THE NATIONAL CONFERENCE
ON
NEXT GENERATION ASSESSMENT SYSTEMS

An Assessment System for the United States: Why Not Build on the Best?

Marc Tucker

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 Assessment System for the United States: Why Not Build on the Best?

Marc Tucker

National Center on Education and the Economy

Part 1: Design of the Proposed System

Though no one would have predicted it 5 years ago, the country appears to be headed toward agreement on voluntary national standards for its elementary and secondary schools, at least for English and mathematics literacy. The U.S. Department of Education is pressing hard for the release of those standards to be followed soon after by tests aligned with those standards. It has set aside $350 million, to go to coalitions of states prepared to develop and implement those tests. It has also made clear its intention to require the states to use the data produced by those tests as important factors in the decisions made about teacher and principal tenure, promotion, compensation, and retention. No one doubts that the implementation of these standards and tests will have a profound effect on instruction and on the professional school workforce in the United States. The question this paper addresses is: What should a testing system for the United States look like?

Our Perspective

For 22 years, the National Center on Education and the Economy (NCEE) has been benchmarking the education systems of those countries that have consistently demonstrated the best performance in international comparisons of student achievement. Educational standards, instructional systems, and assessment have been a particular focus of our work. *Benchmarking* is a term borrowed from private industry. Its origin lies in the late 1970s, when American business was severely challenged by the Japanese. In response, the best American firms identified their most formidable competitors, carefully researched their practices, borrowed the best practices from each of the best performing firms, added some ideas of their own, and implemented their own version of the resulting product and process ideas. They did not seek to reinvent the wheel. Wherever they found a better product or process, they adopted it.

Many people are under the illusion that they can understand another nation's education system after a brief benchmarking visit. National education systems are particularly complex and constantly changing. They are easily misunderstood, and brief visits will inevitably produce portraits that are misleading.

Over the last 22 years, my staff and I have visited 23 countries, including all of the countries that consistently place at the top of the world's league tables for academic achievement. We have visited some of them many times, often for weeks at a time. The recommendations made in this paper are largely based on that research, as well as the research of others.

The Design of This Paper

In the next section, I lay out the goals that I think a system of testing should meet. The following section describes a system of testing at the high school level that would meet these goals. That section is divided into two subsections. The first proposes a way to get started, by taking advantage of the best of the world's existing examination systems. The second subsection on high school testing describes a program of investments that might be made to further improve, over time, the already very good examination systems that are available for use in the United States today. The next section describes a strategy for developing a greatly improved K-8 testing system, which is then followed by a description of the ways in which the new testing systems could be used to drive our accountability systems. Finally, in the last section, I describe a structure that could be used to govern and regulate a voluntary assessment system involving many, perhaps most, states.

What We Should Want From Our Testing System

The answer, of course, depends mainly on our goals for it. I believe that we should want:

Assessments that, first and foremost, advance student learning. Measuring student learning accurately is essential, but it is not enough. The assessment system itself must make it clear to both student and teacher what is to be learned. It must signal what kind of student work will meet the standard. It must enable teachers to figure out what the students are learning and what they are having trouble with in real time, so the teacher can correct course. It must do the same for the student. The assessment system itself must be constructed so that it models and rewards good teaching and discourages the kind of teaching that enables students to get the right answer without really understanding the material. To that end, among other things, is must model the kind of teaching and encourage the kind of learning that is needed if the students are to succeed. Put that way, then….

Assessments that are conceived as part of a highly integrated instructional or learning system. This is the single most important finding from our 22 years of research. It is borne out by years of research by John Bishop and by a brilliant paper by two German researchers, Fuchs and Woessman (2007), who used the data from the Programme for International Student Assessment (PISA) assessment program to identify the most important factors accounting for the success of those countries with the best student achievement at scale. Two factors account for a very large proportion of the variance: teacher quality and the presence or absence of a high quality national or state instructional system. Such systems consist of a core program of required subjects at the high school level. Each course in that program comes with a thoughtful, well-constructed syllabus. Each syllabus is accompanied by instructional materials matched to the course description in the syllabus, as well as a high quality examination that is directly derived from the syllabus. These systems are almost always accompanied by training for the teachers that is also matched to the syllabus and designed to enable them to teach the material successfully to students from many different backgrounds. This finding is actually plain common sense. When all students get a powerful instructional program that teaches what they have been told they have to learn, when the instructional materials are designed to support that learning program, when what is taught is what is measured, and when teachers are taught to teach well what the students are

supposed to learn, then students learn much better than when these things are not true. This leads directly to the next observation….

Assessments that are curriculum-based as well as standards-based. Even detailed content and performance standards are not very clear about what students need to learn and what student work looks like that meets the standard. That is not because they are poorly written. It is because abstract statements about what sort of writing is expected or what good reasoning and analysis looks like in science cannot be conveyed very clearly by statements of the form *students should know this and be able to do that*. The mysterious becomes clear, however, when students are told: here is what the course is going to be about, here is what you need to read, these are the papers you must write, and here is how your work is going to be judged—the sort of thing one finds in any good syllabus. It is clearer still if the student can be shown the kinds of questions that were asked in prior years and examples of work that got good grades, which leads to the next point….

Assessments that are fair, because students have an opportunity to learn the material on which they will be examined. So here is student A, who is taking an exam that was based on the syllabus of the course that student just took. And here is student B, who is taking the same exam, but her teachers got only a page of standards saying "the student should know this and be able to do that." The second teacher did her best to cobble together a course from some ideas she got on the Internet from the state department of education, but it is no match for the thoughtfully designed course that the first teacher was using, nor were the materials that were listed on the Internet anywhere near as well aligned to the course as were the materials that were actually designed to support it. When two students take the same exam, and both experienced the same course, and both had teachers who were well-trained to teach it, then the difference in scores will be explained mostly by the difference in effort made by the two students. But, when one took a course on which the exam was based and the other did not, the student whose course did not serve as the driver of the test is at an enormous disadvantage. That is not fair. And it will not produce the kind of improvement in student performance that all of us want.

Assessments that promote and measure critical thinking, strong analysis, and real imagination and creativity. There is no country with a consistent record of superior education performance that embraces multiple-choice, machine-scored tests to a degree remotely approaching our national obsession with this testing methodology. That is not because they are backward. It is because they are interested in measuring the acquisition of not just basic skills, but also analytical skills, critical thinking, creativity and imagination. They recognize that the only way to find out if a student can write a competent 20-page history research paper is to ask that student to write one, and the only way to find out if a student has the knowledge and skill needed to design and build a robot to certain specifications is to ask the student to build one. Multiple-choice tests will never be able to find out if the student has a better or more original answer than any of the answers imagined by the test constructors. Machine-scored tests are cheap, constitute a very efficient and accurate way to measure the acquisition of most basic skills and can produce almost instant results, all important advantages, but they have a way to go before they will give either e.e. cummings or James Joyce a good grade, or recognize a competent robot or good painting when they see one. We need to find a way to measure what is important to measure,

rather than continue to do what we have been doing, which is confine our curriculum to the things that are cheap to measure.

Assessments that are balanced, in the sense that they can support instruction as it is taking place, by providing immediate, targeted feedback on student performance as well as provide summative information to a variety of audiences, when the student has completed a course or a program of study. There is broad agreement that teachers should have available to them assessments that they can use to find out, as they are teaching a course, whether their students are learning what they are teaching, so that they can take appropriate action if that is not the case. And there is also increasing recognition that it would be highly desirable to be able to have summative assessments that could include, along with the results of performance on a timed test, assessments of major pieces of student work of a kind that cannot possibly be done during a timed test. Some people have realized that the dividing line between these two ideas is very thin indeed. That is, the grades on major pieces of work that are done during the year, which could be incorporated in the final course grade, could also be used as indicators of whether the students are learning what the teacher is trying to teach.

In a sense, these developments constitute not so much a revolution in thinking in assessment as a counterrevolution. Teachers have always given frequent quizzes to find out where their students are as the semester is progressing, so that they could correct course if necessary, and have always given major assignments during the course, which they have graded, and those grades have been typically given some weight in the final course grade, along with the final exam. But none of this has been included in the externally administered tests administered by our districts and states that are now used for accountability purposes. What appears to be called for now is a restitution of the sensible balance found in ordinary classrooms prior to the general focus (for accountability purposes) on the timed test at the end of the process. Thus we return to a more holistic testing regime, one that uses a variety of assessment methods, and in which different kinds of assessments are used both during and at the end of the course, both to focus the teaching on what the students most need and to be able to find out what they have learned by the time the course is over.

Assessments that are both valid and reliable. The United States, by comparison with those countries that are the world's best performers, has emphasized reliability at the expense of validity. In less technical terms, we appear to have valued testing systems in which different students answering a given question in the same way will receive the same scores much more than we have valued tests that accurately measure what we want them to measure. The countries with the highest performance value both of these goals, but they recognize that one gets one at the expense of the other, and so are more interested in establishing a better balance between the two. In any case, it is hardly clear that this country has actually achieved the high degree of reliability that it believes it has achieved. Which one of us has not had the experience of taking a multiple-choice, machine-scored test and, knowing that one could actually reasonably interpret the question and the possible answers provided in more than one way, found ourselves trying to figure out which of the possible right answers was the one that the test constructors had in mind? It is very likely that we get less reliability from our tests than we think we do,

and have exchanged it for even less validity than we think, because our tests, when compared to those in wide use overseas, emphasize the measurement of basic skills at the expense of more advanced skills.

