

Designing and Operating a Common High School Assessment System

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

Kevin Sweeney, Executive Director, Psychometrics, The College Board

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

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

Stephen Lazer, Vice President, Assessment Development, ETS

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

April 2010



I. BACKGROUND AND INTRODUCTION

Today, America's high school graduates need postsecondary education and training to remain competitive for high skill and high wage jobs in the global economy. Depending on the source, between 28 percent and 40 percent of first-time freshmen in four-year public institutions, and between 42 percent and 63 percent of first-time freshmen in two-year public institutions, enroll in at least one remedial course (Olson, April 25, 2006). Students who require remediation are more at risk of dropping out of college and not earning a degree (National Center for Education Statistics, 2004). Research has shown that individuals with a college education are employed at rates 50 percent higher than those with no postsecondary education, and there are even greater advantages for college graduates in terms of income. Policymakers, researchers, and educators have come to recognize the economic and social impact that results when an increasing number of students are not prepared for college and career success and fail to attain postsecondary credentials. To address these challenges, one of the goals of the current U.S. Department of Education is to sponsor efforts to develop common standards and assessments across states. However, summative assessments in high school present unique issues and complexities that must be considered in designing a common core assessment system.

The College Board, ETS, and Pearson have formed a collaboration to explore how innovative approaches and best practices in high-quality assessments can be applied to the creation of a common assessment system. In previous papers, we have shared high-level ideas related to the design of a common core assessment and the use of computerized adaptive testing. In this paper, we highlight and address some of the technical issues surrounding the design and implementation of a common high school assessment system.

II. DESIGN OF HIGH SCHOOL ASSESSMENTS AND ALIGNMENT TO STANDARDS

Common standards for college and career readiness (Common Core State Standards Initiative, 2009) should drive the design of high school assessments. The current standards efforts have largely defined common standards for high school at the level students will need to succeed in entry-level freshman credit-bearing courses in college or postsecondary career training programs that will lead to high skill jobs. This is an important distinction from secondary standards currently established in many states, which target the minimum level of skills and knowledge students are expected to master after two or more years of high school courses. The focus is on determining the college and career readiness (CCR) of students when they are in high school, which has major implications on the rigor of the standards, the design of the assessment, and the evidentiary basis to support the validity of these new common high school assessments. When assessments are primarily used as an exit criterion to certify previous learning or achievement, or as an admissions test to select among a large number of potentially qualified students for a smaller number of openings, different features and evidence demonstrating the validity of resulting decisions are required.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), as well as best practices in educational assessment, note that test development and accumulation of validity evidence begins with the explicit definition of how test scores will be interpreted.

There are many potential uses for high school assessments, and the U.S. Department of Education identified nine potential purposes in their call for comments:

1. *To inform teaching and learning*
2. *To determine school effectiveness*
3. *To determine teacher and principal effectiveness*
4. *To determine student readiness for college and careers*
5. *To determine if a student is on track for college and career readiness*
6. *To measure student growth or change in achievement*
7. *To determine high school graduation*
8. *To determine college course placements*
9. *To inform college admissions*

A single summative assessment cannot serve all of these secondary purposes. In fact, there are tensions between many of these uses, and there are constraints that impose significant operational requirements for other uses. At this time it appears that the primary purpose of high school summative assessments will be to determine if students have the skills and knowledge to succeed in college and postsecondary vocational training programs. Therefore, test scores will be used to inform individual students about their readiness for college and career training programs.

From the above list, we also believe that the most effective secondary purposes for a high school assessment will be: (1) to determine if a student is on track for CCR, and if not, to provide some information on the gaps and weaknesses, (2) to determine the effectiveness of schools and districts in preparing students for CCR, and (3) to potentially provide initial information that would assist in prequalifying for college credit courses in specific institutions or systems.

Each separate purpose requires specific types of evidence and each has large implications for assessment design. As noted above, high school assessments used to certify that students are college or career ready, or on track toward that goal, must be closely aligned to college- and career-ready standards. The assessment framework must include a variety of item and task formats that measure the breadth and depth of the skills required for success in college and career-training programs, not just those skills that are most convenient to measure (as is currently done through multiple-choice items).

