



COMMENTS ON THE RACE TO THE TOP ASSESSMENT PROGRAM

General Assessment Input

Submitted by the College Board

December 2, 2009

Wayne Camara, Vice President for Research and Development, The College Board

(wcamara@collegeboard.org)

Kevin Sweeney, Executive Director of Psychometrics, The College Board

(ksweeney@collegeboard.org)

COMMENTS ON THE RACE TO THE TOP ASSESSMENT PROGRAM
General Assessment Input
Submitted by the College Board

The College Board is a national non-profit membership association of more than 5,600 schools, colleges and universities with more than a century of experience in the areas of standards and assessment. The College Board’s mission is to connect students to college success and opportunity, and it sponsors the SAT and SAT Subject Tests, PSAT/NMSQT, Advanced Placement (AP), ACCUPLACER, CLEP and other national assessments that reach more than seven million students annually. The College Board has strong partnerships with hundreds of states and school districts that rely on its assessments and other teaching & learning programs to prepare students for enrollment and success in college. The College Board has been a major participant in the Common Core State Standards project.

In the following text we would like to address seven major areas that the United States Department of Education should consider as it gathers information to inform the components of a request for proposals from states for a collaborative summative assessment program. These areas are:

- 1 Use the AERA/NCME/APA standards as authoritative guidance
- 2 Clearly define the purpose of the assessment system
- 3 Devise alternative methods of measuring student growth
- 4 Design a unified and integrated assessment system
- 5 Understand the importance of validity evidence in the design of the high school assessment
- 6 Incorporate innovation in the assessment system
- 7 Include teacher involvement, where appropriate

I. Use AERA/NCME/APA standards

The College Board recommends that the Department of Education formally recognize the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) as providing definitive professional guidance on the development and use of any assessments related to this initiative. In the request for input, the Department has called for “high quality summative assessments” that are based on “best practices in assessment.” In addition, the request appropriately requires that such assessments provide evidence relating to their validity, reliability, and fairness. The *Standards for Educational and Psychological Testing* have served as the definitive source for assessment professionals across a variety of applications (e.g., education, employment, licensure, psychological), and they delineate the appropriate types of evidence that are required to support statements by test publishers and users concerning these and other claims (e.g., comparability, use of cut scores). The Department of Education should ensure that any proposed summative assessments appropriately address these standards, and a technical oversight group should be established to review the proposed use(s) and evidence.

The *Standards* recognize that new assessments initially may not have all the documentation and evidence required to support inferences about validity. However, such evidence can be gathered over time and should be required of any assessment or accountability system. Indeed, because of the likely political and other pressures that will be placed on such a system, high quality validity evidence is essential to maintaining the integrity of the assessment system. In addition, the *Standards* note that “the applicability of the Standards to an evaluation device or method is not altered by the label applied to it, . . . the degree to which stimulus materials are standardized . . . or the type of response format (p. 3).” The *Standards* have been widely recognized as the definitive guidelines for the development, validation and use of assessments in a wide range of settings. They provide appropriate guidance and definitions on important technical issues associated with test development and use. It is important that the Department prominently cite the *Standards* in order to maintain a consistent level of quality, ensure common understanding about the types of evidence and documentation required, and ensure that any assessment practices adhere to current scientific findings and best practices. The alternative would be to allow each organization or consortium to define validity and fairness in its own way and thereby threaten the integrity and quality of assessments.

We recommend establishing a technical advisory committee of national assessment and content experts whose role would include adherence to the standards. The National Technical Advisory Committee (NTAC) or some other similar group could provide the Department with advice in developing RFPs and establishing criteria for their evaluation and use. It is important to note that, because not all aspects of the assessment system are driven by technical and psychometric issues, this committee should be advisory in nature and not a committee to determine final policy, although any ultimate policy committee should have representation from this technical advisory group.

II. Clearly define the purpose of the assessment system

Specifying the intended purposes of the summative assessment is the first step in designing a quality assessment. At least nine purposes have initially been mentioned in the Department’s call for inputs for the summative assessment:

1. To inform teaching and learning
2. To determine school effectiveness
3. To determine teacher and principle 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 or assessment system cannot serve all of these purposes equally well. There are tensions between many of these uses, and there are constraints that impose significant operational requirements for other uses. For example, summative assessments are not

designed to provide instructionally rich and actionable information. Typically, results are not available until the end of a school year, while diagnostic information is needed from the beginning and throughout the year.