Americans often observe that many foreign countries depend on teacher-scored examinations, with all that implies about low reliability of the scores obtained in that way. Those national systems that rely heavily on grades and scores provided by classroom teachers are typically low- or no-stakes systems, for both the students and the teachers. In those countries that use their national tests and examinations for high-stakes purposes, either for students or teachers, or both, the examinations are given under high-security conditions, and the exams are professionally scored. Work done by the students during the year and scored by their teachers is sampled and rescored, if necessary, by trained scorers. Often, the organizations that score them employ school teachers to do the scoring; these teachers are professionally trained to do the scoring and are monitored carefully for reliability. If they cannot score reliably, they are dismissed and the exams they have scored are rescored by someone who can score reliably.

Assessments that cannot be test-prepped. Teachers in the United States, operating under conditions of high-stakes accountability in recent years, have strong incentives to prepare students to recognize questions just like those they have drilled on, and to employ rote procedures to provide the answers. This can be done successfully even if the students do not actually understand the material. The process is known as *test prep*, and it is insidious, because it encourages a form of teaching to the test that undermines good teaching and deprives students of the understanding of the material on which their further education depends. Good assessment makes this kind of teaching unrewarding, because it gives students questions that do not look like the ones they have practiced on, but which call for the kind of understanding and skill that they should have mastered. Students cannot succeed on assessments of this kind unless they actually have thought about and understood the material they have studied and can apply it to a wide variety of problems calling for that skill and understanding.

Assessment standards that the states cannot fiddle with. Both the states and their critics are demanding that the performance standards embedded in whatever assessment system is developed be comparable across assessments and across states, so that the individual states cannot game the accountability system by lowering their standards. Thus the national system of testing will have to incorporate a method of setting national content and performance standards that no state can fiddle with.

Assessments that can be used to motivate students to take tough courses and study hard in school. It should come as no surprise to us that secondary school students in the best-performing countries are much more inclined to take tough courses and to study hard in school than their counterparts in the United States. That is because they have strong incentives that are missing here. They cannot go on to the next stage of their education or embark on a rewarding career unless they have earned the appropriate qualification to do so. A *qualification* is a piece of paper attesting that the holder has earned the requisite grades in the requisite courses needed to go on. It is hard to believe that the performance of students in the United States will ever equal the performance of students in the best-performing countries unless we engage their energy in the same way that these other countries have engaged the

energy and commitment of their students. That will require us to develop assessments that can be used to create our own qualification system. More on this in a moment.

Assessments which, when combined with the accountability system, do not produce a very narrow curriculum. No Child Left Behind (NCLB) was designed primarily for the purpose of improving the performance of low-income, minority children. It focused on their most glaring problem: poor performance in basic literacy. That may have been appropriate, but the operation of this system, in combination with the design of many state accountability systems, has been to narrow the curriculum to the testing of English and mathematic literacy (and to some degree, secondary school science). No one would regard this as a complete core curriculum. A new national assessment system must recognize that what gets measured gets taught, and provide incentives to the states and schools to make sure that the whole core curriculum, not just this very limited set of subjects, is taught.

Assessments that make reasonable accommodations for the disabled. This is the right thing to do, and, besides, it is the law. The U.S. Department of Education has led an extensive and rewarding effort to fund innovation—especially the application of advanced technologies—in this arena in the states. Results from these studies should be taken into account in the design and implementation of new assessment systems.

Assessments the country can afford. Here we find what appears at first glance to be an irresolvable dilemma. The superior assessments that the best performing countries are using are expensive, costing perhaps 2 to 3 times what our typical state accountability tests cost to administer. And school finance in the United States is in a shambles. Even if we could get widespread agreement that we would get much better student performance if we adopted assessments of the sort being used in the best performing countries, many people think we could not afford it. I think we can. Our proposal to resolve this dilemma can be found below.

This is a formidable set of requirements. But every one of them can be met, including having a system that is both state-of-the-art and affordable. Here's how…

A High School Assessment System—Part One: Getting Started

Some have proposed that the United States design and build an assessment system with most of the characteristics just described, from scratch. To do that would take many years and cost an enormous sum of money. If that were the only way to get the job done, then so be it. But it is not necessary.

Let's start now by using the world's best examinations and instructional systems. As I noted above, the highest performing countries have had systems of this sort for many years. Many of them are first-rate, themselves the result of many years of work and sizeable national investments. The organizations that produce them have vast experience doing this sort of work. Not all of them, of course, are written in English and available for use outside the country in which they were developed. But some are.

So we have done the obvious. We used our own research and have checked with other knowledgeable researchers to identify the best such assessment systems available in English for use in the United States. We were not looking for tests. We were looking for what the best research says we should be

looking for: complete standards- and curriculum-based instructional systems that include very high quality examinations. Nor were we seeking to identify the very best such system. We discovered years ago that the British (more precisely in this case the education authorities in England, Wales, and Northern Ireland) have a system that enables them to offer several different complete instructional systems to their high schools. All are set to the same standards. That struck us as the perfect solution for the United States. We need a system that is curriculum-based, but no one wants a single high school curriculum imposed by anyone on the states, and few states want a curriculum imposed by the states on all its schools. Why not offer a choice of curricula, as long as we can be sure that an A is and A is an A, no matter which curriculum is chosen? The British have perfected their system of making sure that the standards are the same across programs (a process called *moderation*) over many years. There is no reason to believe that that process or something very like it would not work here.

So we are proposing that the states adopt these outstanding instructional programs, with their associated examinations, and offer their high schools a choice of them.

But do the world's best instructional systems meet the standards we just laid out? Well, actually, they do. They are curriculum-based. They are complete instructional systems. They provide syllabi, associated instructional materials, matching exams, professional scoring, and teacher training. They address not just basic skills, but the full range of desired knowledge and skill, including critical thinking, complex analytical skill, imagination and creativity, and the ability to apply what one knows to real world problems. They employ a range of assessment techniques. Some incorporate the assessment of important performances that cannot be included in a timed test. They include tools for formative assessment as well as summative assessment. They provide strong instructional support to both teachers and students. They model the kind of instruction that teachers should use if they want their students to perform well on the exams. They are much more valid for assessing higher order skills and knowledge than the typical American test, and they satisfy the reliability standard in the countries in which they are used all over the world. They are vastly better on virtually all of these dimensions than any state accountability test now in use, and they are available today. The syllabi that come with these programs describe what the student is supposed to do, in the same way that American-style standards do, but they also publish the prior year exams and examples of the student work that got high scores, so the standards have a concreteness for students and teachers of a kind that we rarely see in the United States.

How we would use these instructional systems and examinations. But there is more to it than that. I mentioned above that other countries motivate their students to take tough courses and to study hard by offering them qualifications that they can use to get on with their lives. These standards-based certificates are earned by students, who have to achieve a certain grade for certain courses in order to get them. They can't get these qualifications just for showing up, in the way that our high school students get their diplomas. They have to perform at a certain level.

There is another idea that is important that we have borrowed from the countries that we have studied. In most of them, a sharp distinction is made between what we think of as the lower division of high school (the freshman and sophomore years) and the upper division (the junior and senior years). Most students are expected to complete a common curriculum by the time they are 16. Then, when they have

mastered it, they can go in different directions. In math, then, they will be expected to have studied the same topics until they are 16 (and can do so at a higher challenge level if they choose to do so), but the math they study after they have mastered lower division math depends on what they want to do with their lives. Thus, for example, those students who want to go on to the science, technology, engineering, and mathematics (STEM) professions would study the topics in Algebra II in their upper division program, because those topics forge a path to calculus that is critical to the STEM professions. But students with other interests might take courses in statistics and data analysis. That is precisely the system, for example, in Singapore, which may have the most admired mathematics curriculum and examinations in the world.

Let's put these ideas together and see how they might be combined into a structure that would fit the American scene.

The move-on-when-ready system. We have identified three first-class comprehensive instructional systems that can be used at the lower division level. These are the QualityCore program from ACT, the International General Certificate of Secondary Education program from the University of Cambridge International Examinations, and the International General Certificate of Secondary Education Program from Pearson/Edexcel. We are proposing that states tell their high schools that they must adopt at least one of these programs for their lower division students.

In most of the countries we have studied, the exams on which the qualifications are based produce a range of scores or grades. This range is used to sort students out, so that students' choices after the age of 16 are constrained by their performance on these exams. But, we realized, these courses and exams do not have to be used that way. They can be used to expand opportunity, not constrict it.

Here's how.

All of these programs would be set to a single pass point (through a mechanism I will describe below). That pass point would be based on empirical data on the actual requirements of the initial credit-bearing courses in the nation's open-admissions 2-year and 4-year colleges. These programs would include, however, not just courses in English literacy and mathematics, but also courses in the sciences and technology, literature, history and civics, and the arts.

