As they work through the solution to a problem, mathematically proficient students maintain oversight of the process, while attending to the details. They continually evaluate the reasonableness of their intermediate results.

6 Reason abstractly and quantitatively.

Mathematically proficient students make sense of the quantities and their relationships in problem situations. Students bring two complementary abilities to bear on problems involving quantitative relationships: the ability to *decontextualize*—to abstract a given situation and represent it symbolically and manipulate the representing symbols as if they have a life of their own, without necessarily attending to their referents—and the ability to *contextualize*, to pause as needed during the manipulation process in order to probe into the referential meanings for the symbols involved in the manipulation. Quantitative reasoning entails habits of creating a coherent image of the problem at hand; considering the units involved; continually attending to the meaning of quantities, not just how to compute them; and having multiple images of a concept and being flexible in transitioning among them.

7 Model with mathematics.

Mathematically proficient students can apply the mathematics they know to solve problems arising in life, society and the workplace. In early grades this might be as simple as writing an addition equation to describe a situation. In middle grades, a student might apply proportional reasoning to plan a school event or analyze a problem in the community. By high school, a student might use geometry to solve a design problem or use a function to describe how one quantity of interest depends on another. Mathematically proficient students who can apply what they know are comfortable making assumptions and approximations to simplify a complicated situation. They are able to identify important quantities in a practical situation and map their relationships using such tools as diagrams, 2-by-2 tables, graphs, flowcharts and formulas. They can analyze those relationships mathematically to draw conclusions. They routinely interpret their mathematical results in the context of the practical situation and reflect on whether the results make sense, possibly improving the model if it has not served its purpose.

8 Use appropriate tools strategically

Mathematically proficient students consider the available tools when solving a mathematical problem. These tools might include such things as pencil and paper, concrete models, ruler, protractor, calculator, spreadsheet, computer algebra system, statistical package, or dynamic geometry software. Proficient students are familiar enough with tools appropriate for their grade or course to make sound decisions about when each might be helpful, realizing the limitations of the tools and the output that they generate. For example, mathematically proficient high school students are able to apply their understanding of limits of technology output to interpret graphs of functions and approximate solutions generated using a graphing calculator. They detect possible errors by using mathematical understanding and estimation strategically. When making mathematical models, they know that technology can enable them to vary assumptions, explore consequences and compare predictions with data. Mathematically proficient students at various grade levels are able to identify relevant external mathematical resources, such as digital content located on a website, and use them to solve problems. They are also able to use technological tools to explore and deepen their understanding of concepts.

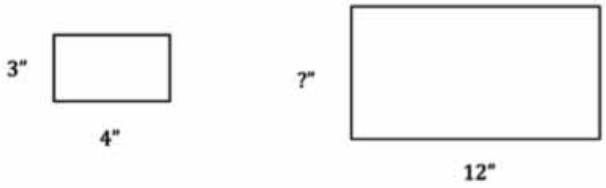
Mathematics Exam—6th Grade

1. The candy jar contains Jolly Ranchers (the rectangles) and Jawbreakers (the circles). Solve each of the following problems:

   a. What is the ratio of Jolly Ranchers (the rectangles) to Jawbreakers (the circles)? [Standard 1]
   b. Write a ratio equivalent to the ratio in part a. Explain how you got it. [Standard 2]
   c. Suppose you had a new candy jar with the same ratio of Jolly Ranchers to Jawbreakers as shown above, but it contained 100 Jolly Ranchers. How many Jawbreakers would you have? Explain how you found your answer. [Standard b]

   Adapted from Smith, Silver, Stein, Boston, Henningsen, & Hillen (2005, p.26)

2. Katie and Jacob are enlarging pictures for the school yearbook and they want to make sure they do not distort the image so they need to make sure that the ratio of the new larger photo is the same as the ratio of the original smaller photo. They have a photograph that is 3" by 4", and they need to enlarge it so that the length of the photo is 12". Katie says that the width of the enlarged photograph will be 9", but Jacob thinks the width will be 11". Who do you think is correct? Explain how you know. [Standard 2]



3. A secondhand store will trade 4 of their comic books for every 5 a customer brings in. Kyle constructed a table so he could find how many comic books they will trade for 35 of his, but as you can see, he did not fill in all the values. Complete the table for Kyle. [Standard a]

Number of Books Kyle Brings	Number of Books Kyle Gets From the Bookstore
5	4
10	8
20	
30	24
35	

4. Two friends buy broccoli at different booths at a farmer's market.
   Amanda buys 5 pounds for 6 dollars.
   Zoe buys 6 pounds for 5 dollars.