The common core standards initiative has released drafts of its college and career readiness standards (Common Core State Standards Initiative, 2009). In English language arts (ELA), these standards include skills that address listening, speaking, writing, and complex, high-order skills, which appear to demand some use of student produced responses (e.g., constructed-response tasks). While multiple-choice items can measure a student's knowledge of concepts such as evaluating potential sources for research or understanding steps in the writing process, extended constructed-response prompts should be used to directly measure a student's writing proficiency in such areas as point of view/perspective, use of evidence and examples, and critical thinking. The interrelatedness of reading and writing skills can be more strongly emphasized and directly measured by having students read a long passage or excerpt and react to it in writing (e.g., literary analysis, synthesis of source material). While some speaking, listening, and media skills can be measured with multiple-choice items, a direct measure of the application of those skills should include actual listening and speaking tasks, such as listening to a conversation or lecture in order to determine the main idea and supporting details and giving a speech with certain goals in mind (e.g., persuasion or explanation of a process). Hence, such an assessment system derived for this purpose is likely to contain a mix of item types.

In the current draft of the College and Career Readiness Standards for Mathematics (Common Core State Standards Initiative, 2009), a primary stated goal is to enable students to achieve mathematical proficiency — a term that includes five components: conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition. The assessment framework, then, must provide the ability to evaluate students' work and the process students use to solve problems. For example, students must demonstrate how to construct mathematical proofs. Assessment frameworks that include student constructed responses expect students to be able to understand the work they do to reach an answer, to think strategically, to reason logically, and to be able to explain, through their work, the steps taken to achieve an answer. Again, this implies a radically different mix of assessment types (and associated management systems) than those currently in place. We have assumed that summative assessments will be one component within a comprehensive assessment system that includes formative and interim assessments. In this way, some skills (e.g., those that require performance on work over an extended period of time, such as multiple revisions of a writing sample or research) may be better suited for the latter assessments than a summative test.

Since high school assessments are also likely to serve as a tool within an accountability system, the assessments must have sufficient reliability and validity at the school level (within major subgroups) to permit comparisons of group performance within a school, district, and state over time. In addition, since between-school, district, and state comparisons will be desired, the assessments must also provide comparability across these educational units. An upcoming paper will address many of the features and issues related to the comparability of assessments.

Properly designed high school assessments can also serve a signaling function to inform students and educators about what is required for CCR. As individual students and schools understand the skills and the related achievement-level descriptors associated with readiness, it will help to inform and change the curriculum through K – 12. One should be careful, however, when tying assessment results to specific teachers or when attempting to make statements about student growth. The skills that students acquired in their ninth-grade courses will certainly impact their results on tests completed in 10th or 11th grade. Thus, failure to reach proficiency in a 10th-grade course may, in some instances, be a more accurate reflection of a failure to master prerequisite ninth-grade material than a failure of the 10th-grade teacher to prepare his or her students.

Finally, high school assessments that are tied to the skills required for college and career readiness may have the potential to inform college placement decisions. The CCR standards should provide a rigorous articulation between high school and credit-bearing courses in the first year of college. States may eventually desire to use common assessments to directly bridge the transition from high school to college. As such, assessment results may be used to inform students about whether or not they are prepared to enter college-level courses without remediation (e.g., English composition, freshman literature, college-level algebra, statistics). In fact, the California State University system and K – 12 public education systems have established similar systems, and other K – 16 collaborations are attempting to implement similar articulated assessments. One important issue to consider is the interval between completing the assessment and placement in college courses. Assessments administered during 10th grade may have limited utility in traditional course placement decisions but can provide early indications of readiness and areas in need of development. Even students who have the skills to enter a college-level algebra course at this time may actually lose ground if they do not continue to take math during their junior and senior years. These issues would need to be addressed in local and state policy. But knowledge that the high school assessment should inform higher education, and college requirements should inform what is included in the assessment, is an important first step in the design of these assessments.

III. END-OF-COURSE (EOC) OR END-OF-DOMAIN (EOD) ASSESSMENTS

Overall, there are two general approaches to summative assessment in high schools: (1) end-of-course (EOC) assessments (e.g., algebra I) or (2) end-of-domain (EOD) assessments (e.g., high school mathematics). Each approach has different strengths and weaknesses. Ultimately, the decision on which model to adopt by a state or consortium of states should be based on the intended purpose of the assessment (and this does not preclude a melding of or use of both together). This paper provides a brief discussion of the assumptions, benefits, and constraints with EOC and EOD assessments in terms of major features of a college and career assessment system.