All of these examinations for these courses would be available to the students by the end of their sophomore year in high school. If they demonstrate the requisite level of math and English literacy, and get satisfactory grades in the courses in the core curriculum, the students would be awarded a high school diploma, and would be able to enroll immediately in a public open-admissions 2-year or 4-year college, where they could take either a college transfer program or a terminal 2-year degree or certificate program, without having to first take any remedial programs, because they would have shown that they do not need remediation.

Alternatively, if they pass their examinations, they could stay in high school and go on to enroll in an approved upper division program of study designed to help them gain admission to a selective college. The programs we have identified for this purpose include the International Baccalaureate (IB) Diploma Programme, a program made up of selected advanced placement courses, the University of Cambridge

Advanced International Certificate of Education (AICE) program, the Pearson/Edexcel "A Level" program, and the ACT QualityCore program.

If the students fail to pass their examinations at the end of the sophomore year, their high school would be obligated to prepare a customized program of studies addressed to the areas in which the exams showed them to be weak. They could retake the exams at the end of their junior year, or, if necessary, at the end of their senior year. But no student would ever definitively fail their exams. They could continue to challenge the exams for the rest of their lives. The idea would be to use these programs to screen students in, not out—to get as many students as possible to the college-ready standard whenever they were ready.

A performance-based continuous improvement system—available now. What I am describing is a performance-based system of education. The exams would not be 10th grade exams. They would be exams that could be used to demonstrate college readiness that could be taken whenever the student feels ready to take them. And, once they are passed, the student could then move on.

If the experience of other countries is any guide, this system would be immensely motivating to students. They would know exactly what they need to do to get on with their lives and they would do it, in very large numbers. The widespread boredom and waste of everyone's time that now characterizes many of our high schools would disappear. Our high school teachers would gain something they have never had: a motivated student body ready, as high school students in other countries, to take tough courses and to study hard.

Though these courses and examinations are more expensive (and higher quality) than those now in use, the system I have just described would, in just a few years, cost no more than the current system. As more and more students left high school before the end of their senior year, the costs of those classrooms would disappear from the high school budget. The funds released could be used to provide more and more support to the students who have a hard time reaching the college-ready standard, in the form of before-school, after-school, weekend, and summer programs. High schools could offer more tutoring. They could build more carefully tailored programs for the students who do not succeed on their exams the first or second time. Students who come to high school well prepared could sprint ahead. It would all cost about the same amount of money, but that money would be much better spent.

What I have just described is a system in which assessment would be used as the basis of a continuous improvement system focused on student learning.

All of this could be accomplished with relatively small changes to these world-class instructional systems. All of the organizations that provide the lower division programs have agreed to modify their offerings to the extent necessary to comply with the Common Core Standards now being developed under the auspices of the Council of Chief State Officers and the National Governors' Association. The British organizations have indicated that they are willing to make the changes in spelling, diction, and reading lists necessary to adapt their offerings for use in the United States. In fact, the Cambridge system is now in use in high schools scattered across the nation, with a large concentration in Florida, where state incentives have encouraged its use.

With these modifications, the United States could adopt the world's best instructional systems and use the world's best assessments at a small fraction of what it would cost to develop them from scratch and have them available for use in our schools in far less time.

On being both college-ready and work-ready. One of the strong impediments to the development of a system for making students college-ready and work-ready is the fiction that there is one standard for college-ready. This plan clearly distinguishes between what it means to be ready for open-admissions colleges and highly selective colleges. Acknowledging that difference makes it possible to develop a system that will get students ready for one or the other or both, as this one does.

But the difference between being ready to begin a career as a nurse's aide and a career as a medical doctor is no less great than the difference between the local community college and the Ivy League or the Big Ten. In this plan, being work-ready means having a 2-year degree that is recognized by an industry as qualifying one for entry level work in an occupation that pays enough to enable one to support a family above the poverty line.

One of the unintended effects of the standards movement has been to drive serious technical and career education out of the high school curriculum. It has been able to survive only by providing curricular alternatives that profess to use technical and career education as a vehicle for motivating certain students to acquire academic skills.

The reality is that few high schools can afford the specialized teaching personnel or the equipment to provide high quality technical and career education. Those resources are typically available only in specialized regional high schools or, more broadly, in 2-year community and technical colleges. The proposals made here would enable many students who want those educational opportunities to leave high school 2 years earlier than they can now and prepare them to succeed in serious technical and career programs, rather than fail in large numbers because they lack the literacy skills to succeed in those programs. They could leave these 2-year institutions at the age of 18 with a high school diploma and a 2-year degree or certificate, fully prepared to embark on a rewarding career, a very different prospect than most such students now face.

A High School Assessment System—Part Two: Improving on the Best the World Now Has to Offer

Saying that the instructional systems we have identified are among the best in the world is not to say that they are everything they could be. Once these standard-setting instructional systems are in place, there are many possible improvements that might be made in them, most having to do with the application of advanced technology to the assessments and the curriculum. Much has been written on this subject, so I will only illustrate the point here, first for assessment, then for curriculum and instruction.

Advance the state of the art in assessment technology. Computer-based systems employing simulations and other forms of dynamic modeling make it possible to pose test and examination problems for students to solve that require them to demonstrate their understanding of the way the variables in all

kinds of systems interact to produce a wide variety of results. These systems might include everything from biological systems to mechanical systems, from economic systems to social systems. In the course of responding to problems of this sort, students could demonstrate their capacity to frame the relevant problems in mathematical terms and to solve these multi-step problems. They could also demonstrate their understanding of the relevant scientific principles and their application in specific situations.

With such dynamic models, students could be asked to change the value of the variables and to explain why the changes in those values produce certain observed results, thereby enabling the assessor to judge whether the students really understands the processes involved, the conceptual underpinning of the relevant phenomena, and the ways in which cause and effect are connected. As part of their assessments, students could be given access to data that could not possibly be included in a conventional timed testing environment and asked to access and use that data to address the question posed in the assessment. Art students could be asked to respond to various works of art with narrative commentary, or to create new works of art or alter images presented to the student, using the painting and drawing tools in the computer software. The same could be done with musical composition and performance and graphic design. Technology students could be asked to manipulate simulated systems to improve their efficiency, output, or effectiveness.

Expert systems could be used to engage students in problem-solving exercises in which the student engages in dialogue with the computer, solving complex, multi-step problems in continuous interaction with the computer. Similar systems could go beyond their current rule-bound limitations to grade original essays that are now beyond their capacity to score reliably.

In all these and other ways, computer-based systems could overcome many of the current limitations of multiple-choice, computer-scored tests, and go way beyond them, enabling the grading of important kinds of student performances that cannot now be reliably scored even by human scorers. If that could be achieved, all examinations could result in scores or grades that were nearly instant, and the range of learning and knowledge that could be assessed would be greatly increased. It is even possible that these computer-based assessments could be administered and scored at lower cost than at present, once the initial research and development costs are amortized.

I believe that this sort of research and development should be done through close partnership between the organizations now producing the world's best high-performance learning systems and assessments and technologists who are working at the leading edge of the available technologies. Though it will be essential to include people whose primary expertise is in psychometrics, I do not think they should lead this work. Their role should be to assist the curriculum people and the technologists as this creative partnership takes shape.

Cost is always a consideration in assessment, and the returns to scale that can be obtained from computer-based testing vastly exceed any returns to scale that can be obtained from the various approaches to human scoring. So is speed of response, another arena in which computer-based testing wins hands down. And computer-based testing is often handed the baton on reliability as well, though the situation there is less clear-cut. I argued above that these considerations need to be weighed against

the overriding need to assess the full range of knowledge and skill that is important to us. It is not clear at all whether technology-delivered assessment will ever be able to assess the full range of outcomes that is important to us. But it is worth pushing as hard as possible on the boundaries. In my view, responsible policy will pursue this option by supporting the necessary research and development, acknowledging the risks involved, rather than bet the ranch on outcomes that may or may not be achieved.

Advance the state of the art in instructional technology. All of the possibilities just described could and perhaps should produce a world in which virtually all summative assessment could be embedded in the curriculum itself. The circumstances in which summative assessment would have to take place would have to be controlled so that the assessors could be sure that the work was the work of the person being assessed, but the process of assessment would be nearly indistinguishable from the process of instruction. As students tackled ever more sophisticated problems and solved them in these technology-rich environments, their progress would be recorded by the very software that provided the learning environment. Pushed to its logical conclusion, these could be envisioned as environments in which students of any age or background could interact with machines in virtually any setting, alone or with others. In this imagined environment, the student is interacting through a portal with an infinitely capacious education system.

The Australian government is now engaged in the development of a new national instructional system of the kind proposed in this paper. The vision animating its principal designer is the possibility of creating a complete instructional system capable of delivery in all its aspects by computer-based technologies, a vision that lies on the road to the vision I just described. Students anywhere in Australia within reach of the Internet would be able to access the program descriptions, course syllabi, instructional materials, and assessments by computer and all scoring would be done by the computer. The instructional materials would include myriad links to resources that would themselves include myriad links to yet other materials and resources, from dictionaries to thesauruses, from articles to paintings, from tabular data to historical references, from live authorities to ancient manuscripts, all arranged in such a way that the student could progress from an initial overview of the topic at hand to an ever deeper understanding of the underlying conceptual structure and specialized knowledge. The entire system would be interactive. Dynamic models of complex phenomena would be available to the students, enabling them to explore the relationships of the variables and see the consequences of changes in their values. The line between assessment and learning would grow very thin. Instruction would be liberated from schools and classrooms and could take place anywhere where there was an Internet connection. Credentials could be earned by anyone at any age, when the students were ready.