   Amanda says,
   "My price per pound is closer to $1 than yours is."

   Zoe says,
   "No, my price per pound is closer to $1 than yours is."

   A farmer overhears them and says,
   "No, your 'per pound' prices are equally close to $1."

   Who is right? Explain your answer so that all three people can understand why your answer is correct. [Standards 3, c]

5. Louis rode his bike to his grandmother's house. It took him 40 minutes. The next time he drove to his grandmother's house with his mom he noted that it was 4 miles from his house. What was Louis' speed on his bike trip? Explain how you got your answer. [Standard 4]

6. Find the cost unit price for each of the following: [Standard c]
   10 oz of cereal for $3.20
   3 cans of peas for 99¢
   5 pounds of hamburger for $13.00

7. In a recent poll of 150 TV viewers, 100 said that they watched American Idol every week while the other 50 said they have never watched the program. Describe the ratio of TV viewers who do and do not watch American Idol. [Standard d]

8. If a phone call costs 20¢ per minute, how much will the call cost if it lasts (a) 2 minutes (b) 3 minutes (c) 4 minutes (d) 5 minutes? [Standard c]

9. Four cars were driving in a race. The first car traveled 120 miles in 80 minutes. The second car traveled 60 miles in 40 minutes. The third car traveled 180 miles in 120 minutes. The fourth car traveled 100 miles in 75 minutes. For each car, plot an (x, y) pair in the coordinate plane where x is the time and the y coordinate is the distance traveled. [Standard a]

10. At a tennis camp, there were 90 right-handed players and 5 left-handed players. How many right-handed players were there for each left-handed player? [Standard 3]

Appendix B

Sample English Language Arts Examination

Ninth Grade Informational Texts

The students will be given three speeches to read (these are attached):

"I Have a Dream" Martin Luther King; August 28, 1963

"Remarks to the Convocation of the Church of God in Christ" William J. Clinton; November 13, 1993

"Ending Racial Inequality" George W. Bush; NAACP Annual Convention; July 10, 2000

Please answer the following questions:

1. Write a 1-2 paragraph summary of the key ideas in Bush's speech. What language in the speech (give quotes and line numbers) provides evidence that you have identified the key ideas? (Standard 1: Draw from evidence throughout the text to support an analysis of what the text says explicitly as well as to draw inferences from the text.)

2. Choose one key idea in Clinton's speech. Write why you think it is key. Use evidence from the speech to support your response. (Standard 2: Trace in detail the development of an idea, including how it emerges and is shaped and refined by specific details.)

3. Pick a key metaphor that King uses and explain how it contributes to his argument. Use evidence from the text to support your response. (Standard 4: Analyze how an author uses analogies, metaphors, or other comparisons strategically to frame an argument, explanation, or description.)

4. Bush and King both argue for ending racial inequality. Complete the chart below to tell how their solutions are alike, and how their solutions are different. (Standard 6: Compare and contrast how different authors construct and develop different points of views or perspectives on similar events or issues, assessing their assumptions, evidence, and reasoning.)

5. Select a key idea from King's speech that you think Clinton may be responding to. Give the line numbers in which this idea is stated. Write a paragraph in which you summarize the idea and explain how Clinton is responding to it. Use evidence from both speeches to support your response. (Standard 9: Analyze how authors argue or respond to one another's ideas or accounts of key events, evaluating the strength of each author's claims.)

The standards reprinted by permission from the January draft of the ELA Core Standard for Informational Texts for students in 9th-10th grade.

Vocabulary

I Have a Dream

Martin Luther King, Jr.

Delivered on the steps at the Lincoln Memorial in Washington D.C. on August 28, 1963. Source: Martin Luther King, Jr: The Peaceful Warrior, Pocket Books, NY 1968

1. Five score years ago, a great American, in whose symbolic shadow we stand signed the Emancipation Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of captivity. But one hundred years later, we must face the tragic fact that the Negro is still not free.

2. One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination. One hundred years later, the Negro lives on a lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later, the Negro is still languishing in the corners of American society and finds himself an exile in his own land.