End-of-course (EOC) Assessments. EOC assessments offer a tremendous intuitive appeal to various stakeholders. They are generally considered to be more closely aligned to instruction and curriculum. Assessment results can more directly inform curriculum and assist teachers in addressing areas of weakness. Students would complete the EOC assessments during a testing window that is likely to be closer in proximity to their instruction than EOD assessments, which often assess content that has been included in curriculum over several years.

As their name suggests, EOC assessments may also be directly tied to performance in high school and college credit-bearing courses. For example, the EOC assessment may play a role in determining the final grade of a student in a particular course. In addition, students are more likely to complete different courses at different times, making it more complex to aggregate results to inform growth. However, results from EOC assessments may be quite useful in helping schools to evaluate and improve particular courses (e.g., algebra II, ELA 10th grade) if a pre/post design were implemented. For the same reasons, EOC assessments may offer greater utility in evaluating teacher effectiveness if adequate controls are available to account for prior knowledge and differences in student ability across courses.

In an EOC system, it will be difficult to support meaningful statements of student growth by comparing performance on different EOC tests. Obviously, the greater the extent to which the tests measure different content, the less meaningful any comparisons become. Some comparisons, on the face of them, seem nonsensical, such as comparing performance on a geometry EOC test to an algebra II EOC test. Perhaps some statements of growth could be made as students progress from, say, algebra I to algebra II or from a less complex ELA course through more complex ELA courses. For such statements to be truly meaningful, these courses should be designed, from a content perspective, to support such cross-course comparisons. Here, content matter experts should be the final arbiters as to the reasonableness of such course design, and such a decision is not independent of the goals of the standards.

In high school, EOC assessments can be aligned to specific college credit-bearing courses (e.g., college algebra, composition), and the results may be incorporated into placement or prequalification for college courses. The difference between placement and prequalification is discussed next. Placement tests are typically used to determine whether students can be placed in specific courses and usually occur in very close proximity to course enrollment. If high school assessments are administered during 10th grade (or even 11th grade), they may have limitations for placement decisions because they will not account for the additional math courses students may or may not be exposed to in their last year(s) of high school. Consider the student who is proficient in math but does not continue to take math in high school and has to relearn much of what was lost, or the student who is not a proficient writer at the end of 10th grade but develops these skills later in high school. Colleges may be less comfortable in making precise placement decisions when tests have been taken 16 – 28 months prior to enrollment. However, it seems reasonable that EOC tests can be used as a

prequalification that certifies students who have a specific level of mastery in math or ELA will be placed in credit-bearing courses if they retain their level of proficiency. EOC tests could serve as a prequalifier, and students may then simply be required to complete additional math courses with grades above B- or take a short test that would verify the student is still proficient. Such a use could inform students whether they are on track to be CCR, whether they are CCR, or whether substantial developmental work is needed to become CCR. A similar system has been implemented by the California State University system in cooperation with the state board of education (CSU, 2009).

One of the primary advantages of an EOC system is that the assessment is closely aligned with instruction and curriculum. This strength is also the source of one of the primary drawbacks of an EOC system. Because a tight alignment between the test and the curriculum is desirable, there may be limited flexibility at the local level in curriculum, sequencing, and instructional content of a course with a state-mandated EOC assessment. Ideally, states within a consortium would first reach agreement on common courses, including the curriculum framework, sequencing, and pacing. However, if states and districts insist on retaining significant flexibility in high school course content and sequencing, EOC tests providing comparable student-level data will be more challenging. For example, if some districts insist on integrated math courses, other districts mandate algebra I for all eighth-graders, and remaining districts value a model that limits eighth-grade algebra to only the most able students, with other students taking algebra I in ninth or 10th grades, the decision on which tests to use for high school accountability is more challenging and will raise concerns about fairness and comparability of data.

One possibility is that states could agree that algebra II and ELA III benchmark courses form the basis for any decisions concerning CCR. In this way, the EOC assessments for these two courses would be used for CCR determinations. Thus, only two assessments would be required, and students in each subject would complete the assessment when completing the course. Certainly there would be differences in when students took these courses and assessments. Additionally, states would need to reach consensus on the instructional frameworks and content for each course. If algebra II and ELA III were the benchmark courses, alternative curriculum offerings (e.g., integrated math, literature-only course) may be more difficult to explain to the parents, and some schools would have to address how to assess students who today do not take algebra II in high school. Differences in the scope and sequence of curriculum and courses could introduce variability in assessment results used to compare school or state performance.