Another constraint and conflict exists between the desire for innovative assessments that take advantage of technology and the use of the same assessments for very high-stakes individual decisions. Many state assessments are delivered by computer (although very few, if any, have achieved the desired goal of delivery exclusively on computer), but only when states permit schools to administer the same form (and/or items) over an extended testing window. There are simply not enough computers in schools to administer the same test to all 8th graders, for example, in a state on a single date (or even 3-4 different dates). School calendars also vary greatly within a state and flexibility in administration is required to accommodate local demands. Contrast this requirement with the security demands placed on tests used for college admissions, college credit and college placement. National testing programs have extensive procedures to ensure the security of test content and results for such high-stakes programs. The same items and forms cannot be administered over an extended window without greatly compromising security. In addition, the number of item pools and items required to maintain security of adaptive programs that offer the same level of flexibility for administrative dates would be cost prohibitive. These and other trade-offs need to be considered in determining the final requirements and purposes for an assessment system. The Department should identify a limited number of desired uses for a summative assessment system. In each instance, the consortium of states should then describe the types of evidence that will be used to support the validity of inferences that will be made for each purpose.

Testing at different grade levels may also need to take on different purposes. We believe that a summative assessment is not the best vehicle for providing diagnostic information to teachers and schools, and this issue is addressed later in the paper. However, a summative assessment in earlier and middle grades *can* be used to determine if students are on a path that will lead to college readiness. A summative test would ideally provide comprehensive information about student skills and mastery at a particular point in time, a measure of student growth during the academic year, an indication of whether a student has the knowledge, skills, and abilities required for success at the next grade level, and a metric that can be used as part of an accountability system for schools and teachers. At the high school level, a summative assessment may ideally be administered at the end of 10th grade to serve the above purposes, as well as to determine whether a student is prepared for college and career success. We believe that states should avoid attaching high stakes for students to this type of assessment during any transitional period. Moreover, when tests are used to determine graduation or college admissions, many operational and technical constraints arise that will reduce the flexibility and innovation desired for this assessment program. Graduation and admissions tests include a significant incentive to perform well on the test at all costs. Such proposed uses would require significantly more test items, test forms, and security, and they would also introduce significant operational constraints (limit dates of testing, require longer tests, greater reliability) and significant additional costs (more test items, more test forms).

III. Alternative methods of measuring student growth

Measuring student growth has long been an explicit goal of many state testing programs. However, because of technical, logistical, and cost constraints, this goal has been achieved with only mixed success. There are many lessons to be learned from the attempts to measure student growth, and we encourage the Department to speak with states and technical experts who have done it successfully, as well as with those who have not. Done properly, student growth data can be useful in both accountability programs and in providing information about individual student achievement; done poorly, student growth data will distort (either exaggerating or disguising) the amount of growth obtained.

Any meaningful discussion of student growth, however, requires a careful use of language. The term “student growth” is sometimes used as if its meaning were clearly understood by all parties and has a common definition. A cursory review of the research literature indicates that this clearly is not the case. Minimally, for example, student growth can be defined as relative to an achievement standard (e.g., student X scored five points closer to proficiency than on a previous test), relative to content standards (e.g., student X has displayed mastery on 4 of 5 objectives compared to mastery on 2 of 5 objectives on an earlier test), or relative to other students (e.g., student X is now at the 75th percentile, compared to the 60th percentile on a previous test). To be meaningful, all of these examples require that there be at least two points in time at which a student is assessed and that the results of these assessments be compared. It goes without saying that, for such comparisons, the results of the two tests must be comparable. A full discussion of what makes test results comparable is beyond the scope of these comments, but there are many important technical and logistical issues to be considered, and any assessment system purporting to measure student growth will need to work through these.

Each measurement of student growth provides answers to slightly different questions. There are at least three different questions that one can ask about student growth:

- 1) How much did student X learn this year?
- 2) How much more does student X know this year compared to last year?
- 3) How does student X compare to other students?

It is important to note here that each of these questions focus on different aspects of growth, and one cannot substitute for another. Consequently, we recommend that the RFP be clear as to what is meant by growth and what types of student growth are important.