3. So we have come here today to dramatize an appalling condition. In a sense we have come to our nation's capital to cash a check. When the architects of our republic wrote the magnificent words of the Constitution and the Declaration of Independence, they were signing a promissory note to which every American was to fall heir.

4. This note was a promise that all men would be guaranteed the inalienable rights of life, liberty, and the pursuit of happiness. It is obvious today that America has defaulted on this promissory note insofar as her citizens of color are concerned. Instead of honoring this sacred obligation, America has given the Negro people a bad check which has come back marked "insufficient funds." But we refuse to believe that the bank of justice is bankrupt. We refuse to believe that there are insufficient funds in the great vaults of opportunity of this nation.

    - In paragraph 1, what does the underlined word "momentous" mean?
    - In paragraph 2, what does the underlined word "prosperity" mean?
    - In paragraph 3, what does the underlined word "magnificent" mean?
    - In paragraph 3, what does the underlined phrase "promissory note" mean?
    - In paragraph 4, what does the underlined word "defaulted" mean?

I Have a Dream

Martin Luther King, Jr.

Delivered on the steps at the Lincoln Memorial in Washington, D.C., on August 28, 1963. Source: Martin Luther King, Jr.: The Peaceful Warrior, Pocket Books, NY 1968

1. Five score years ago, a great American, in whose symbolic shadow we stand signed the Emancipation Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of captivity. But one hundred years later, we must face the tragic fact that the Negro is still not free.

2. One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination. One hundred years later, the Negro lives on a lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later, the Negro is still languishing in the corners of American society and finds himself an exile in his own land.

3. So we have come here today to dramatize an appalling condition. In a sense we have come to our nation's capital to cash a check. When the architects of our republic wrote the magnificent words of the Constitution and the Declaration of Independence, they were signing a promissory note to which every American was to fall heir.

4. This note was a promise that all men would be guaranteed the inalienable rights of life, liberty, and the pursuit of happiness. It is obvious today that America has defaulted on this promissory note insofar as her citizens of color are concerned. Instead of honoring this sacred obligation, America has given the Negro people a bad check which has come back marked "insufficient funds." But we refuse to believe that the bank of justice is bankrupt. We refuse to believe that there are insufficient funds in the great vaults of opportunity of this nation.

5. So we have come to cash this check -- a check that will give us upon demand the riches of freedom and the security of justice. We have also come to this hallowed spot to remind America of the fierce urgency of now. This is no time to engage in the luxury of cooling off or to take the tranquilizing drug of gradualism. Now is the time to rise from the dark and desolate valley of segregation to the sunlit path of racial justice. Now is the time to open the doors of opportunity to all of God's children. Now is the time to lift our nation from the quicksands of racial injustice to the solid rock of brotherhood.

6. It would be fatal for the nation to overlook the urgency of the moment and to underestimate the determination of the Negro. This sweltering summer of the Negro's legitimate discontent will not pass until there is an invigorating autumn of freedom and equality. Nineteen sixty-three is not an end, but a beginning. Those who hope that the Negro needed to blow off steam and will now be content will have a rude awakening if the nation returns to business as usual. There will be neither rest nor tranquility in America until the Negro is granted his citizenship rights.

7. The whirlwinds of revolt will continue to shake the foundations of our nation until the bright day of justice emerges. But there is something that I must say to my people who stand on the warm threshold which leads into the palace of justice. In the process of gaining our rightful place we must not be guilty of wrongful deeds. Let us not seek to satisfy our thirst for freedom by drinking from the cup of bitterness and hatred.

8. We must forever conduct our struggle on the high plane of dignity and discipline. We must not allow our creative protest to degenerate into physical violence. Again and again we must rise to the majestic heights of meeting physical force with soul force.

9. The marvelous new militancy which has engulfed the Negro community must not lead us to distrust of all white people, for many of our white brothers, as evidenced by their presence here today, have come to realize that their destiny is tied up with our destiny and their freedom is inextricably bound to our freedom.

10. We cannot walk alone. And as we walk, we must make the pledge that we shall march ahead. We cannot turn back. There are those who are asking the devotees of civil rights, "When will you be satisfied?" we can never be satisfied as long as our bodies, heavy with the fatigue of travel, cannot gain lodging in the motels of the highways and the hotels of the cities. We cannot be satisfied as long as the Negro's basic mobility is from a smaller ghetto to a larger one. We can never be satisfied as long as a Negro in Mississippi cannot vote and a Negro in New York believes he has nothing for which to vote. No, no, we are not satisfied, and we will not be satisfied until justice rolls down like waters and righteousness like a mighty stream.