Another alternative is to offer multiple EOC assessments in more than one course within a domain. A state consortium may prefer to have several EOC assessments (e.g., algebra I, algebra II, geometry), and students may then either be required to take multiple tests as the basis for determining college readiness or be given some choice among assessments. The former model has been implemented in several states that use EOC assessments. Students may be required to take an assessment at the end of their math and ELA courses in ninth, 10th, and 11th grade. For example, Texas has implemented such a model, where performance on each EOC test counts toward 15 percent of the final course grade, and students are required to pass each test. Texas has chosen to focus solely on topics in advanced algebra as the measure of college preparedness in math, and to then distinguish the content associated with college preparedness from high school performance (TEA, 2009). Because the common core state standards for CCR include multiple content strands (e.g., geometry, algebra), it is more likely that performance across more than one EOC assessment would be required to determine college and career readiness. However, student performance on each assessment could be used in a conjunctive model to determine college readiness in such major strands or content domains within math and ELA.

One of the benefits of EOC designs is that the assessment is administered toward the end of the course, and the gap between instruction and assessment should be minimal in time and content. The downside to this is that if students must attain a minimum score (e.g., for graduation), then repeat testing is more of a challenge as the gap between instruction and assessment increases. This is less of a concern with a general summative assessment that surveys topics across math and ELA.

In discussing any EOC option, it is important to remember that the assessment system will need to measure the CCR content standards. Implementing EOC assessments, in and of itself, does not guarantee that the CCR standards will be assessed. Consequently, there will need to be an explication of the relationship between the course content assessed in an EOC system and the CCR standards. This is a critical step to the validity of the system and should play a primary role in deciding which EOC assessments to include and what content is covered in those courses. It is important to note that establishing the relationship between the CCR standards and the EOC assessments is a nontrivial point that will require a great deal of careful planning around course content and sequencing. We believe that the stakeholders who are advocating acceptance of the common core standards, and racing to join assessment consortia, will do well to think through this sticky alignment issue.

As described above, an EOC assessment system, while offering a relatively straightforward measurement of student capabilities related to individual courses, becomes complex as one considers issues of accountability, student growth, and assuring that the tests within the assessment system represent the content embodied within the CCR standards. Thus, for an EOC system, simplicity of interpretation at the course level is traded off with complexity of interpretation and implementation at a system level, or accountability level, or even college readiness level. By comparison, an EOD system is simpler from a systems and implementation perspective but offers fewer direct linkages between curriculum and instruction and student performance.

End-of-Domain (EOD) Assessments. This option is a more traditional design, where an assessment is administered at the end of a grade or range of grades (e.g., 9 – 10) and is more consistent with the current state assessment systems in grades 3 – 8. An EOD assessment system would presumably be based on a single ELA and math assessment.¹ The majority of states still employ this type of model to comply with No Child Left Behind (NCLB); however, states have increasingly been adopting EOC designs in the past several years. This approach is also typical of European assessment approaches in secondary schools, which rely on knowledge and skills across domains (e.g., math, biology).

Many of the same issues described previously also impact this design and will not be reviewed in the same detail as before. An EOD assessment design should provide more flexibility for states and districts in determining particular curriculum offerings because the content, scope, and sequence of math or ELA courses do not need to have the same level of consistency as long as students have an opportunity to learn the skills and knowledge included on the assessment prior to its administration. Schools, districts, and states would have more discretion in determining the types of courses student enroll in (e.g., integrated math vs. an algebra-geometry-algebra sequence) and the sequence of courses with a traditional EOD model.

EOD assessment designs will normally use a single assessment in each subject area and may, depending on relevant policy decisions, allow schools flexibility in when students complete the test. That is, it may be possible for some students to take the test during ninth grade if they have completed the relevant instruction, while students in less advanced courses may first test in 11th grade. Because a single math or ELA summative test is less dependent on instruction in a specific course (e.g., geometry), retesting does not present as great a challenge.

¹ EOD tests could also include multiple grade-level assessments (9, 10, 11; 9 and 11) that are not tied explicitly to specific courses and curriculum associated with the EOC model discussed previously.