Measuring student growth does not require the establishment and use of a vertical scale (i.e., placing test results from all grades onto a single scale). A vertical scale, while useful in many circumstances, has some limitations in measuring student growth in a K-12 standards-based assessment. Chief among these limitations is that in comparing, for example, the end of grade 3 to the end of grade 4, there is an important assumption that the grade 3 test is a good measure of grade 4 content (and vice versa). Because the content taught in grade 3 differs from that taught in

grade 4, this will rarely, if ever, be the case. With a vertical scale, any content that a student may have learned in grade 4 that does not overlap with content in grade 3 will not be captured in any measures of growth comparing end of grade 3 to end of grade 4. Grades 3 and 4 are used as examples here, but the logic applies to any pair or sequence of grades and may be of even greater concern in middle school and high school, where separate courses are taught (e.g., Algebra, Geometry). Several researchers (e.g., Lissitz & Huynh, 2003, Schaeffer, 2006) have discussed this issue extensively and make a compelling case for not using vertical scales in K-12 standards based assessments.

Despite their limitations, one of the reasons for the desirability of vertical scales is that they allow for statements of cross-grade growth. Often vertical scales have been adopted for this type of efficiency, and the instructional and curricular differences across grades have been overlooked. However, cross-grade growth can be measured in other ways (e.g., vertically moderated standards, growth percentiles), all of which have pluses and minuses. Whichever cross-grade growth model is employed (should one be employed at all), it is important that it be consistent with the stated purpose of the assessment system and that the strengths and limitations be clearly articulated.

In addition to cross-grade growth models, student growth can be measured—and depending upon the stated purpose of the assessment system, arguably, should be measured—via a within-grade growth model. This is consistent with a notion put forward by Laress Wise during his testimony in Boston. The measurement of within grade growth is a simple idea:

At the beginning of each school year, assess students on the material to be covered that year and use this initial measure as a baseline. At the end of the year, compare the end of year, summative test to the baseline measure to determine how much a student grew that year.

Various metrics can be established to ascertain how much improvement is adequate growth. This approach is direct in the interpretation of results and removes the troublesome problem of placing tests that measure different content standards on the same scale. Done properly, this approach can also provide initial diagnostic, actionable information about a student's areas of strengths and weaknesses at the beginning of the school year, when teachers can use that data to help students.

IV. Design a unified and integrated assessment system

We believe that the goals and intended purposes of this new assessment will be best served through an integrated assessment system that includes summative, interim and formative tests. In addition, we believe that the integrated assessment should be strongly aligned with the curriculum and that professional development will be essential to assist educators in connecting these elements. However, we will restrict our comments to the assessment system. The summative assessment can best provide useful information to students, parents, and schools on college and career readiness. Valid and reliable inferences can be produced for student and school level decisions. This information may also inform other decisions in time, such as course

placement, teaching and learning, and student growth or changes in achievement, if additional information is incorporated into the system beyond that collected during a single summative assessment. For example, a math test administered in 11th grade may not be the most precise way to predict how well a student will perform in a college math class some 18 months into the future. This is especially true when students score close to the cut point or when they fail to continue to take a math course in their senior year. Interim assessments can provide snapshots of how students are doing in mastering skills or providing more in-depth analysis of student weaknesses at a point in time. The formative components of such an integrated system can complement the summative and interim assessments and provide instructionally actionable information to schools and districts. A carefully designed integrated system is needed to ensure all components are complementary and consistent. Formative and interim assessments could utilize a common bank of assessment tasks and scoring rubrics available for teacher use.

The way in which the components of an integrated system are designed and work together will contribute greatly to the success or failure of the entire system. Consequently, although the current guidance is focused upon summative assessments, it would be short-sighted to not specify certain critical aspects of how the summative, interim, and formative components should work together.