11. I am not unmindful that some of you have come here out of great trials and tribulations. Some of you have come fresh from narrow cells. Some of you have come from areas where your quest for freedom left you battered by the storms of persecution and staggered by the winds of police brutality. You have been the veterans of creative suffering. Continue to work with the faith that unearned suffering is redemptive.

12. Go back to Mississippi, go back to Alabama, go back to Georgia, go back to Louisiana, go back to the slums and ghettos of our northern cities, knowing that somehow this situation can and will be changed. Let us not wallow in the valley of despair. I say to you today, my friends, that in spite of the difficulties and frustrations of the moment, I still have a dream. It is a dream deeply rooted in the American dream.

13. I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident: that all men are created equal." I have a dream that one day on the red hills of Georgia the sons of former slaves and the sons of former slaveowners will be able to sit down together at a table of brotherhood. I have a dream that one day even the state of Mississippi, a desert state, sweltering with the heat of injustice and oppression, will be transformed into an oasis of freedom and justice. I have a dream that my four children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character. I have a dream today.

14. I have a dream that one day the state of Alabama, whose governor's lips are presently dripping with the words of interposition and nullification, will be transformed into a situation where little black

boys and black girls will be able to join hands with little white boys and white girls and walk together as sisters and brothers. I have a dream today. I have a dream that one day every valley shall be exalted, every hill and mountain shall be made low, the rough places will be made plain, and the crooked places will be made straight, and the glory of the Lord shall be revealed, and all flesh shall see it together. This is our hope. This is the faith with which I return to the South. With this faith we will be able to hew out of the mountain of despair a stone of hope. With this faith we will be able to transform the jangling discords of our nation into a beautiful symphony of brotherhood. With this faith we will be able to work together, to pray together, to struggle together, to go to jail together, to stand up for freedom together, knowing that we will be free one day.

15. This will be the day when all of God's children will be able to sing with a new meaning, "My country, 'tis of thee, sweet land of liberty, of thee I sing. Land where my fathers died, land of the pilgrim's pride, from every mountainside, let freedom ring." And if America is to be a great nation, this must become true. So let freedom ring from the prodigious hilltops of New Hampshire. Let freedom ring from the mighty mountains of New York. Let freedom ring from the heightening Alleghenies of Pennsylvania! Let freedom ring from the snowcapped Rockies of Colorado! Let freedom ring from the curvaceous peaks of California! But not only that; let freedom ring from Stone Mountain of Georgia! Let freedom ring from Lookout Mountain of Tennessee! Let freedom ring from every hill and every molehill of Mississippi. From every mountainside, let freedom ring.

16. When we let freedom ring, when we let it ring from every village and every hamlet, from every state and every city, we will be able to speed up that day when all of God's children, black men and white men, Jews and Gentiles, Protestants and Catholics, will be able to join hands and sing in the words of the old Negro spiritual, “Free at last! free at last! thank God Almighty, we are free at last!”

from REMARKS TO THE CONVOCATION OF THE CHURCH OF GOD IN CHRIST

William J. Clinton

November 13, 1993, 11:51 A.M.

1. If Martin Luther King were to reappear by my side today and give us a report card on the last 25 years, what would he say? “You did a good job,” he would say, “voting and electing people who formerly were not electable because of the color of their skin. You have more political power, and that is good.”

2. “You did a good job,” he would say, “letting people who have the ability to do so live wherever they want to live, go wherever they want to go in this great country.”

3. “You did a good job,” he would say, “elevating people of color into the ranks of the United States Armed Forces to the very top or into the very top of our Government.”

4. “You did a very good job,” he would say, “creating a black middle class of people who really are doing well, and the middle class is growing more among African-Americans than among non-African-Americans. You did a good job; you did a good job in opening opportunity.”

5. “But,” he would say, “I did not live and die to see the American family destroyed. I did not live and die to see 13-year-old boys get automatic weapons and gun down 9-year-olds just for the kick of it. I did not live and die to see young people destroy their own lives with drugs and then build fortunes destroying the lives of others. That is not what I came here to do.”