EOD assessments may be less costly and complex to design and operate simply because it is cheaper and easier to operate a single assessment (e.g., math, ELA) across high school grades than multiple assessments (e.g., algebra I, geometry, algebra II). Currently, many states allow students to retest a year or more later on high school summative assessments. EOD tests can accommodate this requirement because they are not based on content in a specific course but rather on content across a subject domain. Schools have greater flexibility in structuring and sequencing content. Tenth-grade students who must retake a summative test in geometry may be at a disadvantage 6 – 10 months after completing the course. In addition, EOD assessments may also be more consistent with a computer-adaptive test (CAT) design because subject based tests generally assess mastery of all skills and content in a course. There are relatively few large-scale subject tests that employ a fully adaptive model in education, while adaptive tests are frequently employed in assessments of skills or larger content domains. This does not mean to imply that EOC tests cannot be adaptive, only that few currently exist.

Today, all national college placement tests are course or subject based (e.g., algebra vs. mathematics), as are most institutionally based tests. In addition, national placement tests in math and reading employ computer adaptive models, and writing tests are computer delivered with automated scoring (Achieve, 2007). Therefore, EOD tests appear to be less consistent with the design of current college placement tests, while EOC tests may be more easily migrated to such uses. Irrespective of the design and delivery model, any summative assessment should be designed with this use in mind if we anticipate scores to provide valid and reliable information for college placement. As with summative EOD assessments, EOC assessments are also not ideally suited to measure student growth, unless course sequence and content are comparable across schools and results across years and courses are tracked. This would require more central control over course sequencing and curricular offerings than many states and districts currently accept but could provide one method of tracking growth across different content strands and domains (e.g., algebra, geometry). In addition, pre/post testing can be employed as a simpler means of tracking learning and mastering in specific courses.

Following is a table briefly summarizing the advantages and/or limitations of EOC and EOD models for particular aspects of an assessment system:

Feature	EOC	EOD
Implications on courses offered	High. Will likely influence districts to have more consistency in the curriculum and fewer alternative courses in a domain (one algebra I course vs. three or four different levels for introductory algebra).	Low. Will not likely have any immediate influence on courses offered or sequencing.
Link between instructional and assessment models	Moderate to High. Can have a closer integration with formative and interim assessments, which are developed at the course level. Many EOC tests are used as a portion of the final course grade and can reinforce instruction.	Low. EOD tests in high school represent skills and knowledge associated with a content domain across grades and courses. In this way, they are less likely to provide a close integration with formative and interim assessments developed for a specific course.
Flexibility in sequence and structure of curriculum or courses	Low. The greater the variation in courses and content within a course, the less comparable test results will be. In addition, the link between assessment frameworks and curriculum will also decrease. An EOC model assumes greater consistency within and between schools in matters of curriculum and course content than does an EOD model.	Moderate. Clearly, variation in curriculum and content provide challenges to EOD assessments. However, these variations can be addressed more easily with EOD tests. CAT designs would be useful and could more efficiently focus on both the ability level of the student and major content areas.
Instructional impact	High. Can have a very tight alignment to curriculum and have closer integration with formative and interim assessments, which are developed at the course level. Many EOC tests are used as a portion of the final course grade and can reinforce instruction.	Low. Traditionally, EOD tests in high school are less likely to be aligned to specific courses but rather represent skills and knowledge associated with a content domain across grades and courses.
Number of potential exams	Most likely three (in mathematics: algebra I, algebra II, geometry; in reading/language arts: ELA I, ELA II, ELA III), although additional EOC assessments in pre-calculus or statistics may be appropriate for advanced students and more precise uses by higher education. Ultimately, this decision will be up to the stakeholder.	Most likely one summative exam in reading/language arts and one summative exam in mathematics at the end of grade 10 or 11.
Opportunities for students to retest	Low. Students will be best prepared for an EOC test at the end of the course. Students required to retest at a later date will face a large gap between instruction and assessment that reduces the effectiveness of using test results as an indication of learning (i.e., difficult to find frequent remedial opportunities).	High. In most EOD designs, students can retake tests for several years. Because the test is much less dependent on a specific course, the gap between instruction and assessment is not as important (i.e., remedial opportunities are likely to be far greater and more frequent).
Potential for using test results for college placement	Moderate to High. This is contingent upon the extent of K – 16 collaboration and data systems.	Low to Moderate. Results from general tests may not have sufficient coverage of specific skills in math (e.g., advanced algebra) to provide precise information for some college placement decisions.
Motivation factor for students	When performance on the EOC assessment is a component of a final course grade, it is more likely to motivate all students, including high-ability students.	EOD tests that are linked to outcomes like graduation may be more motivating than tests used only for school accountability purposes. However, high-ability students may not necessarily be as motivated on tests if they are not challenging and have no positive impact.
Measuring student growth	Low. Pre/post designs can be used to measure growth within a course only with dubious links between courses.	Low. Pre/post designs can be used to measure growth within a grade only.
Evaluating teacher performance	Moderate. Test results from a pre/post design would offer utility for evaluating the effectiveness of instruction in a specific course. This approach may be more useful in courses with specific content (e.g., algebra).	Low. Test results from a pre/post design may offer some utility for evaluating the effectiveness of instruction in a specific grade, but prior learning would influence student performance.