Consistent with the above comments, we recommend an integrated assessment system comprised of the following three inter-related components:

- 1) Summative end of year
 - a. Grades 3-8: end of year
- 2) HS: end of domain (administered in grade 10)
- 3) Interim/Benchmark
 - a. Grades 3-8: minimum of 2 tests: baseline (at beginning of year) and midterm
 - b. HS: minimum of 4 tests:
 - i. Grade 9: baseline and end-of-year
 - ii. Grade 10: baseline and midterm
 - c. HS interim tests are not course specific but focus on college readiness
 - d. Test items are calibrated onto the same scale as the summative tests
- 4) Formative.
 - a. Most teacher involvement
 - b. Teacher scored

Ultimately, the summative and interim tests should be computer administered and, if possible, the summative assessment should be computer adaptive. The interim tests should be content focused and may not need to be adaptive. For the summative tests, the item types would be designed so that they are computer scorable. This summative design would facilitate: (a) quick turnaround of results; (b) increased use of innovative item types; (c) lower operational costs with higher fidelity items; and (d) greater ‘diagnostic-type’ information on college and career readiness.

For the interim and formative assessments, we recommend that decisions be made locally as to the item types and degree of teacher involvement in scoring. Allowing such local or state control will promote greater buy-in to the entire system and allow schools and districts to make the determination of valuing quicker turnaround time over teacher involvement in scoring more complex item types. Projects and performances can easily be integrated into interim assessments, and once they have been refined and evaluated, they could be integrated as a component of a summative assessment. However, this type of transition will require additional time to ‘try-out’ and evaluate the model and tasks, which is best done before they are incorporated into a summative assessment.

In the proposed system, items available for the interim assessments would be scaled onto the same theta metric as the summative test to allow for growth comparisons. These items would come from a common item bank, which would accept contributions from teachers and others. Projects, performances and extended tasks (e.g., out-of class assignments, in-class research) could also be included if standardized with well-developed rubrics for scoring. Additionally, it should be possible for off-the-shelf tests that demonstrate content and psychometric congruence to be used as interim assessments. In this instance, these instruments must be scaled (via a special study) to the summative scale.

Within this proposed system, all components should be designed to assess the same content standards. In this model, within-year growth can be measured by comparing the interim baseline assessments to other interim tests and baseline to the summative end of year test.

In this proposed system, the high school summative assessments would focus exclusively on college and career readiness and not be course specific. This is in keeping with one of the stated purposes of the assessment system.

The main advantages of the system outlined above are that it allows for measuring student growth, enables the measurement of performance against standards, and has the capacity to track students for college and career readiness. Additionally, it includes the capability of teacher scoring, but does not require it for interim assessments.

The main disadvantages of the system are that it requires universal access to technology, requires innovative item types to be developed and piloted, and requires a sophisticated database infrastructure to support relationships between interim and summative assessments.

V. Design of the high school assessment and the importance of validity evidence

The Department has stated that demonstration of college and career readiness is a priority of the RTTT assessment system, and that the high school test should focus on college and career readiness (CCR). We fully support this position and believe that the high school assessment should be consistent with this vision. A focus on CCR in high school, coupled with the options for differential course taking patterns in high school, is logical for this component of the assessment system. We must recognize that high school assessments must begin to evolve in a

different manner and design than the K-8 portions if assessments are to be relevant for higher education and career training programs.

Because the Department desires a system that supports the assessment of CCR, we recommend that assessments be made at the beginning and end of each grade to assess each student's status relative to college and career readiness. End-of-course assessments, while valuable, do not assess the same standards at the same level as an assessment focused on college and career readiness. If one wants to know the status of a student relative to college and career readiness, then assess that directly. End-of-course tests will present additional challenges to measuring student growth and obtaining agreement across schools, districts, and states. True "opportunity to learn" requires that students are allowed to take an end of course test at the completion of a course and not have to wait several years to take the test. This means that some students in middle schools may be taking the same Algebra and Geometry end-of-course tests as students in upper high school grades. It also means that schools may be administering different tests to different students in the same grade. All of these issues will complicate the use of such test results for school or teacher accountability. Our research often illustrates that students taking a test in 9th and 10th grade outperform students taking the same test as 11th or 12th graders. This phenomenon is more related to differences between the students than differences in school or teacher effectiveness. Students who are taking advanced math courses in earlier grades are generally at a higher ability level than the population of all high school students.

To assure that the defined purpose(s) of the assessment are being met, a comprehensive program of validity research must be established. 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. Such a state of affairs ultimately undermines the credibility of any testing program. The purpose of validity evidence is to establish the parameters for what are legitimate and illegitimate interpretations to be made from test scores, as it provides an empirical basis for the veracity of the claims being made. If the evidence does not support a particular interpretation, then there exists an empirical basis to refute bogus claims. Similarly, if the evidence does support a particular interpretation, then there exists an empirical basis to support such claims. 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.