6. “I fought for freedom,” he would say, “but not for the freedom of people to kill each other with reckless abandon, not for the freedom of children to have children and the fathers of the children walk away from them and abandon them as if they don't amount to anything. I fought for people to have the right to work but not to have whole communities and people abandoned. This is not what I lived and died for.”

7. “My fellow Americans,” he would say, “I fought to stop white people from being so filled with hate that they would wreak violence on black people. I did not fight for the right of black people to murder other black people with reckless abandon.”

8. The other day the Mayor of Baltimore, a dear friend of mine, told me a story of visiting the family of a young man who had been killed -- -18 years old -- on Halloween. He always went out with little bitty kids so they could trick-or-treat safely. And across the street from where they were walking on Halloween, a 14-year-old boy gave a 13-year-old boy a gun and dared him to shoot the 18-year-old boy, and he shot him dead. And the Mayor had to visit the family.

9. In Washington, DC, where I live, your Nation's Capital, the symbol of freedom throughout the world, look how that freedom is being exercised. The other night a man came along the street and grabbed a 1-year-old child and put the child in his car. The child may have been the child of the man. And two

people were after him, and they chased him in the car, and they just kept shooting with reckless abandon, knowing that baby was in the car. And they shot the man dead, and a bullet went through his body into the baby's body, and blew the little bootie off the child's foot.

10. The other day on the front page of our paper, the Nation's Capital, are we talking about world peace or world conflict? No, big article on the front page of the Washington Post about an 11-year-old child planning her funeral: "These are the hymns I want sung. This is the dress I want to wear. I know I'm not going to live very long." That is not the freedom, the freedom to die before you're a teenager is not what Martin Luther King lived and died for.

11. More than 37,000 people die from gunshot wounds in this country every year. Gunfire is the leading cause of death in young men. And now that we've all gotten so cool that everybody can get a semiautomatic weapon, a person shot now is 3 times more likely to die than 15 years ago, because they're likely to have three bullets in them. A hundred and sixty thousand children stay home from school every day because they are scared they will be hurt in their schools.

12. The other day I was in California at a town meeting, and a handsome young man stood up and said, "Mr. President, my brother and I, we don't belong to gangs. We don't have guns. We don't do drugs. We want to go to school. We want to be professionals. We want to work hard. We want to do well. We want to have families. And we changed our school because the school we were in was so dangerous. So when we showed up to the new school to register, my brother and I were standing in line and somebody ran into the school and started shooting a gun. My brother was shot down standing right in front of me at the safer school." The freedom to do that kind of thing is not what Martin Luther King lived and died for, not what people gathered in this hallowed church for the night before he was assassinated in April of 1968. If you had told anybody who was here in this church on that night that we would abuse our freedom in that way, they would have found it hard to believe. And I tell you, it is our moral duty to turn it around.

13. And now I think finally we have a chance. Finally, I think, we have a chance. We have a pastor here from New Haven, Connecticut. I was in his church with Reverend Jackson when I was running for President on a snowy day in Connecticut to mourn the death of children who had been killed in that city. And afterward we walked down the street for more than a mile in the snow. Then, the American people were not ready. People would say, "Oh, this is a terrible thing, but what can we do about it?"

14. Now when we read that foreign visitors come to our shores and are killed at random in our fine State of Florida, when we see our children planning their funerals, when the American people are finally coming to grips with the accumulated weight of crime and violence and the breakdown of family and community and the increase in drugs and the decrease in jobs, I think finally we may be ready to do something about it.

15. And there is something for each of us to do. There are changes we can make from the outside in; that's the job of the President and the Congress and the Governors and the mayors and the social

service agencies. And then there's some changes we're going to have to make from the inside out, or the others won't matter. That's what that magnificent song was about, isn't it? Sometimes there are no answers from the outside in; sometimes all the answers have to come from the values and the stirrings and the voices that speak to us from within.

16. So we are beginning. We are trying to pass a bill to make our people safer, to put another 100,000 police officers on the street, to provide boot camps instead of prisons for young people who can still be rescued, to provide more safety in our schools, to restrict the availability of these awful assault weapons, to pass the Brady bill and at least require people to have their criminal background checked before they get a gun, and to say, if you're not old enough to vote and you're not old enough to go to war, you ought not to own a handgun, and you ought not to use one unless you're on a target range.