As argued above, both EOC and EOD assessment systems have distinct advantages and limitations. The decision on which goals and elements to optimize should help in choosing an assessment design for high schools. Alternatively, states may choose to develop both EOC and EOD assessment systems to take advantage of the strengths associated with each design. For example, the EOD assessments could be used primarily as a summative assessment for school accountability, while EOC assessments would be useful in providing greater instructional impact and specificity for student level decisions and placement decisions in higher education. Of course, this approach would be more expensive, and potentially result in situations where results from the two systems provide conflicting information about students.

IV. INTEGRATED ASSESSMENT SYSTEMS

The goals and intended purpose of accountability and instructionally actionable information cannot be met by a single assessment. An integrated assessment system is essential for a high school assessment and should incorporate both summative and formative components.

The assessment system should include a summative assessment that primarily serves the need for accountability and can provide an indication of students' preparedness for college and careers. High school assessments could provide criterion-related information on proficiency in major content or skill domains, normative information (e.g., school, state, and national percentiles), and projections for CCR (e.g., probability of success). It is important that the summative assessment reflect the skills and knowledge required for CCR, and its format (e.g., administration, design, tasks, projects) reflect the types of performance required in college and career training programs. Whether the ultimate summative assessment model employs an EOC or EOD design, we believe innovation must be introduced as results from research become available. For example, while we have described the desire to incorporate student performance captured during the year in a summative grade, as is currently done in many high-performing nations, we do not expect this capability will be incorporated in the first year of operation (Darling-Hammond & McCloskey, 2008). Interim assessments and projects can be incorporated into a summative performance score with end-of-course assessments or used to provide interim results about student progress in mastering major competencies required in the CCR standards. In either case, performance tasks that resemble the types of performance (e.g., research papers, laboratory experiments, scenario, or project based learning) expected of students when they reach college or career training programs can increase the fidelity and relevance of an integrated assessment system. However, we cannot assess these types of tasks until they become ubiquitous in the instructional flow. Interim assessments should immediately involve objective and performance tasks, as well as in-class and extended projects as they can be delivered by computer. Yet some assessments and projects will not likely be completed on computer. Hence, a combination of local scoring (by the teacher) and distributed scoring (within the school or state) can be used, depending on the task and requirements for scope and sequence. What is essential is that student performance is entered into the system so results can be tracked and instructionally actionable reports are provided to teachers, students, and parents. Involving teachers in scoring these tasks and projects will both provide valuable professional development and also give them insight into the meaning of the same standards used for the summative assessment. Interim results will inform students of their performance on major skills, their projected performance on a summative test, areas of strength and weakness, and how this relates to specific CCR standards.

Formative assessments should be an integral component of the assessment system and be incorporated into the curriculum (Perie, Marion, & Gong, 2009). In many models, formative assessments are used exclusively by teachers for instructional purposes. However, formative and interim assessments may take on other roles,

and these should be clearly stated prior to their development and use. In any case, the tasks and questions should be part of an easily interchangeable pool or system that involves a quality control peer review process and extensive professional development. Some assessments may be scored via computer, but complex, constructed-response tasks and projects will likely be scored locally by the classroom teacher.