It is not clear how far the Australians will get, but the vision is very exciting. It will not be achieved by a date certain, and may not ever be fully achieved. But this vision has the potential to transform instruction and assessment in ways that could greatly advance learning, personalize it, and make it far more intrinsically interesting and engaging than it has ever been before.

The key to success here lies neither in technology nor in assessment. It lies in curriculum, in a clear and powerful conception of what is worth learning and how to create an environment in which powerful

instruction can take place to increase the probability that the student will learn it. The right sort of technology and assessment can enable that vision but cannot produce it. It is essential that these systems be curriculum-driven.

A K-8 Assessment System

Here, I believe, two contending visions are worth considering. One dominates the scene in the best-performing countries. The other dominates the scene in the United States.

Low-stakes and high-stakes assessment system models in high-performing countries. In those countries with the best records of student achievement in the world, there are high stakes for the students in their high schools that are connected with their performance on their qualifications examinations.

In the lower schools in those countries, the stakes for the students are either low or nonexistent, except in those countries that use qualifications exams for students in their mid-teens that determine which high schools they can be admitted to, not, of course, a policy we would advocate.

In almost all of these systems, there are no stakes or very low stakes for the teachers that are connected with student performance on the assessments. It is not the case that the schools are not accountable for their performance, but, in many of these countries, the student assessment systems, if they play any part at all, are used to send a signal to government that triggers a visit by school inspectors, who then make a definitive finding with respect to the course of action that a school found deficient must take. The achievement scores of students are used to signal the possibility that there might be a problem worthy of the attention of the inspectors, rather than as the basis of adverse action directly.

In most of these countries, there may be national tests or exams for K-8, but they typically come at the end of grade bands, not individual grades. Because there are no stakes or low stakes, they are mainly scored by the students' teachers. Sometimes those scores are moderated by the scores of other teachers in other schools. Generally, teacher scoring in these countries in the lower schools is regarded as an important part of teacher professional development, because it focuses their attention on the standards, on being able to recognize work that meets the standards and on talking with other teachers about how they enable students to reach the standards.

These countries' assessment methodologies sometimes include multiple-choice, computer-scored assessment, but, when they do, that methodology accounts for a fairly small proportion of their overall testing regime. Most assessment is done with prompts calling for essay-type responses and are scored, as I have noted above, by the students' own teachers.

High-stakes testing in the United States. The United States stands nearly alone among the advanced industrial countries in its approach to high-stakes school assessment. Whereas most of the rest of the world has been interested in examining students to see to what extent each has mastered a particular curriculum, Americans developed tests that were supposed to be insensitive to the curriculum the student had taken. Multiple-choice, computer-scored testing systems were developed for reporting student progress on the acquisition of basic skills, but these tests made very little difference in the lives of college-bound students, who were expected to take examinations built on the European model (like

the advanced placement tests) or tests of general intelligence (which were actually tests of the kind of high literacy and general knowledge that these students possessed in abundance compared to less well-off students in non–college-bound tracks).

Virtually all of the basic skills tests and the general intelligence tests were computer-scored, and mostly multiple-choice. It was natural that, when the accountability movement began a few years ago, policy makers and testing directors turned to the companies that had long serviced the school testing market to build similar tests. Their virtues—reporting speed, reliability, and low cost of administration—are real and important. But Americans have gotten used to tests that cost one third to one quarter what the tests used by the highest performing countries use. They have gotten used to tests that most accurately measure the basic skills and have tended to reserve examinations of higher order skills for our elites, not for use in the mass education system. They have gotten used to tests that are unrelated to the curriculum taught in the schools (which is, of course, why American teachers, almost alone in the world, hate teaching to the test). They have gotten used to very rapid turnaround in test reporting. And they have gotten used to using their tests for high-stakes purposes, which they believe rules out involving teachers substantially in grading and scoring student work when it really counts.

The accountability movement greatly increased the use of American-style tests for high-stakes purposes, a development that was greatly accelerated in the George W. Bush administration with the passage of the NCLB legislation, by requiring the grade-by-grade use of these tests by the states in mathematics and English literacy and additional tests in science and in high school. It now appears that the Obama administration may wish to greatly expand the use of these computer-scored, multiple-choice tests to other subjects and other grade levels in order to implement a system that would base teachers' tenure, promotion, compensation, and retention on the growth in the performance of their students as measured by these tests or similar tests.

While the administration has offered to pay up to $350 million to create these tests, it has not offered to pay for their continued administration. It has indicated its preference for the use of computer-adaptive testing, which will substantially increase the number of test items required. If the tests are released each year, there will be a very large requirement for new tested items each year, far larger than is currently the case, because more students will be tested in more subjects at more grade levels using more test items per student.

Given that current projections show that school and state budgets are likely to be under pressure for many years to come, this could further decrease the quality of the tests this country uses. It is certainly not a formula for increasing the quality of our tests, which is what I think is needed.

The Board of Testing and Assessment of the National Academies has raised serious questions about the use of data from currently available tests or other tests similar to them for the development of indicators of the growth in student achievement attributable to the work of individual teachers.

It seems to me that the fundamental decision that must be made is whether to adopt the pattern that is prevalent in most of the high-performing countries or to pursue the course this country has been on since the advent of the accountability movement.

Because the countries we compete with have systems that are high stakes for students only in high school and have only low to no stakes for teachers at any time, they can afford to develop and administer very high quality assessment systems tied to their standards and curricula in high school and to have a system of assessment deliberately designed to support student learning and teacher professional development in their K-8 systems. Because the national or state assessments they use at the K-8 level do not have to be high stakes assessments, the costs of these assessments can be relatively low and still be of higher quality than the ones we use, because they are teacher-scored, and that scoring is typically part of the teacher's job.

The result is that these other countries may be spending more or less what we are spending, but they are using their money very differently. They are getting much better quality testing, meaning that they are able to test a much wider range of desired outcomes more accurately than we, without spending any more money. The difference is the stakes attached and the testing methodologies used.

In my opinion, these other countries have got an advantage on us. There is no empirical evidence from any source that I am aware of that American-style testing and accountability systems produce better results than the systems in the countries NCEE has been studying for more than 2 decades. The contrary is true. The general performance of these other countries and the specific studies I referenced at the beginning of this paper strongly suggest that the high-performance instructional systems they use, which include their testing and examination systems, are major contributors to their superior performance.

So it will come as no surprise that I believe that the next generation of K-8 testing in this country should emulate the design of the best systems in Europe and Asia, rather than continuing to pursue a low-cost testing regime that is alienating teachers and parents and, in some cases, students as well. While technology can no doubt offer some efficiencies, I think we would be better off employing it in the service of creating high-concept, high-fidelity, high–cognitive-demand assessments, thus lowering their cost and concurrently improving their contribution to the operation of a robust instructional system.

Briefly, this would mean dividing the grades into grade spans and developing summative assessments for the end of each band. I would begin with an assessment to be used at the beginning of kindergarten, by kindergarten teachers. The purpose would be to give kindergarten teachers a clear picture of the arenas in which their new students are strong and weak and, in particular, to identify problems, especially in literacy, that need priority attention. That data would also go to the state, to enable the state to make policy for early childhood education that would take into account the actual profiles of the children entering the formal school system at the time the assessments are given. These children cannot be given a formal test of the usual sort, so the assessments would have to be appropriate to the age of the students, but I would want, at a minimum, a measure of their vocabulary.

Beyond that, summative assessments would be provided at the end of the third, fifth, and eighth grades. Third is the end of primary school and the age at which students should have mastered the foundation skills in English and arithmetic. It is essential that schools, teachers, parents, and the state know where students are at the end of that crucial year, so that they have a chance to correct major problems before

they get worse, whether at the individual student or school level, or both. Fifth grade is the end of elementary school and eighth is the end of middle school, and both are appropriate marking points to take the temperature of the students and the school.

All of these assessments would be set to a progression of standards in the core subjects in the curriculum defined by the most recent knowledge we have concerning the desired progression of learning in these disciplines. By this I mean what is now being learned about how children actually learn these subjects when the topic is logically organized so that each topic is presented as the next logical building block in a progression toward a defined end. In this case, the end in view would be the knowledge and skills required to succeed in the freshman year of any of the high-performance instructional systems named above in the section on high school assessment. In the first instance, this progression would be the one embedded in the Common Core Standards now being formulated under the auspices of the Council of Chief State School Officers and the National Governors' Association.

The summative assessments would include computer-scored, multiple-choice questions, but would also include constructed responses, more frequently used and of greater length than is typically found in American state accountability tests. They would be administered under secure conditions and scored professionally. And they would be designed to provide levels of reliability and validity acceptable in the United States. To the extent feasible, they would make use of the more advanced forms of technological support in delivering and scoring the assessment.