While validity evidence is among the most important information about a testing program, currently most state testing programs provide a very limited amount of validity evidence to support the claims made from statewide test scores (Sweeney, 2009; Sireci, et al, 2009). There are a variety of reasons for this state of affairs, chief among them being the costs and difficulty in obtaining good data to do strong validity work. Another limitation with current state assessments lies in the criteria. Current state assessments are designed to measure state standards not future outcomes. Therefore, the vast majority of validity evidence to support state assessments comes solely 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 justified 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 score 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, we believe high school assessments should be evaluated in large part by their relationship to student performance in college and career training programs. That is, states must break away from relying solely on subject matter experts to decide if their assessment frameworks are comprehensive and if their proficiency levels are rigorous. Instead, empirical results from future performance must be incorporated in this validation plan. Consequently, we urge the Department to make validity research a fundamental component of any assessment program and to provide funding specific to the collection of validity evidence. Validation efforts at the state and local level should not be used as the primary focus because, for example, students in New Jersey do not just think about going to college in New Jersey, but are often applying to public and private colleges throughout the country. College readiness results must be generalizable across colleges and states, and meta-analysis is a far more robust and superior validation strategy than supporting local studies that will produce slightly different results because of sampling and other methodological issues (Hunter and Schmidt, 2004).

In sum, we believe that 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 Bachelorette 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 freshmen courses across a wide range of colleges and universities. If students are deemed proficient based primarily on current content-based evidence and judgmentally derived standard settings, but do not achieve these outcomes, then it is evident that the tests and proficiency levels are simply not established as CCR. Similarly, if the proficiency levels are set so high that students who are successful on these external metrics are not deemed CCR, then the tests and proficiency levels are set at a level beyond what is currently required for post secondary academic success. The best way to evaluate the validity of high school assessments is by conducting large scale meta-analytic studies of performance across institutions (2-yr, 4-yr, 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. We do not believe that funding hundreds of local or state validation efforts will be effective, because the conflicting data by state and institution will lead to a false perception that CCR differs by state and institution and will lead to greater confusion among students and parents. In today’s global environment we must establish college and career readiness indicators that generalize across state and national lines.

VI. Incorporate innovation in the assessment system

Innovation can be realized most efficiently in a large scale testing program if it is delivered exclusively on computer. Innovative item types, extended performances and different response formats can be more efficiently captured and scored with the use of technology. Innovation in large scale assessment has been hampered by the requirement to produce comparable forms on paper. If the assessment is administered solely on computer (with the exception of paper administration as a special accommodation), it will be easier to introduce new item types such as simulations, scenario-based tasks, or performance tasks. Ideally such tasks in the summative assessment can largely be scored by computer to increase efficiency and reduce turnaround time. Teacher scoring of formative and/or interim assessments can be best utilized in a distributed scoring network or through an audit function.

Many of the emerging skills contained in the draft Common Core State Standards can likely not be measured with paper-based assessments alone. Maintaining parallel paper and computer systems would likely limit innovation and the range of emerging skills that could be measured. This is another example of the trade-offs that must be considered in the final design of assessment systems that will be proposed by state consortia.

Another option is to incorporate results from interim assessments or actual student performances that occur throughout the year into the summative assessment score. Currently, summative assessments are based on what a student does at the end of the year on a single test date, while some high performing nations have incorporated student performance at several different points in time into their summative assessment. Results from interim assessments or tasks completed during the year or student performance on a highly structured in-class or out-of-class assignment (e.g., research paper, literary report, laboratory report, presentation) that is scored by teachers using a detailed scoring rubric could be incorporated into the results of summative assessments. Clearly such models present operational challenges in terms of security and when students transfer into a school midway through the year, yet such models could increase the instructional relevance of assessments and work for the vast majority of students.

“Computer scored” items does not necessarily mean multiple choice items, but may also include simulations and other performance tasks that can be objectively scored.