17. We want to pass a health care bill that will make drug treatment available for everyone. And we also have to do it. We have to have drug treatment and education available to everyone and especially those who are in prison who are coming out. We have a drug czar now in Lee Brown, who was the police chief of Atlanta, of Houston, of New York, who understands these things. And when the Congress comes back next year, we will be moving forward on that.

18. We need this crime bill now. We ought to give it to the American people for Christmas. And we need to move forward on all these other fronts. But I say to you, my fellow Americans, we need some other things as well. I do not believe we can repair the basic fabric of society until people who are willing to work have work. Work organizes life. It gives structure and discipline to life. It gives meaning and self-esteem to people who are parents. It gives a role model to children.

19. The famous African-American sociologist William Julius Wilson has written a stunning book called "The Truly Disadvantaged" in which he chronicles in breathtaking terms how the inner cities of our country have crumbled as work has disappeared. And we must find a way, through public and private sources, to enhance the attractiveness of the American people who live there to get investment there. We cannot, I submit to you, repair the American community and restore the American family until we provide the structure, the values, the discipline, and the reward that work gives.

20. I read a wonderful speech the other day given at Howard University in a lecture series funded by Bill and Camille Cosby, in which the speaker said, "I grew up in Anacostia years ago. Even then it was all black, and it was a very poor neighborhood. But you know, when I was a child in Anacostia, a 100 percent African-American neighborhood, a very poor neighborhood, we had a crime rate that was lower than the average of the crime rate of our city. Why? Because we had coherent families. We had coherent communities. The people who filled the church on Sunday lived in the same place they went to church. The guy that owned the drug-store lived down the street. The person that owned the grocery store lived in our community. We were whole." And I say to you, we have to make our people whole again.

21. This church has stood for that. Why do you think you have 5 million members in this country? Because people know you are filled with the spirit of God to do the right thing in this life by them. So I say to you, we have to make a partnership, all the Government agencies, all the business folks; but where there are no families, where there is no order, where there is no hope, where we are reducing the size of our armed services because we have won the cold war, who will be there to give structure, discipline, and love to these children? You must do that. And we must help you. Scripture says, "you are the salt of the Earth and the light of the world, that if your light shines before men they will give glory to the Father in heaven." That is what we must do.

22. That is what we must do. How would we explain it to Martin Luther King if he showed up today and said, yes, we won the cold war. Yes, the biggest threat that all of us grew up under, communism and nuclear war, communism gone, nuclear war receding. Yes, we developed all these miraculous technologies. Yes, we all have got a VCR in our home; it's interesting. Yes, we get 50 channels on the cable. Yes, without regard to race, if you work hard and play by the rules, you can get into a service academy or a good college, you'll do just great. How would we explain to him all these kids getting killed and killing each other? How would we justify the things that we permit that no other country in the world would permit? How could we explain that we gave people the freedom to succeed, and we created conditions in which millions abuse that freedom to destroy the things that make life worth living and life itself? We cannot.

23. And so I say to you today, my fellow Americans, you gave me this job, and we're making progress on the things you hired me to do. But unless we deal with the ravages of crime and drugs and violence and unless we recognize that it's due to the breakdown of the family, the community, and the disappearance of jobs, and unless we say some of this cannot be done by Government, because we have to reach deep inside to the values, the spirit, the soul, and the truth of human nature, none of the other things we seek to do will ever take us where we need to go.

24. So in this pulpit, on this day, let me ask all of you in your heart to say: We will honor the life and the work of Martin Luther King. We will honor the meaning of our church. We will, somehow, by God's grace, we will turn this around. We will give these children a future. We will take away their guns and give them books. We will take away their despair and give them hope. We will rebuild the families and the neighborhoods and the communities. We won't make all the work that has gone on here benefit just a few. We will do it together by the grace of God.

25. Thank you.

Ending Racial Inequality

George W. Bush

NAACP Annual Convention, Baltimore, Maryland

July 10, 2000

1. The history of the Republican Party and the NAACP has not been one of regular partnership. But our nation is harmed when we let our differences separate us and divide us. So, while some in my party have avoided the NAACP, and while some in the NAACP have avoided my party, I am proud to be here today.

2. I am here today because I believe there is much we can do together to advance racial harmony and economic opportunity.