But these timed assessments would be accompanied by test items and prompts that could be used by teachers throughout the year, some for formative purposes, to enable the teacher to find out at moments of his or her own choosing, how the students were doing relative to the standards, and to produce pieces of student work produced during the year, the grades on which could be taken into account when deciding on the final grade for the course. Just as in the British system, these pieces of student work would be of the sort that cannot be accommodated in a timed assessment administered at the end of the course.

The system would also produce such items, keyed to the standards, and available to teachers whenever they wished to use them, in all the off years in which there was no summative assessment. The system would be designed so that teachers would be able to assemble these assessment tasks into mini-tests at any time, and to use them to compare the progress of their students to what the research shows is the progress to be expected of a typical or modal student at any particular point in time.

Just as in the systems whose design I have been describing, this K-8 system would be keyed to a scaffolded curriculum, one that is not confining, but which gives teachers a structure on which to build their own curriculum and gives the publishers of instructional materials a very good idea of what needs to be in their materials at each grade level in each core subject in the curriculum. If this were done, as it is almost everywhere else in the world, texts for a given grade in mathematics would contain only a handful of topics and treat each one at considerable length, instead of treating many only superficially. Students of many ability levels would all spend enough time on each of those topics to master each one before going on to the next, and students generally would move on to the next grade ready to do the

work. The fact that this is not generally the case in the United States, but is generally the case in the best-performing countries, explains the success of those systems.

As in the case of my suggestions for a high school assessment system, we would do well to start with the best of what is already in use, in this country and abroad, and then build on it systematically. In this case, one could build on the British system for the pre–high school years, which has the advantage of being an integral component of a complete instructional system, or one could begin by building on the best of the American K-8 testing systems, such as the New England Common Assessment Program (NECAP) assessment or the Massachusetts Comprehensive Assessment System (MCAS).

At the outset, one would have to alter whatever system was chosen as the base to reflect the Common Core Standards, and to make sure that this K-8 system was aligned with the high school system proposed above.

After that, all the options are available for further improvements that were suggested with respect to the high school system. Again, it would make sense to put the best available system in place at the outset nationally, work out the implementation problems, and then proceed to improve it over time.

One might reasonably ask why I proposed a system that offers choices and multiple pathways at the high school level and did not do so for the lower schools.

Grades K through 8 are the grades in which the foundation is laid for everyone. Few would argue that that foundation should be different for different children. Some might go through faster. Some will need more support. Some might get an enriched curriculum. But everyone needs the basics, and that is what these grades are about.

The big variable should be time, not content. Because, at every grade level, so many are so far behind where they need to be, they will need more time and more support to be ready for the next grade when the next grade begins. The United States should do this the way Singapore does it. Singapore expect svery high-level performance from everyone and they get it. But some students get more help before school. Some get more help after school or on Saturdays or during the summer. But students are all ready to start school the following year on the same page. Our assessment system should make the assumption that that will happen, but our policymakers and educators have to make sure it does happen.

High school is another matter. The common curriculum should end with a qualification that says one is ready to start a 2-year or 4-year open-admissions college, ready to do the work that is required there. And then paths should diverge, because the common curriculum has been mastered. Even in the lower division of high school, we need to recognize that students are young adults and need choices that really don’t need to be available at the lower grade levels. From our research on this matter, I believe there is more consensus on this point among educators than one might think.

On Accountability

The system described above was designed, in part, to facilitate a tough accountability policy.

The data collected from the high school design will make it possible to hold the educators in the school accountable for the proportion of students that reach the college-ready standard, in absolute terms and relative to other schools with similar populations. The same thing can be said of the proportion of students that reach that standard at the end of the sophomore year, the junior year, and the senior year. The system will be able to report what proportion of the students enroll in the upper division programs and what proportion achieve high grades on their upper division exams. It will be able to relate these outcomes to the achievement of the students when they entered the high school, as well as the socioeconomic background of the students and their membership in certain protected classes, thus making it possible to measure the growth of the student body while in high school, in absolute terms and relative to schools enrolling similar student bodies.

What the data will not do is provide accurate measures of the contribution of each teacher to the growth of that teacher's students. But, because it will provide growth measures for the school as a whole (or the board examination program, if it operated as a program within the school), it could be used to reward the entire faculty for its contribution to the growth of the students, and those rewards could be distributed by the teachers and principal to the most effective among them. In an accountability system that put the whole school (or the faculty of this program) at risk if the students were not learning, the faculties of the schools would have a strong incentive to retain their most effective teachers and, therefore, to provide them with disproportionate rewards.

In high school and in Grades K-8, I would recommend that the data collected by the recommended system be used to trigger visits from professional school inspectors employed by the state, when that data appear to indicate performance problems in that school that need further attention. Student performance data alone cannot tell the state whether the school is in the midst of a precipitous slide or has just acquired a new principal who is doing everything needed to turn the school around, whether the school has the same sort of student body it has had for years, or whether it has just gotten an influx of students who speak no English, whether it has had stable funding, or the bottom has recently dropped out of a previously stable funding pattern. These points are all material, and only a visit to the site by people who know what questions to ask can reveal the problems and whether they are being competently addressed.

The K-8 assessments proposed herein will be able to provide the data that can trigger such visits. That data can also be used as the basis of a whole school reward system of the sort described above for the high school, and the scores on the summative tests could be broken down in the same way.

Summing up, the assessment system I have described could be used to report on the performance of all the students and subgroups of students in just the same way that the NCLB data are reported. Those reports could be supplemented by on-the-ground reports from professional school inspectors that put the raw data in perspective. The assessment system data can be used to report on the degree to which the school's students are on a path toward being truly ready for college and work. They can also be used

to calculate the value being added by the faculty of the whole school to the whole student body, by comparing student achievement when the students enter the school to their achievement when they leave the school, and comparing that rate of growth to the rate of expected growth and to the rate of growth for students in schools with similar student bodies.

Additionally, as mentioned above, these comprehensive instructional systems include formative assessments that provide teachers with regular feedback on whether students are on target at any point in the course, or, if they are not, extra supports to help them meet their performance goals.

Governing and Regulating the System

The system just described will not run itself. Someone has to decide what subjects are assessed at which grade levels, and what qualifications will be awarded, based on what evidence. Someone must decide what the cut scores will be and what standards must be met by organizations proposing to construct and offer high-performance instructional system to the schools. Someone must decide which of those proposed systems meet those standards. Someone must decide what content and performance standards the students should meet, subject by subject. These are not necessarily the same decision makers. Different decisions may need to be made by different bodies.

In this section, I sketch out what such a system of governance and regulation might look like. In doing so, I have tried to be sensitive to the realities of American culture, tradition, and law. No one is interested in having the federal government decide on a national curriculum. Many do not want the federal government writing national student content and performance standards. Some want as much choice for teachers, students, and their parents as possible. I have tried to accommodate all of these interests.

The NCEE is currently assembling a consortium of states committed to demonstrating the worth of the high school design I shared above. We have made a preliminary selection of the high-performance instructional systems named earlier. We have assembled a distinguished technical advisory committee (TAC) composed of some of the world's leading cognitive scientists, psychometricians, and curriculum experts to construct and employ the methods needed to array the selected examinations along a common scale. They will also design the methods that will be needed to ascertain the empirically determined level of literacy needed to succeed in the initial credit-bearing courses in the nation's open-admissions 2-year and 4-year colleges. This work is under way, supported by a grant from the Bill and Melinda Gates Foundation.

There is no reason why, once this research is done, these tools that we are developing cannot be applied to other fully developed high-performance instructional systems and their embedded examinations. Thus, other consortia or individual states could use these tools to align their systems with ours, under the general guidance of the TAC we have assembled. We would welcome that development.

As this is written, we are completing the work needed to determine the initial membership of states in our consortium. When that is done, we will assemble a governing board for the program, largely from the states that join the consortium. That governing board will represent the major constituencies that need to be represented in such a body, including the chief state school officers, governors, state school

boards, superintendents of schools, teachers, business leaders, legislators, and others. Our intention is that the voice of the chief state school officers will predominate in this group, since they have the lead responsibility in their states for the schools. This governing board will make the important policy decisions that have to be made for the group of states in the consortium.

Staff support for the work will be provided by the NCEE.

This structure will, I think, work well to get the system of high school assessment recommended in this paper off the ground. But, over time, if a large group of states joins this venture, and it begins to acquire the status of a voluntary national system, something else will be needed.

Should that happen, the Congress could charter a new not-for-profit corporation called the National Examinations Board. The charter could describe a process by which the Council of Chief State School Officers and the National Governors Association could appoint the members of the National Examinations Board, in such a way that the kinds of constituencies I just mentioned would be represented on the National Examinations Board. In this way, the National Examinations Board would be national but not federal, since its membership would not be appointed by federal officials, but rather directly by the states. This new corporation could be chartered by the Congress so that it is eligible for appropriated funds, to be distributed to it by the U.S. Department of Education, but without program direction from the Department.