Some final thoughts about an integrated assessment system. First, we believe that all assessments can be best served if they are delivered and completed using technology. We see this as most essential for the summative assessment because of the constraints paper would place on the use of innovative item types and how a construct is measured. Second, we also see the opportunity to provide immediate or near-immediate score reporting to students and schools if assessments are delivered exclusively on computer. Third, the potential savings in time and effort for the teachers, administrative staff, and others involved with the management of the physical assets (like test books and administrative materials) when moving to online is unknown but likely quite large. Associated “green savings” potential is also likely to result in the lack of transportation and production of physical testing materials once the assessment moves online.

In all instances, we believe teachers should be integrally involved in the design and construction of assessments. Teachers should be used in scoring all projects and extended tasks, and we have only suggested minimizing this for the summative assessment because of the requirements for higher stakes, which require greater reliability and integrity (e.g., teacher effectiveness, graduation, CCR), more rapid turnaround of results, lower costs, and consistency of results and scoring. Professional development is another essential ingredient of the integrated system and effective use of all assessments for instructional purposes.

All three types of assessments must be closely aligned and should not employ significantly different designs, formats, and interfaces. Finally, we believe that innovative assessments require significant amounts of research to evaluate new designs and item types. States interested in innovation must be willing to think out of the box and plan for both the collection of data to validate the assessment system and evaluate the consequences it has on students, schools, and the quality of education provided. As discussed in Section V below, states must have an aggressive plan to pilot and evaluate new item types using multiple psychometric and educational criteria and go beyond the current models, which often employ limited innovation on computer-delivered tests.

V. VALIDITY

The key to successful interpretation and use of the scores resulting from the high school assessment, particularly in light of multiple purposes, is a comprehensive program of supporting validity research. Because there will be many pressures for test scores to be used for purposes that the system was not designed to support, it is important that validity evidence exist to support each intended test use and to refute possible improper uses. Otherwise, appropriate and inappropriate test score uses become a matter of opinion and not a matter of fact. Furthermore, without a comprehensive portfolio of validity evidence, the perceived value of results from the system may be called into question. For example, it is probably not enough to show the relationship between rigorous course content in high school (like algebra II) and college success. Rather, other issues impacting this course content, such as sequencing and when in a student’s educational career he or she received the instruction, are also likely to be important.

The purpose of validity evidence is to establish the parameters for what are legitimate and what are illegitimate interpretations made from test scores and provide an empirical basis for the veracity of the claims being made. If the evidence does not support a particular interpretation, then an empirical basis to refute bogus claims exists; similarly, if the evidence does support a particular interpretation, then an empirical basis

of support for such claims exists. In the best case, this evidence becomes foundational data on which solid policy decisions are made. Without these data, important policy decisions are based on untested beliefs and hearsay.

Current state assessments are designed to measure state standards, not future outcomes. Therefore, the vast majority of validity evidence to support state assessments comes from a content validation strategy. States review assessment frameworks to ensure that they adequately map to state standards. When gaps are found between state assessments and standards, they are often identified as constructs that cannot be measured with a summative assessment. Contrast this validation evidence with the type used in other settings (e.g., admissions, employment) where the outcome of the test is compared against empirical outcome data. That is, admissions and employment tests incorporate concurrent and predictive validity evidence in their design because they are used to predict performance in a future setting (e.g., college, organization).

Given that the primary purpose of the high school assessment will be to determine if students are college and career ready, it is reasonable that high school assessments should be evaluated in large part by their relationship to student performance in college and career training programs. That is, to provide meaningful and comprehensive validity evidence, we must expand beyond relying solely on subject matter experts to decide if the assessment frameworks are comprehensive and if the proficiency levels are rigorous. Empirical results from future performance must also be a fundamental component in this validation plan. Validity research related to high school assessments and CCR should include rigorous predictive evidence that demonstrates a relationship with academic success at college and career training programs. External evidence must be collected in order to establish the validity of high school assessments in measuring CCR. Students who are considered college or career ready based on these assessments should be able to demonstrate college proficiency on a variety of external indicators. For example, students who are considered college ready in 11th grade should be able to attain a grade of 3 or higher on an AP course the subsequent year (or a corresponding grade in the International Baccalaureate degree program). They should also be able to attain the prerequisite score on most college placement tests, and ultimately, they should be much more likely to attain higher grades in freshman courses across a wide range of colleges and universities.