An important question is how states can transition from their current assessment system to a more innovative and integrated system as proposed by the College Board and other leaders in education. Clearly each state will need to address the specific mechanisms for transitioning data systems, proficiency standards, and reports. State specific approaches may ultimately be required for many of the operational issues, but we believe that such a transition can be accomplished more easily if technology is incorporated in the assessment system and there is additional federal support to prepare states for such a migration. A transitional or interim approach may be required but it is doubtful that it will meet many of the desires expressed by the Department for comparability and a new generation of assessments that are directly mapped to the Common Core State Standards. It may be more effective to spend limited resources to fund development

on the integrated assessment and supporting curriculum for 2014 than to build an interim solution for 2012 that will fall far short of most goals and objectives.

VII. Include teacher involvement where appropriate

Teacher involvement will be a critical component of the success of any assessment system of the type proposed. However, current state testing program experiences have taught us that teacher involvement is most beneficial when it allows teachers to learn from each other and to develop skills needed in the classroom.

Several states have learned that teacher involvement in operational scoring of summative tests is neither cost- nor time-efficient. Including teachers in the scoring process for constructed response items is problematic from both a technical and logistical perspective. Logistically, getting teachers out of the classroom for the necessary training time and scoring time within the bounds of an operational testing program will likely be challenging. Technically, the characteristics that make one a good teacher are different from the characteristics that make one a good scorer; and good scorers are needed to assure that the assessment is scored validly and reliably. This is not to say that teachers cannot be good scorers, but merely points out that one's ability to teach and one's ability to score are independent attributes.

If the intent to include teachers in the scoring process is for the professional development of teachers, then this should be done outside of the operational scoring window. Goals of professional development can be better met by including teachers in the development of the constructed response or performance task items or in an audit function of the scoring of those items. Teachers play an integral part in the development and scoring of Advanced Placement exams. We believe that professional development that provides teachers with greater insight into the assessment frameworks and student performance levels can be accomplished through involvement in the development of assessment tasks and scoring of formative and integrative components in the short term. As described in an earlier section of this paper, the interim or formative tests are much better suited for teacher involvement than are the summative tests. As noted earlier, once such a system has been in place we can then examine ways to more effectively integrate interim or benchmark tasks or projects into the summative component, as well as use teachers for scoring these elements.

In closing, the College Board is pleased to have the opportunity to share these views, and we would welcome the opportunity to respond to any questions you might have about our comments. We deeply appreciate the Department's strong leadership in pursuing common core state standards and establishing common state assessments through this competitive grant process.

As an organization, the College Board has a history of working with states, districts, and schools in a variety of capacities related to assessment practices. For example, we currently work with the state of Maine in providing the SAT for use as its high school NCLB assessment and have statewide agreements for use of the PSAT/NMSQT and Advanced Placement exams.

Wayne Camara is the Vice President of Research and Analysis at the College Board. He serves on several state technical advisory committees and is a member of the technical advisory committee for the Achieve end-of-course algebra assessment. Dr. Camara is the president-elect of the National Council of Measurement in Education (NCME), which works closely with state leaders on issues of assessment.

Kevin Sweeney recently joined the College Board as the Executive Director of Psychometrics. Prior to joining the College Board, he worked for 11 years providing psychometric expertise in K-12 statewide assessments. In that work, Dr. Sweeney was responsible for the psychometric design and implementation of NCLB assessments for several states, including the Massachusetts Comprehensive Assessment System (MCAS) and the New England Common Assessment Program (NECAP).

References:

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, D.C.: American Educational Research Association.

Hunter, J.E., and Schmidt, F.L. (2004). Methods of meta-analysis: Correcting error and bias in research findings. Thousand Oaks: SAGE.

Lissitz, R. W. & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research, and Education*, 8, 1-10.

Schaefer, W. D. (2006). Growth scales as an alternative to vertical scales. *Practical Assessment, Research and Evaluation*, 11, 1-6.

Sireci, S.G., Meng Y., Hanwook Yu, H. and Zenisky, A. (2009). Building Validity Arguments for Educational Testing Programs. Paper presented at the annual conference of the Northeastern Educational Research Association, October 22, 2009, Rocky Hill, CT.

Sweeney, K. P. (2009). Focus on Validity: State Practices in Obtaining Validity Evidence. Paper presented at the annual conference of the Northeastern Educational Research Association, October 22, 2009, Rocky Hill, CT.