The National Examinations Board would take up where our own governing board for the program had left off, with responsibility for the same functions at the high school level, as well as updating the Common Core Standards and curriculum frameworks for mathematical and English literacy. It would continuously update the research on evolving literacy requirements for success in the initial credit-bearing courses in our nation's open-admissions 2-year and 4-year colleges, and it would use that information to update the cut scores for passing the literacy requirements on the national examinations.

At the K-8 level, the system might evolve the same way. The U.S. Department of Education might fund one or more consortia at the K-8 level to start the process. My guess is that, over time, only one or a very few of these would turn out to be viable systems, attracting the funds and state support needed to sustain the effort. Each of these consortia would need to have a governing board of some sort to make the necessary policy decisions for the member states. In time, as the work at the K-8 level consolidated, the system that ended up growing most strongly could be brought under the National Examinations Board proposed above. It would update the Common Core Standards at the K-8 level, approve the curriculum frameworks for the core subjects in the K-8 curriculum, contract to have the summative assessments developed and administered in English and mathematical literacy and in science in Grade 8. It would supervise the system by which the formative and curriculum-embedded tasks, items, and mini-exams are produced. It would be responsible for collecting the data produced by the system at the school level and would report on the performance of the whole system. If state policymakers agree that the nation will not ultimately need multiple K-8 assessment systems, then the National Examinations Board would not have to moderate the standards among multiple examination or testing systems at the

K-8 level. However, if the National Examinations Board saw advantage in having multiple providers, it would continue to play this role.

There is another challenge that the National Examinations Board would have to deal with. As I pointed out at the beginning of this paper, NCLB, by focusing almost exclusively on mathematics and English literacy, and to some degree science, while creating high stakes for teachers to improve student performance in these subjects, has pushed the other subjects in the core curriculum to the background in many schools. No one believes that the other subjects in the core curriculum are unimportant, and no one I know thinks that students should get a performance-based school-leaving certificate without demonstrating some level of competence in these other subjects. How this is dealt with will be a function of what is tested and what kinds of stakes are attached to student performance in those subjects.

One way to deal with this problem would be for the National Examination Board to develop, over time, indicative curriculum frameworks for these other subjects and high school level exams for them, including syllabi. The states would not be required to adopt these frameworks, but they would be required to select from and administer high school exams in these subjects approved by the National Examinations Board and to report student performance on those exams to the National Examinations Board. The states could set their own pass points on these exams if they wanted to use them to determine who gets a high school diploma in their state. The National Examination Board would publish the state scores on these exams and, in this way, state residents would be able to compare student performance at the high school level in these subjects to performance in other states or the nation as a whole, and high school students would have powerful instructional systems available to them to enable them to succeed in these subjects.

This approach is not the only strategy available for dealing with this issue, and others might work as well or better, but it will have to be dealt with.

The National Examination Board would also be responsible for supporting an extensive program of research on assessment, on assessment technology, on curriculum and on the effects of the system on student performance, so as to be in a position to continually improve the system itself.

Many will ask how the functions of the proposed National Examination Board might be related to the functions of the National Assessment Governing Board (NAGB) and its National Assessment of Educational Progress (NAEP). I believe that the greatest value of NAEP comes from its audit function, as an independent check on the assessments used by the states. In this design, NAEP would continue to play that function. In countries that have high-performance instructional systems of the kind I have described, there are sometimes fierce arguments over whether the government in power has manipulated the standards for the exams in such a way as to enable them to report improved student achievement when in fact the standards were lowered. There is no way to resolve such arguments unless there is some way to independently assess student progress with stable instruments such as the NAEP tests. To ask NAGB to assume the functions just described as belonging to the proposed National Examinations Board would be to rob it of its independent audit function.

In Sum

With this approach, we can swiftly adapt and adopt the world's best assessment programs, as well as the tools they use to raise student performance to world-class levels. We can do this without taking the time they took to develop those tools or spend the money they spent, because it has all been done. In this way, we could get from the middle of the pack (at best) to the front of the pack in record time. In doing this, we would not be beating a new path. This is the way Singapore did it, by asking the University of Cambridge to build a customized version of the British O Level exams (these are the exams on which the International General Certificate of Secondary Education exams are based), along with the associated curriculum. No nation has leaped from the back of the pack to the front faster than Singapore.

But this approach does not foreclose the possibility of getting far out in front of the pack. There is, as I have pointed out, no reason why, once this system is in place, we cannot make the investments needed to build on what we have done to build the curriculum delivery system of our dreams and an assessment system to match. We have the experts, the technology, and the drive to do it.

What we should not do is behave as if we have to invent it all ourselves.

It is no small matter in these difficult times that the United States can afford to do all this, if we combine the proposed high school examinations system with the proposed move-on-when-ready system. We cannot afford not to, because the implementation of the system would itself generate the savings required to invest heavily in the extra, targeted instruction that our secondary students need to greatly improve their performance. No other proposal now on the table will do that.

Part 2: Responses to Guiding Questions

The proposals just made represent the author's best effort to conceive of the broad shape of a state-of-the-art system that the United States could use to develop, over time, a national voluntary instructional system that would include a voluntary system of examinations. As it happens, the organization that the author heads, the NCEE, has assembled a consortium of states interested in demonstrating the merits of the proposals made here for instruction and assessment at the high school level. The responses to the Guiding Questions received from Educational Testing Service found below apply to the examinations that this NCEE consortium will offer at the lower division level of our high school program.

Rigorous Standards and Good Instructional Practices

The system we propose is based on the use of the best instructional systems available worldwide in English for use in the United States. The organizations that offer these programs—the University of Cambridge International Examinations, Pearson/Edexcel, the International Baccalaureate Organization, ACT, and the College Board—have helped to set the standard for instructional practice for many years all over the globe.

The upper division exams (the ones we propose for use in the junior and senior year of high school) do not simply reflect international standards; they actually *set* the international standard. All over the world, students who apply to and are accepted by the world's leading universities take these exams, having taken the courses these systems offer. As the standards of these universities change, the standards for these exams change with them.

The exam systems we have selected for the freshman and sophomore year programs are designed to prepare students for the upper division programs leading to the upper division exams. Students who do well on them should do well on the upper division exams.

But the lower division examinations will be set to pass points determined by the actual requirements of the initial credit-bearing courses in the nation's open-admissions 2-year and 4-year colleges. There have been earlier efforts to establish empirically the requirements of college and work, but all suffer from important, sometimes crippling shortcomings. We are in the midst of a program of original research designed to greatly advance the nation's understanding of the actual requirements in this important arena. It will build on parallel efforts by others and combine what is learned from those efforts with original research we are conducting, using multiple data sources and multiple research methods, to produce a more precise and well-grounded set of standards for college readiness and work readiness than has heretofore been available.

Open-admissions colleges that accept students who have passed these board examinations will not have to administer placement exams to these students to determine who needs remedial courses, and students who pass them will not have to take remedial courses to begin credit-bearing courses. Because it is our community and technical colleges that provide most of the quality technical education in the United States below the bachelor's degree level, and because it is these institutions that provide almost all of the education and training for virtually all jobs that pay enough to support a family above the poverty line that do not require a bachelor's degree, we can say with confidence that students who pass our lower division exams will be on a course that will enable them to succeed in the programs they need if they are going to be prepared for further education or for work.

But the board examination systems that we plan to offer are not just assessments. They are entire instructional systems incorporating state-of-the-art instructional programs and practices. They consist of programs of study that constitute a full core curriculum, a thoughtful syllabus for each course, instructional materials matched to the syllabus, very high-quality examinations derived from the syllabus, professional scoring of the exams, and high-quality training for the teachers who will teach these courses.

The systems include both summative and formative evaluation. The summative scores are based both on timed examinations given at the end of the courses and, for some systems, on grades given on extended assignments contained in the syllabus. The examinations are typically based on prompts intended to elicit extended essay-type responses. Though some multiple-choice, computer-based items are used in some of the exams, this testing methodology, if it is used at all, typically constitutes only a small portion of the testing methodologies employed. Most use a variety of assessment methods, and

many include methods for assessing student work products that cannot possibly be assessed in timed tests or examinations, such as 20-page history research papers, robots designed and constructed to specifications, or a work of fine art. They typically include a rich assortment of released test items that can be used by the teacher during the course to assess student progress and to monitor the progress of the class and of individual students so as to course-correct whenever necessary and to provide individual students the support needed to make sure that every student succeeds.

Because the assessments are based on syllabi that call for learning the skills required to do complex analysis, as well as synthesize equally complex material; to apply what one has learned to unfamiliar, real world problems; to be creative and imaginative and so on, the curriculum and instructional materials provided by these organizations are designed to support the development of these skills, and the examinations are designed to assess them.

In all cases, the instructional systems offered at the lower division level will be aligned with the Common Core Standards.

Technology

The providers of the board examination systems we will be using make it their business to incorporate the most advanced technology available, consistent with their objectives. Each of them, however, does this in different ways.

In the case of QualityCore, ACT offers the program in two versions: a paper-and-pencil format and a computer-delivered/computer-scored version. Assessment core reports are delivered online through a secure, user-friendly interface that presents individual and group achievement data relevant to the user's role. Online reports provide local, state, and national comparisons of students' performance within each course, as well as evaluation of student progress toward college readiness on a course-by-course basis. Tests that include a constructed-response component are not available online for mathematics, science, or social studies because of the complexity of the constructed response system. Teachers and institutions are given online accounts to create rosters and input student information. These accounts grant access to example preparation materials and quizzes and later for online reporting of student scores.