3. But before we get to the future, we must acknowledge our past. In the darkest days of the Civil War, President Lincoln pleaded to our divided nation to remember that "We cannot escape history...[that] we will be remembered in spite of ourselves." One hundred and forty years later, that is still true.

4. For our nation, there is no denying the truth that slavery is a blight on our history. And that racism, despite all our progress, still exists.

5. For my party, there's no escaping the reality that the Party of Lincoln has not always carried the mantle of Lincoln.

6. Recognizing and confronting our history is important. Transcending our history is essential. We are not limited by what we have done, or what we have left undone. We are limited only by what we are willing to do.

7. Our nation must make a new commitment to equality and upward mobility for all our citizens.

8. This is a great moment of national prosperity. But many still live in prosperity's shadow. The same economy that is a miracle for millions is a mystery to millions as well.

9. From the beginning of this campaign, I have said that prosperity must have a purpose. The purpose of prosperity is to ensure that the American Dream touches every willing heart. We cannot afford to have an America segregated by class, by race or by aspiration. America must close the gap of hope between communities of prosperity and communities of poverty.

10. We have seen what happens when African-American citizens have the opportunity they have earned and the respect they deserve. Men and women once victimized by Jim Crow have risen to leadership in the halls of Congress. Professionals and entrepreneurs have built a successful, growing African-American middle class.

11. It must be our goal to expand this opportunity – to make it as broad and diverse as America itself. And this begins with enforcing our civil rights laws.

12. Discrimination is still a reality, even when it takes different forms. Instead of Jim Crow, there is racial redlining and profiling. Instead of "separate but equal," there is separate and forgotten. Strong civil rights enforcement will be a cornerstone of my administration.

13. I will confront another form of bias – the soft bigotry of low expectations in education.

14. Several months ago I visited Central High School in Little Rock, where African-Americans confronted injustice and white Americans confronted their conscience. In 43 years, we've come so far in opening the doors of our schools.

15. Yet today we have a challenge of our own: while all can enter our schools, many are not learning there. There is a tremendous gap of achievement between rich and poor, white and minority. This, too, leaves a divided society. And whatever the cause, the effect is discrimination.

16. My friend Phyllis Hunter, a teacher in Texas, calls reading "the new civil right." Equality in our country will remain a distant dream until every child, of every background, has a chance to learn and strive and rise in the world. No child in America should be segregated by low expectations… imprisoned by illiteracy… abandoned to frustration and the darkness of self-doubt.

17. And there is reason for optimism. A great movement of education reform has begun in this country, built on clear principles: Raise the bar of standards. Give schools the flexibility to meet them. Measure progress. Insist on results. Blow the whistle on failure. Provide parents with options to increase their influence. And don't leave any child behind.

18. I believe in these principles. I have seen them turn around troubled schools in my state. I've seen them bring hope into the lives of children – inspiring confidence and ambition. I'm especially proud that the performance of minority students in my state is improving at one of the fastest rates in the country. African-American fourth-graders in Texas have better math skills than any other state.

19. We can make the same kind of progress at the national level. A central part of my agenda is changing Title One to close the achievement gap. All students will be tested. Low-performing schools will have three years to produce results. If they do not, then these resources will go directly to the parents.

20. Every child can learn. Every child in this country deserves to grow in knowledge and character and ideals. Nothing is more important to our prosperity and goodness than cultivated minds and courageous hearts. As W. E. B. Du Bois said a century ago, "Either the United States will destroy ignorance, or ignorance will destroy the United States."

21. Education is the essential beginning – but we must go further. To create communities of promise, we must help people build the confidence and faith to achieve their own dreams. We must put

government squarely on the side of opportunity.

22. This is a higher and older tradition of my party. Lincoln argued that "every poor man should have a chance." He defended a "clear path for all." He financed colleges, welcomed immigrants, promoted railroads and economic development. Through the Homestead Act, he gave countless Americans a piece of land a start in life.

23. I have proposed a New Prosperity Initiative that reflects the spirit of Lincoln's reforms. A plan to remove obstacles on the road to the Middle Class. Instead of helping people cope with their need, we will help them move beyond it.

24. We must provide a Family Health Credit that covers 90 percent of the cost of a basic health policy for low-income families.