A particularly effective way to evaluate the validity of high school assessments is by conducting large-scale meta-analytic studies of performance across institutions (two-year, four-year, career training) using external data on CCR. Certainly, validation evidence based on content, construct, instruction, and consequences are also important in this effort, but predictive evidence is directly relevant in supporting future predictions. Local validation studies at a college or state level are frequently conducted; large multiple institution or national studies are more robust and relevant. Students in a particular state do not attend only institutions within that state or even limit their applications to regional public institutions. Local studies can often provide contradictory results based on small and restricted samples, and results may not generalize across all colleges and universities. National studies that incorporate institutional selectivity, institutional size, control (public, private), and region can provide appropriate contexts when interpreting scores and setting college readiness benchmarks. Results are strongest when they generalize across colleges and states, and meta-analytic approaches can provide a more robust strategy to supplement local studies (Hunter & Schmidt, 2004).

Achieve employed an innovative mixed-methods approach in establishing validation evidence for its multistate algebra II examination, which is particularly relevant to the types of summative assessments we propose for college and career readiness. Their evidence included more traditional content evidence (what skills and knowledge are associated with college readiness) and judgmental processes in recommending performance levels and descriptors. In addition, they introduced empirical evidence and relationships between scores on

the algebra II tests and other relevant measures of college success and mathematical achievement. Together, all evidence was evaluated for consistency in a more holistic process that included multiple sources of evidence and multiple methods (Achieve & Pearson Educational Measurement, 2009).

One final line of validity evidence that must be considered is the collection and use of consequential validity data. Consequential validity refers to the impacts that result from a testing program, both positive and negative. In high school, for example, an EOD assessment may result in improved teaching techniques, greater student learning, and more students being college and career ready. However, it may also result in a higher drop-out rate among a certain segment of the students, and in some advanced courses not being taken as frequently because students who have mastered the CCR content are focusing on more basic skills to ensure they pass. None of these impacts are directly related to the interpretation of test scores but are examples of potentially important consequences of the testing program. Although some testing experts dispute that addressing consequential validity data is the responsibility of test developers, current testing standards state that unintended consequences that result from test use should, at the very least, be investigated to determine whether they are related to issues of construct underrepresentation or construct-irrelevant variance. For this reason, consideration of consequential validity issues should be a fully integrated component of any comprehensive validity program of research.

VI. CONCLUSION

There are many decisions to be made and issues to be resolved in the establishment of high school assessments that focus on CCR standards. The focus on CCR standards is a significant departure from past statewide high school assessments. One of the fundamental decisions to be made is whether the assessment program is built from an end-of-course or end-of-domain framework. That fundamental distinction has implications for every other aspect of the testing enterprise, including, among other things, the number and characteristics of the examinations, how the assessment(s) relate to the CCR standards, the ways in which student growth can be measured, and the types of validity evidence to be collected.

This paper attempts to lay out some of the important issues to be faced by states and consortia as they consider implementation of a high school assessment system within the current Race to the Top framework. As with the other white papers, this one should be viewed as a starting point for discussions, not an endpoint or definitive solution.

REFERENCES

- Achieve. (2007). *Aligned expectations: A closer look at college admissions and placement tests*. Washington, DC: Author.
- Achieve, & Pearson Educational Measurement. (2009). *American diploma project algebra II end-of-course exam: Standard setting briefing book*. Iowa City, IA: Author.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. New York: AERA.
- California State University. (2009). *About the early assessment program*. Retrieved from <http://www.calstate.edu/eap/about.shtml>
- Common Core State Standards Initiative. (2009). *College and career readiness standards*. Retrieved from <http://www.corestandards.org/Standards/index.htm>
- Darling-Hammond, L., & McCloskey, L. (2008). Assessment for learning around the world: What would it mean to be internationally competitive? *Phi Delta Kappan*, 90(4), 263–272.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.
- National Center for Education Statistics. (2004). *The condition of education 2004*. Retrieved from <http://nces.ed.gov/pubs2004/2004077.pdf>
- Olson, L. (April 25, 2006). Views differ on defining college prep. *Education Week*. Retrieved from <http://www.edweek.org/ew/articles/2006/04/26/33college.h25.html?r=1196246650>
- Perie, M., Marion, S., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5–13.
- Texas Education Agency. (2009). *What's coming ahead: TEA update on end of course assessment*. Retrieved from <http://ritter.tea.state.tx.us/student.assessment/eoc/presentation/TAC2009-EOC.ppt>