Cambridge International Examinations' website provides teachers with direct portal access to the syllabi for the courses in the program and a wide range of other teaching resources. Among these resources are copies of previous years' examinations, examples of student responses that received A's and other grades, lesson plans, and grading guides. The teacher support website also includes teacher discussion groups that make it possible for teachers to communicate with other teachers about the challenges they face teaching the Cambridge courses and the solutions they have found to those problems. A similar set of resources is available to students over the Web that is specific to the courses they are taking. Assessments are currently in development to allow for digital portfolios of student work and computer-based assessment. Cambridge's scoring systems use scanning of student exams and other scorable work products to put them in digital form so that they can be transmitted worldwide to scorers in England.

Pearson/Edexcel is a world leader in online learning. The company is one of the world's largest users of online marking, administration, and onscreen testing. For its International General Certificate of Secondary Education (IGCSE) exams, Pearson/Edexcel offers a complete line of courses for teachers that are available on the web, as well as in face-to-face trainings. Reports on examination results for students, teachers, and schools are instantly made available online as soon as they are ready for release. While Pearson/Edexcel has introduced computer-delivered examinations in its General Certificate of Secondary Education (GCSE) program in the United Kingdom, it has yet to do this for the IGCSE program; however, we suspect that it is only a matter of time and scale before Pearson/Edexcel brings its leading edge technologies developed for its domestic audience to its international work, including its U.S. offerings.

Summative Assessments That Measure Growth and Readiness

The pass points of all of our lower division examinations will be set to the level of literacy in English and mathematics required to be successful in the nation's 2-year and 4-year open-admissions colleges. There is widespread consensus that students leaving high school should be ready for college and work, and assertions that the standards for college and work are the same. But much of the discussion has failed to distinguish between the standards for success at Harvard and the standards for success in the local community college, and they have failed as well to define what is meant by *work*.

Our program will define *college readiness* as having the literacy skills needed to enter 2-year and 4-year open-admissions colleges with a high probability of being successful in the initial credit-bearing courses in a wide variety of college transfer programs and terminal 2-year technical and career programs. Because students who succeed in their initial credit-bearing courses in these institutions without first having to take remedial courses have a much higher probability of completing a 2-year degree than those who must take remedial courses, and because these institutions offer transfer programs to 4-year colleges, we will define college readiness as having the literacy in mathematics and English needed to succeed in the initial credit-bearing courses in these institutions.

We will define *readiness for work* in the same way, because most jobs not requiring a 4-year degree that pay enough to support a family of four above the poverty line are jobs that require at least a 2-year college degree. Few high schools can afford either the specialized staff or the modern equipment needed to prepare students for jobs not requiring a 4-year degree that pay well.

However, the research has not yet been done to establish the level of literacy in mathematics and English needed to meet this standard. NCEE, as part of this program, is currently conducting the research needed to establish this literacy standard, which will be used to set the pass points on the board examinations we will offer to our consortium states. This standard will be based in part on original empirical research on the specific topics required in the initial credit-bearing courses in mathematics in a sample of programs in a sample of institutions in a sample of states, the reading challenge level of instructional materials required in a sample of courses in a sample of institutions in a sample of states, the writing demands found in a sample of courses in a sample of institutions in a sample of states, the conceptual frameworks of the most widely used placement tests, and the research that has been done

on the ACT and SAT® scores of students who enjoy success on the initial credit-bearing courses at the colleges they attend. We are confident that the research that we are doing on what it actually means to be college- and work-ready will make an important contribution to the tools available to connect high school and college seamlessly.

The examinations we will be using will be end-of-course examinations. They will be administered as high-stakes examinations under secure conditions and every effort will be made to make them highly reliable. But no such examinations will be administered at the beginning of these courses. By that standard, it will not be possible to calculate the value added by the teacher to the skills and knowledge of the individual students in these courses. But data will be available in most cases on student performance on the exams taken at the end of the preceding course in the course sequence, making possible calculations of value added from those comparisons. In some cases, for example, the case of a course in algebra preceding a course in geometry, grades on the preceding course may not be as useful in establishing value added as one might wish, but in many others, these comparisons should be very helpful. The British program providers typically package their courses as 2-year courses, so that the boundary between one year's program and the next year's program in a given subject matter should be almost seamless. It is also true that the British convention is not to divide mathematics the way we do into courses based on particular branches (algebra, geometry, trigonometry, and so on), but to integrate these branches every year. Here, too, it will be easier to calculate value added using the end-of-course examination data.

Data will be available on the backgrounds of the students in each class, and on their success rates on the examinations they take. It will be possible to calculate the scores of the students in the class on each examination relative to the scores obtained by other classes of students from similar backgrounds. It will also be possible to track the proportion of students that reaches the pass points on the lower division examinations at the end of the sophomore, junior, and senior years, and these proportions and these trend lines can be compared to the same metrics for other teachers and schools serving students of similar backgrounds.

It would, of course, be possible to develop secure examinations for all of these courses for the beginning of the courses as well as the end of them, but that would double the cost of the assessment program, in order to make it possible to track the value added by each teacher of each course. But we think that would be poor use of available funds.

Accessibility

All of the providers of the board examination systems we are offering are prepared to make reasonable accommodations for students with special needs and currently do. Pearson/Edexcel participates in England's Joint Council for Qualifications on Access Arrangements, Reasonable Adjustments and Special Considerations, which has developed an extensive set of protocols to forthrightly and responsibly address the range of accommodation issues that arise in high-stakes testing environments, and the participants abide by these protocols. They also offer an English-as-a-second-language (ESL) course and other second language courses as well.

Cambridge also offers an ESL course and provides a variety of accommodations for special needs students. These include extra time, adapted test forms, and assistance with reading and writing. ACT QualityCore provides accommodations for its paper-and-pencil tests, including large print, Braille, reader scripts, and audio cassettes. However, it presently does not offer an option for English language learners (ELL).

Technical Quality

The organizations that provide the examinations we are offering are among the most highly regarded suppliers of tests and examinations in the world. They typically employ scores of research scientists in the field of psychology, particularly cognitive scientists and psychometricians, to make sure that their assessment products are fair, valid, and reliable. In addition, NCEE has assembled a technical advisory committee (TAC) to set the examination results provided by these organizations to a common scale and to set the scores for the pass points on these exams to the mathematical and English literacy required to succeed in the initial credit-bearing courses in the nation's 2-year and 4-year open-admissions institutions. Furthermore, the TAC will provide independent commentary on the technical quality of the offerings of the board examination providers. The members of the NCEE TAC are:

Howard T. Everson, City University of New York, *Co-chair*

James W. Pellegrino, University of Illinois at Chicago, *Co-chair*

Lloyd Bond, Carnegie Foundation for the Improvement of Teaching

Philip Daro, America's Choice

Richard P. Durán, University of California, Santa Barbara

Edward H. Haertel, Stanford University

Joan Herman, University of California, Los Angeles

Robert L. Linn, University of Colorado

Catherine Snow, Harvard University

Dylan Wiliam, University of London

Reporting

ACT's QualityCore tests, be they administered by computer or paper and pen, are scored within 2 weeks of submission. Assessment score reports are delivered online to teachers, districts, and the state. Teachers receive student and classroom reports; districts receive student, teacher, school, and district reports; and the state receives all of these plus a state-wide analysis. The QualityCore reports provide local, state, and national comparisons of students' performance within each course, as well as evaluate students' progress toward college readiness on a course-by-course basis.

University of Cambridge reports are currently available online approximately 6 to 8 weeks after the exam dates (November exams are reported in January; March exams are reported in April; June exams are reported in August). Hard copies are sent 10 days later. However, Cambridge has assured us that for the United States they will be able to turn around exams in 10 business days if this is what the states desire. Their reports are available in a variety of forms, including individual student reports and school-wide reports. The University of Cambridge reserves the right to change scores after a review period. The initial score reports are called *provisional*.

Pearson/Edexcel reports are also available online approximately 6 to 8 weeks after the exam dates, and, like Cambridge, the group stands prepared to offer turnaround in 10 business days in the United States. Reports are reported online for students. Edexcel's *ResultsPlus* service offers analysis of exam scores across schools, classes, cohorts, and gender and compared to national averages. It also offers analysis of individual student scores, showing areas of strength and weakness as compared to skill maps (for math and science subjects). These also report on how students performed item by item.

Informing Instruction and Leadership

The strongest argument for adopting board examination systems is that they do not merely inform instruction, but actually provide the instruction students need to succeed on the examinations. They do this in myriad ways. First, they make the standards to which the curriculum and exams are set concrete and vivid. They do this by providing the narrative standards to which the curriculum and examinations are set (in the form of statements that students who complete the courses should know this and be able to do that), a library of the questions asked in previous years, and, perhaps most important, examples of student work that received top grades in response to those questions in past years. This not only provides much more concrete images of the standards, it also makes it possible for both teachers and students to internalize the standards, which has a powerful affect on instruction, because this kind of internalization of the standards leads naturally to both students and teachers being able to easily compare how the student is doing at any point in the instructional process to examples of how well the student ought to be doing, therefore enabling real-time course correction. This kind of internalization of the standards is the most powerful form of formative assessment available.