25. We must make it possible for more people to become homeowners, to own a part of the American Dream. So we'll allow low-income families to use up to a year's worth of Section 8 rental payments to make a down payment on their own home – then use five years of those payments to help with the mortgage.

26. We'll start an American Dream Down Payment Fund, matching individual savings for the down payment on a home.

27. Behind all these proposals is a simple belief: I believe in private property. I believe in private property so strongly, I want everyone to have some.

28. Education helps the young. Empowerment lifts the able. But there are those who need much more. Children without role models. Young people captured by gangs or addiction or despair.

29. Government can spend money, but it cannot put hope in someone's heart or a sense of purpose in their lives. This is done by caring communities – by churches, synagogues, mosques and charities that serve their neighbors because they love their God. Every day they prove that our worst problems are not hopeless or endless. Every day they perform miracles of renewal.

30. What we need is a new attitude that welcomes the transforming power of faith. In the words of a writer who visited the Mott Haven section of the Bronx: "the beautiful old stone church … is a gentle sanctuary from the terror of the streets outside."

31. In city after city, for the suffering and the hurting, the most hopeful passageway is the door to the house of God. We are going to extend the role and reach of charities and churches, synagogues and mosques, mentors and community healers, in our society. As President, I intend to rally these armies of compassion in the neighborhoods of America.

32. I will lift the regulations that hamper private and faith-based programs. I will involve them in after-school programs, maternity group homes, drug treatment, prison ministries. I have laid out specific

incentives to encourage an outpouring of giving in America.

33. Here's an example. More than a million children have one or both parents in prison. These are forgotten children – almost six times more likely to go to prison themselves. And they should not be punished for the sins of their fathers. We should give grants to ministries and mentoring programs that offer support to these children. Let us bring help and hope to these other innocent victims of crime.

34. I'm not calling for government to step back from its responsibilities, but to share them. We'll always need government to raise and distribute funds, monitor success and set standards. But we also need what no government can provide: the power of compassion and prayer and love.

35. These are some of my goals for America – to help make opportunity not only a hope and a promise, but a living reality.

36. The NAACP and the GOP have not always been allies. But recognizing our past and confronting the future with a common vision, I believe we can find common ground.

37. This will not be easy work. But a philosopher once advised: "When given a choice, prefer the hard." We will prefer the hard because only the hard will achieve the good. That is my commitment. That is our opportunity.

Appendix C

Multidimensional Bayesian Latent Variable Model

It may not be possible to string together DAEs into a single unidimensional measurement to which classical reliability calculations apply. We believe the DAEs within a subject will be at least mildly multidimensional; if we consider each DAE within a subject within a year as a measure of one proficiency, five DAEs would be measuring five different but substantively related proficiencies. These proficiencies are likely to be statistically related as well. For example, in NAEP, proficiency subscales within the same subject area are typically correlated 0.8 or higher, and seldom lower than 0.5 to 0.6. We can exploit these correlations by building a multidimensional Bayesian latent variable model to take advantage of proficiency estimates from old DAEs to help produce more precise proficiency estimates for the next DAE, or indeed to shorten the next DAE with no loss in measurement precision.

For example, suppose we wish to estimate a student's proficiency with a margin of error of 0.2 (SEM=0.1), and each item contributes roughly one unit of Fisher information to proficiency estimation (here we are borrowing an IRT formulation for specificity); then the student would need to answer roughly 100 items. However, if we could already predict the proficiency on this DAE with a margin of error of 0.4 using past DAE performance, we would need roughly only 20 more items to obtain a margin of error of 0.2 on this DAE. (The numbers are chosen here mostly for computational convenience and may not reflect the actual values obtained from item precalibration and so on.)

This calculation depends on the student's performance on the new DAE being consistent, in a way that can be made precise using Bayesian modeling, with his or her performance on past DAEs. If the student's responses on the next DAE are inconsistent with older DAE results, we would need to do follow-up testing to get a more precise estimate of the student's proficiency. Thus, for students who learn consistently from one unit to the next, we can exploit past performance to help estimate proficiency on the current unit of instruction. However, for the student who performs unusually well (or poorly) on the current unit, we can use the Bayesian machinery to see the inconsistency, and offer another block of items in order to more precisely assess that student's learning. A similar process is used in online tutoring systems and adaptive testing systems and is an illustration of the kind of useful customization that is discussed in the *Adaptive Mass Personalization* section of this paper.