But there is much more to the influence of board examination systems on instruction than that. Each of these systems comes with its own approach to formative assessment, enabling teachers to assess the progress of their students during their teaching so as to correct course when they find out that their students are having difficulty with some aspect of the course.

Most powerful, however, is the availability of a robust curriculum, with all the implied supports for teachers and students, to help the students succeed on the examinations. No one needs to guess at what sort of classroom materials and activities will enable students to do well on the exams, because that work has been largely done.

Some years ago, leading researchers pointed out that tests could not be valid if the students had not had an opportunity to learn the material on which the test was based. This point was epitomized in the

phrase *opportunity to learn*. What board examination systems uniquely provide is just that: the opportunity for all students to learn the material on which they will be examined.

The implications for teachers are profound. Teachers are expected to be able to teach a particular curriculum, and to teach it well, to students of many different backgrounds.

This singular focus might be interpreted by some as a deprofessionalization of teaching. Actually, in our experience, the opposite is true. The Advanced Placement Program® (AP®) of the College Board, as I pointed out above, has the characteristics of a board examination system. These are the courses that the best of our high school teachers most want to teach. These courses, as defined by the College Board, provide a strong framework for the work of the teacher. But neither the AP courses nor any of the courses offered by the other board examination system providers are paint-by-the-numbers programs at all. There is a lot of room for teachers to figure out for themselves, and in collaboration with other teachers, the best pedagogical approach that will enable their students to get high grades on their AP exams. When I and my colleagues visited a high school in the East end of London serving almost exclusively students from mostly Muslim countries in the Eastern Mediterranean who were taking the GCSE programs, we were stunned to find that most of these students, who came from very low-income households in which English was rarely spoken at home, were going to college after they left high school. The professionalism of the teachers in that school was very impressive. None of them could conceive of making the progress they were making with these students without the instructional tools that the GSCE curriculum, the domestic version of the IGCSE, provided.

For many school leaders, agreement on the curriculum of the kind that is provided by a board examination system spells the difference between success and failure, because the board examination system provides a framework for the instructional core of school improvement that is now lacking in the vast majority of schools. With that framework, school leaders know what instructional materials are needed and what professional development and training would be most useful. They will be able to evaluate the competence of their faculty against a clear standard of performance, because it will be clear what the instructional program, as implemented in the classroom, should look like. School leaders will find it easier to explain the school program to students and their parents, and they will find that it will enable them to define their priorities when it comes time to build the school budget. In these and many other ways, adoption of a board exam system will greatly improve the ability of school leaders to lead and manage.

Leveraging Common Standards and Assessments

It is well-established that, when faced with a choice between teaching to formal standards and teaching to the tests by which their own performance will be measured, teachers tend to teach what is measured. So the Common Core Standards will not affect what goes on in the classroom until those standards are translated into tests or examinations on a widespread basis. The most expensive and time-consuming way to accomplish that goal is to develop a new set of tests or examinations matched to the new standards. The quickest and least expensive way to accomplish that goal is to modify existing

tests or examinations of high quality to reflect the Common Core Standards. That is what the State Consortium for Board Examination Systems proposes to do.

But we do not propose simply to create tests based on the new standards. We will leverage the new Common Core Standards to produce both curricula and examinations based on the new standards, which will have a far larger effect on what goes on in classrooms than simply producing new tests. That is real leverage.

Implementation Timeline

The states in our consortium are planning to start implementing the system in their high schools in the fall of 2011. Most will start in demonstration high schools with student populations representative of the student population of the state as a whole, in order to work out the bugs in their implementation plan before they go statewide. At least one state, however, is now on track to implement statewide beginning in the fall of 2011.

Cost

The examinations we are proposing to use are substantially more expensive than the typical American accountability test. So one would assume that full implementation of the system described here would be substantially more expensive for states, districts, and schools than the systems they are now using. But that is not the case. In fact, though these systems will cost more than the current system at the outset, as the system is scaled up, a point is reached at which the cost of the new system is the same as the cost of the current system. After that, as more students and schools are added to the system, the cost of the new system is actually less than the cost of the current system.

That is because, as the system expands within a school, high school classrooms in which courses for juniors and seniors are offered will begin to empty out. The students who would otherwise be in those classrooms will be in college instead.

We think most of the money that is thereby saved should be returned to the school to be used to pay for the examinations and to provide incentives to teachers to teach these courses; to faculty that succeed in enabling students from disadvantaged backgrounds to score well on the exams; and to students who get high scores, in the form of scholarships to college. A substantial share of this dividend should be returned to schools that use it to make greater investments in the support needed by students who enter high school far behind to take these board examination programs and succeed on them, including more time for them before school, after school, on Saturdays, and during the summer to get focused help on the areas in which they are particularly weak. The same goes for the resources the schools will need to provide targeted assistance to the students who take their examinations at the end of their sophomore year and do not succeed, so they will be able to pass them on the next attempt. After a few years, the average expenditure on our high schools will be pretty much the same as it is now, even though they are using more expensive examinations, but the taxpayers will be getting far more for their money.

Limitations

The strategy proposed here will not provide growth data for each student for each class, as discussed above. It may take a little longer to get the scores and grades back to the school after the test is given than it would if all of the examination questions were administered and scored by computer, though not much longer. It is possible that there might be a modest diminution of reliability (relative to computer-based, multiple-choice tests) in exchange for a considerable increase in validity, but we will not know that until our TAC has completed its analyses. And, lastly, because we will be starting with instructional programs and examinations that exist today, they may not have all of the benefits that a very large, multi-year development program could eventually bring. But, on the other hand, such programs fail as often as they succeed. The more ambitious they are, the more likely they are to fail. And, as I pointed out above, there is nothing to prevent us from investing heavily from the very first day in research and development to improve these existing programs, in which case the costs and the risks of failure are reduced and the speed with which the benefits could reach students would be increased. Think of this as akin to the continuous improvement strategies made famous by Japanese manufacturers in the 1980s and now widely adopted across the globe, where good products are steadily improved until they evolve into excellent ones.

Value Versus Burden

There are two primary components of the high school assessment design shared with the reader in this paper. The first is the use of board examination systems for instruction and assessment in the high school. The second is the use of the move-on-when-ready system to restructure the high school experience for students and to restructure the relationship between high school and college and work.

The burden of implementing the board examination systems as such can be brought into focus by thinking about what it means to implement two board examination systems with which most high school people are familiar, at least by reputation: the AP system from the College Board and the International Baccalaureate Diploma Programme from the IB. It is relatively easy for a school to sign up for the AP program. But it takes 2 years to go through all of the steps required by the IB organization to be certified as an IB school. IB has a set of requirements concerning test security on site and operates a continuing quality control system to make sure that the school provides a quality program that meets the IB requirements. If a school fails to meet these requirements, it loses its certification. Unlike the AP system, a school offering the International Baccalaureate Diploma Programme must offer all of the courses required by IB, and a student must take those courses and perform other requirements stipulated by IB, as well. In addition, IB has certain teacher initial and continuing training requirements for schools to get and keep their certification as an IB school. The other organizations offering board examination programs place somewhere along a dimension line between IB and AP on these points. Thus the burden associated with implementation of the board examinations systems varies with the system chosen.

The states in the State Consortium for Board Examination Systems will all require their high schools to offer complete core programs, rather than select one or a few courses. That will include AP, which does

not offer a diploma program, but which does offer courses that can be assembled into a diploma program.

A high school could, of course, offer one or more board examination systems at either the lower division level or the upper division level, or both, without also using the move-on-when-ready system that NCEE has proposed. Many have been doing so for years. But the states in the State Consortium for Board Examination Systems have signed up for both.

The burden for a state, with respect to the demonstration phase of the program, includes identifying high schools willing to offer one or more board examination systems to their students at the lower division level and at the high division level. It also includes identifying 2-year and 4-year open-admissions institutions willing to admit as fulltime students those sophomores who pass their board examinations and choose to apply to their institutions, and not require those students to take any remedial courses. And it requires the state to pass legislation or amend the regulations to allow the state to grant high school diplomas to students who pass their board examinations at the required level.

As the program goes to scale within a state, that state will have to establish some sort of executive authority to manage its statewide examination system, arrange for the financing of the system (including incentives offered to students, teachers, and schools), establish policies designed to focus secondary school teacher education on what is necessary to enable their graduates to teach these courses well, and set up an accountability plan to take advantage of the structure of the board examination systems and the data those systems will provide on the rate at which schools are preparing their students for college and for work.

These are not modest tasks. But the payoff could be immense, in terms of the improvement in student learning in high school, the reductions in high school dropout rates, the increase in retention rates in our nation's community and technical colleges, and the reduction in the need for remedial courses in our colleges.

References

Fuchs,T., & Woessmann, L. (2007). What accounts for International differences in student performance? A re-examination using the PISA data. *Empirical Economics, 32*(2–3), 433–464.